Lars Gunnar Johnsen

Surgery with Total Disc Replacement Compared to Rehabilitation in Patients with Chronic Low Back Pain and Degenerative Disc Disease

Clinical, Health Economical and
Biomechanical Perspectives

Thesis for the degree of Philosophiae Doctor

Trondheim, February 2014

Norwegian University of Science and Technology
Faculty of Medicine
Department of Neuroscience

NTNU – Trondheim
Norwegian University of
Science and Technology

NTNU
Norwegian University of Science and Technology

Thesis for the degree of Philosophiae Doctor

Faculty of Medicine
Department of Neuroscience

ISBN 978-82-326-0084-7 (printed ver.)
ISBN 978-82-326-0085-4 (electronic ver.)
ISSN 1503-8181

Doctoral theses at NTNU, 2014:80

Printed by NTNU-trykk

Kirurgi med skiveprotese sammenlignet med multidisiplinær rehabilitering.

En randomisert, prospektiv multisenterstudie.

Denne avhandlingen er basert på en sammenlignende studie av to behandlingsalternativer for pasienter med degenerativ skivelidelse i korsryggen og kroniske lave ryggsmerter. Kroniske ryggsmerter utgjør et betydelig helseproblem. På global basis er det anslått at 10 % av alle leveår med funksjonstap pga ikke-fatal sykdom skyldes ryggsmerter. Mange behandlingsformer har vært foreslått men det hersker fortsatt stor uenighet om hva som skal være standard behandling. Flere kirurgiske behandlingsalternativer eksisterer. Skiveprotesekirurgi med fjerning av mellomvirvelskiven og innsetting av skiveprotese tar sikte på å bevare bevegeligheten i de affiserte degenerative segmentene til forskjell fra standard kirurgisk behandling som er avstivning. Sammenlignende studier mellom protesekirurgi og avstivningskirurgi har vist omtrent like resultater. En del studier har sammenlignet avstivningskirurgi mot ikke-kirurgisk behandling. Noen av disse studiene viser best effekt med kirurgi mens andre viser best effekt med ikke kirurgisk behandling. Vår studie er den første som sammenligner skiveprotesekirurgi mot et ikke-kirurgisk behandlings alternativ.

Rammen rundt studien var en randomisert kontrollert multisenterstudie med 173 pasienter rekruttert fra ryggpoliklinikkene ved alle fem universitetssykehus i Norge. Pasientene ble randomisert enten til kirurgi med innsetting av skiveprotese eller til et ikke-kirurgisk behandlingsopplegg i form av tverrfaglig (multidisiplinær) ryggrehabilitering. Oppfølgingstiden etter primærbehandlingen var 2 år med kontroll etter 6 uker, 3, 6, 12 og 24 måneder. En rekke kliniske utfallsmål ble brukt med ODI (Oswestry Disability Index) som hovedeffektvariabel i en klinisk og en biomekanisk studie og EQ-5D som effektvariabel i en helse økonomisk studie. To år etter oppstart av primærbehandlingen ble alle pasienter i tillegg undersøkt og vurdert av to uavhengige observatører blindet for behandlingstype.

Vi konkluderer med at både kirurgi med skiveprotese og ikke kirurgisk behandling med multidisiplinær rehabilitering gir signifikant bedring i livskvalitet etter 2 år hos selekterte pasienter med kroniske ryggsmerter og degenerativ mellomvirvelskive. Bedringen var størst i kirurgi gruppen men selv om forskjellen mellom behandlingsgruppene var *statistisk* signifikant, så var den ikke sikkert *klinisk* signifikant. Samtidig viser vi at det sannsynligvis er andre egenskaper ved protesen enn bevegelighet som bidrar til denne (signifikante) bedringen av livskvalitet. Videre er det usikkert om skiveprotesekirurgi vil være et kostnadseffektivt alternativ til ikke kirurgisk behandling da kostnadseffektiviteten i denne studien var svært avhengig av hvilken livskvalitetsindeks som ble brukt som effektmål. I en analyse av noen sentrale psykometriske egenskaper ved disse indeksene konkluderer vi med at det er stor forskjell mellom dem, at de ikke kan brukes om hverandre og at valg av indeks som effektmål i kost-effekt studier generelt sannsynligvis bør relateres til diagnose og/eller behandlingstype.

Kandidat: Lars Gunnar Johnsen

Institutt: Institutt for Nevromedisin

Veiledere: Ivar Rossvoll, Gunnar Leivseth, Peter Fritzell

Finansiering: Helse Øst HF**,** Extramidler gjennom Helse og Rehabilitering v/ Ryggforeningen i Norge, UNIMED Innovasjons forskningsfond, Samarbeidsorganet Helse Midt-Norge og NTNU

Ovennevnte avhandling er funnet verdig til å forsvares offentlig for graden PhD i klinisk medisin.

Disputas finner sted i Auditorium KA11 i Kunnskapssenteret
Fredag 7. mars 2014
kl.12:15

CONTENTS

AKNOWLEDGEMENT

This study is a result of research collaboration between the university hospitals in Norway. Papers II – IV in this thesis was accomplished at the National Centre of Spinal Diseases (NSSL), University Hospital of St Olav, Department of Neuromedicine, NTNU in Trondheim.

The study was funded by the South Eastern Norway Regional Health Authority, EXTRA funds from the Norwegian Foundation for Health and Rehabilitation, through the Norwegian Back Pain Association, the UNIMED innovation research fund and the Liaison Committee between the Central Norway Regional Health Authority (RHA) and the Norwegian University of Science and Technology (NTNU).

First of all, I wish to thank the patients participating in this study. By sharing your stories with us you have expanded our knowledge of the treatment of low back pain.

Several individuals have contributed to the completion of this thesis. Their reading and comments of the manuscripts have been most valuable. In particular I give my thanks to:

Ivar Rossvoll - My principal supervisor both in the work for this thesis and in my education as an orthopedic surgeon. Thank you for always finding the time to discuss any problem with me and for encouraging me to embark on this project in 2004.

Øystein Nygaard - Leader at the NSSL. Thank you for your enthusiasm and faith in me and for introducing me to the very inspiring environment at the NSSL and for providing me with the best working conditions possible.

Gunnar Leivseth - My supervisor concerning the biomechanical study. Thank you for inspiring conversations – not only about biomechanical issues but also about life in general.

Peter Fritzell - My supervisor concerning the health economic study. Thank you for introducing me to the exciting field of health economics and your enthusiasm during the writing process.

Hege Andresen – Research nurse at the National Center for Diseases of the Spine. Thank you for everything! Without you, this study would never have been possible.

Kjersti Storheim – Central conductor of the study. Thank you for your helpfulness and for conducting the study forward at all times.

I also wish to thank Jens Ivar Brox, Christian Hellum and Margreth Grotle for valuable and instructive comments on my work, Oliver Grundnes, Magne Rø, Marit Pedersen and all the people in the Norwegian Spine Study Group for your cooperation.

Finding time to conduct clinical research next to regular clinical work is a challenge. This process has however been facilitated by my leaders at the Orthopedic Department at St Olavs Hospital throughout these years. I would therefore like to thank Henrik Sandbu, Erik Rødevand, Arild Aamodt, Vigleik Jessen and Robert Buciuto for being very flexible in giving me time off whenever needed.

I would also like to thank all the people at NSSL especially Janne-Birgitte Block Børke, Bjørn Skogstad and Ulrik Schattel for your kindness and your way of letting me feel I am always welcome.

Finally, I would like to thank my wife Kirsti, love of my life and best friend, and our children Aurora, Herman and Selma for making my life worthwhile.

To my father

Dagfinn Georg Johnsen

(1932 – 1989)

- A great surgeon

LIST OF PAPERS

I. Hellum C , Johnsen LG, , Storheim K, Nygaard ØP, Brox JI, Rossvoll I, Rø M, Andresen H, Lydersen S, Grundnes O, Pedersen M, , Fritzell P and The Norwegian Spine Study Group*
Surgery with disc prosthesis versus rehabilitation in patients with low back pain and degenerative disc: Two year follow-up of randomized study
BMJ. 2011 May 19;342:d2786.

II. Johnsen LG, Hellum C, Storheim K, Nygaard OP, Brox JI, Rossvoll I, Rø M, Andresen H, Lydersen S, Grundnes O, Pedersen M, Leivseth G, Olafsson G, Borgström F, Fritzell P; Norwegian Spine Study Group.
Cost-Effectiveness of Total Disc Replacement Versus Multidisciplinary Rehabilitation in Patients With Chronic Low Back Pain: A Norwegian Multicenter RCT.
Spine (Phila Pa 1976). 2014 Jan 1;39(1):23-32

III. Johnsen LG, Brinckman P, Hellum C, Rossvoll I, Leivseth G
Segmental mobility, disc height and patient reported outcome after surgery for degenerative disc disease - a prospective randomized trial comparing implantation of disc prostheses and multidisciplinary back rehabilitation
Bone Joint J. 2013 Jan;95-B(1):81-9.

IV. Johnsen LG, Hellum C, Nygaard ØP, Storheim K, Brox JI, Rossvoll I, Leivseth G, Grotle M
Comparison of the SF-6D, the EQ-5D and the Oswestry Disability Index in patients with chronic low back pain and degenerative disc disease
BMC Musculoskelet Disord. 2013 Apr 26;14:148.

ACRONYMS AND ABBREVIATIONS

CEAC	Cost effectiveness acceptability curve
CLBP	Chronic low back pain
DCRA	Distortion-compensated Roentgen analysis
DDD	Degenerative disc disease
FABQ	Fear Avoidance Believe Questionnaire
EQ-5D	Euro Qol 5 Dimensions
HRQoL	Health-related quality of life
HSCL	Hopkins Symptom Checklist
ICER	Incremental cost-effectiveness ratio
LBP	Low back pain
ODI	Oswestry Disability Index
MDR	Multi-disciplinary rehabilitation
MRI	Magnetic resonance imaging
PRO	Patient Related Outcome
QALY	Quality Adjusted Life Year
ROM	Range of motion
SF-6D	Short Form – 6 Dimensions
VAS	Visual Analogue Scale
WTP	Willingness to pay

DEFINITIONS AND KEY CONCEPTS

Total disc replacement (TDR): a standardised surgical technique where an intervertebral disc of the spinal column is removed and replaced with an artificial implant, a prosthesis**.**

Multi-disciplinary rehabilitation (MDR): a structured education programme (often with varying content) consisting of group-based training covering topics of anatomy, physiology (including pain), and illness, as well as exercises and training. The aim is to increase patient functioning and coping ability[1].

Patient-reported outcome (PRO): questionnaires filled out by patients to evaluate the outcome of a treatment. The questionnaire assesses a single underlying characteristic – a measurement of the health state of the patient that may or may not be of concern to the patient. The measurement characteristic is termed a 'construct' and the questionnaires used to collect them are called 'instruments'. When a questionnaire measures a single construct it can be said to be unidimensional.

Quality-adjusted life year (QALY): a method used to calculate how many extra months or years of life of a reasonable quality a person might gain as a result of treatment. The advantage of using QALYs to measure the effect of treatment(s) is that it allows comparisons across disease areas. One disadvantage is that the use of generic instruments to measure quality of life as opposed to disease-specific instruments could be inaccurate and less responsive to a change in health state (e.g., after treatment).

Cost-utility analysis (CUA): a method of combining economic and clinical effect (outcome) to evaluate treatment of a disease. The cost of the treatment (hospital costs, follow-up costs, social expenditures, etc.) is combined with PROs in the form of a generic quality of life questionnaire, which is then transformed to give a scale value from 0 (dead) to 1 (excellent health) and which, combined with time, gives the unit 'QALY'.

Distortion-compensated Roentgen analysis (DCRA): a method for evaluating sagittal plane motion, translation, and disc height in segments of the spinal column from plain lateral radiographic views. The method compensates for distortion in central projection, off-centre position, axial rotation, and lateral tilt of the spine. The method also comprises a database of disc height, vertebral height, and sagittal plane displacement from lateral radiographic views of the lumbar spine, valid for male and female subjects in the age range of 16–57 years and used as a norm reference in the assessment of sagittal motion and disc height.

SUMMARY

The subject of this thesis is a comparative study of two treatment options for a selected group of patients with degenerative disc disease and chronic low back pain. Chronic low back pain represents a significant health problem. Globally, it is estimated that 10% of all years with functional impairment due to non-fatal disease is caused by low back pain. Many types of treatment have been proposed but there is still great disagreement about what should be the standard treatment. Several surgical treatment options exist. Disc arthroplasty with removal of the intervertebral disc and the insertion of a disc prosthesis aims to preserve the mobility of the affected degenerative segments. This is different to standard surgical treatment, which is fusion of the affected segments. Comparative studies of arthroplasty and fusion surgery have shown approximately even results. Other studies have compared fusion surgery against non-surgical treatment. One of these studies showed the best effect with surgery while others showed no difference. Our study is the first to compare disc replacement surgery with a non-surgical treatment option.

The setting for the study was a randomised controlled multicentre trial including 173 patients recruited from outpatient back clinics at all five university hospitals in Norway. Patients were randomised either to surgery with the insertion of a disc prosthesis or to a non-surgical treatment option in the form of (multi-disciplinary) spinal rehabilitation (MDR). Follow-up time after primary treatment was 2 years, with control after 6 weeks and 3, 6, 12, and 24 months. A variety of clinical outcome measures were recorded at the controls with the Oswestry Disability Index (ODI) as the primary endpoint in a clinical and in a biomechanical study and EuroQol 5D (EQ-5D) and Short Form-6D (SF-6D) as the effect variables in a health economic study. Two years after the index treatment, all patients were, in addition, examined and evaluated by two independent observers blinded to treatment type.
In the first study, we found that surgery with the insertion of a disc prosthesis after 2 years resulted in a statistically significantly better outcome compared with the non-surgical treatment in most clinical outcome measures including the primary efficacy variable. However, this difference could not be said to be *clinically* significant.

In the second study, we looked at the health economic consequences of choosing disc arthroplasty over a non-surgical treatment option. We found that it is uncertain if total disc replacement (TDR) could be a cost effective alternative to MDR. When the EQ-5D was used, TDR was cost effective, but when the SF-6D was used, it was not.

The main purpose of inserting a disc prosthesis is to preserve movement of the segments. This was the subject of a third study. We found no significant difference in segmental motion in the operated segments compared with segments at the corresponding level in the non -operated group. Furthermore, we found no relationship between segmental movement and disc height after the insertion of a disc prosthesis and clinical outcomes.

In the fourth study, we investigated some of the key psychometric characteristics of the EQ-5D and the SF-6D used as efficacy measurements in the health economic study. We found that there were significant differences between these indices in terms of ability to detect change after treatment and that, even though they measure the same construct along the same (overall quality of life) numeric scale, they measure different aspects of this property.

We conclude that surgery with disc prosthesis and non-surgical treatment in the form of MDR provide significant improvement in quality of life after 2 years in selected patients with chronic low back pain and degenerative disc disease. The improvement was greatest in the surgical group. Although *statistically* significant, the difference between treatment groups could not be said to be *clinically* significant. At the same time, we showed that there were probably other characteristics of the prosthesis than segmental mobility that contributed to this (significant) improvement in quality of life after 2 years. Moreover, it is uncertain whether disc replacement surgery would be a cost-effective alternative to non-surgical treatment, as cost effectiveness in this study was highly dependent on the quality of life index used as an outcome measure. In an analysis of some key psychometric properties of these indices, we conclude that there is a big difference between them, that they are not interchangeable, and that the choice of index in cost-effectiveness studies, in general, probably should be related to diagnosis and/or treatment type.

1 INTRODUCTION

1.1 Epidemiology

The lifetime incidence for an episode of low back pain (LBP) has been estimated to be about 80% and the prevalence of *chronic* low back pain (CLBP) is about 23%[2-4]. In a systematic review of the literature by Hoy et al[5], estimates of the point prevalence of LBP ranged from 1.0% to 58.1% (mean: 18.1%; median: 15.0%), and 1-year prevalence from 0.8% to 82.5% (mean: 38.1%; median: 37.4%)[5]. The Global Burden of Disease report from 2010 states that LBP constitutes 10.7% of total global years lived with disability (YLD) and it is ranked as the 9th of the 50 most common global sequelae of diseases[6].

In Norway, the point prevalence has been reported to be 13.4%, the 1-year prevalence 40.5%, and the lifetime prevalence 60.7%[7]. Furthermore, back pain was ranked as the most common type of health problem among individuals of 15–74 years of age with work-related health problems in Norway in 2007 (Statistics Norway 2011). The prevalence of back pain-related problems reported here was 27.1%, with pain in the shoulder and neck in second place with 19.4%. After 2 years' absence from work because of LBP, the likelihood of returning to ordinary work is less than 1% which, in turn, has a great influence on the size of social security payments[8].

Walsh[9] found a rise in low back disability between the ages 20 to 29 and 30 to 39 and then it remained constant up to age 50 to 59. Several authors have reported an increase in LBP prevalence[10-13]. It has been suggested that part of this increase was a result of the transition from an agrarian to an industrial society that took place at the end of the 19th century and Waddell[11,14] states in a paper from 1987 that "…low back-disability as opposed to pain is a relatively recent Western epidemic". The combination of modern diagnostic modalities and improved social support that allows for absence from work without catastrophic economic consequences could at least partly explain this increase[4,14,15]. Harkness et al[16] mentioned increasing rates of psychological distress and increased awareness of certain pain syndromes, not only by patients but also by health professionals as possible explanations.

1.2 Degenerative Disc Disease

1.2.1 The normal disc

The main function of the intervertebral disc (IVD) is to act as a shock absorber and to maintain limited mobility of the spine[17,18]. Three morphologically distinct structures can be identified[18]: a) a thick outer ring of fibrous cartilage termed the annulus fibrosus (AF), organised in a series of 15–25 concentric rings or lamellae[19]; b) a gelatinous core termed the nucleus pulposus (NP), consisting of randomly organised collagen fibres and elastin fibres surrounded by the AF[20,21]; and c) the vertebral end plates at the top and bottom of the vertebrae, inferior and superior to the AF and NP. The end plates consist of a thin horizontal layer, usually less than 1 mm thick, of hyaline cartilage[18,22]. In the healthy IVD, the end plate is usually an avascular and aneural structure[18,23,24].

1.2.2 The degenerative disc

A central concept in understanding disc degeneration is nutrition for the disc[22-25]. There is strong evidence that a fall in nutrient supply is associated with disc degeneration[23]. Discs receive most of their nutrients by diffusion through pores in the vertebral end plate from capillaries at the margin of the end plate[18,23,26-28]. Steep differences in concentrations of glucose, oxygen, and lactic acid provide an exchange of nutrients and waste metabolites into and out of the matrix of the cells in the NP across these pores[23]. It has been documented that cyclic mechanical stimuli in the form of compression and decompression assist in the exchange of large soluble factors across the IVD and its surrounding circulation and apply direct and indirect stimulus to disc cells[29]. Proteoglycans are huge water-binding molecules in the extracellular matrix of the disc. Aggrecan, a highly anionic glycosaminoglycan and a major proteoglycan especially in the NP, is responsible for maintaining tissue hydration through osmotic pressure[30]. Because of its structural alignment, resistance of compression of the disc is essential[31]. The hydrophilic properties of proteoglycans cause the NP to swell which, in turn, increases the resistance to compressive forces[32]. During growth and with increasing age, obliteration of the pores of the end plate by calcification and/or diminished blood supply to the IVD causes diminished nutrition flow to matrix cells [23]. This initiates tissue breakdown[23,33]. There is a fine balance between synthesis, breakdown, and accumulation of matrix macromolecules[18]. This delicate balance is influenced by proteinases and other enzymes[34,35]. Much research of recent years has been done

to clarify the role of such enzymes, especially the metalloproteinases. It is now commonly accepted that their proteolytic action in degradation, especially of aggrecan, plays a central role in DDD throughout life[30,36-38]. The degradation of proteoglycans causes damage of major structural components of the IVD[38,39]. In vitro studies have shown that metalloproteinases may be produced by cells of the discs themselves as well as by cells of the invading blood vessels; the invasion takes place as a part of the degenerative process[18]. The loss of matrix production of proteoglycans including aggrecan and, at the same time, an increase in matrix degradation lead to loss of the water-binding ability of the proteoglycans, which causes the nucleus to be less hydrated[23,24,40]. When fluid pressure within the disc falls, the disc starts bulging radially as a result[41]. The regularity of the annular lamellae is also compromised and the degenerate disc becomes increasingly cracked and fissured[42]. The disarrangement of the cartilaginous tissue structure of the disc eventually leads to loss of both disc height and biomechanical properties such as shock absorption and flexibility in movement[17,18]. The structural failure associated with degenerative changes may also cause spinal instability, which has been considered as one of the significant causes for mechanical LBP[43].

Adams proposed a definition of disc degeneration as "…an aberrant cell-mediated response to progressive structural failure"[54]. As possible causes for this structural failure of the IVD, he mentioned genetic inheritance, age, inadequate metabolite transport, and loading history.

1.2.3 Genetic factors for DDD

Several studies have documented the family aggregation of lumbar disc disease[44-48]. Heredity and linkage studies have demonstrated the correlation between genetics and IVD pathology. From the presence of family aggregation of DDD follows the possibility of the influence of a genetic component[49]. A differentiation between genetic factors and social-behaviour factors is, however, necessary. This has been achieved to a certain degree through twin studies. In a Finnish study based on in-depth interviews of twins and magnetic resonance imaging (MRI) scans, Battie et al[50] were able to show that 77% of the variability in disc degeneration observation scores could be explained by family aggregation. However, the authors stressed the fact that such studies could not separate genetic, anthropometric, and metabolic factors and the effect of shared early environment and lifestyle influences. In a study of twins unselected for back pain, Sambrook et al[51] were able to show that overall heritability for lumbar disc degeneration was 74%. In a UK twin study, Livshits et al[52] showed that one of the main risk

factors for reported episodes of severe and disabling LBP was genetic heritability. McGregor et al[53] reported a significant genetic effect on LBP, with estimates of heritability ranging from 52% to 57%.

Genetic studies on DDD have focused on the genes that code for functioning molecules in the disc[49]. These include aggrecan, degrading enzymes such as metalloproteinase II, and signalling molecules such as Interleukin I (IL-1), known to stimulate nerve endings on nociceptive nerve fibres[49,54,55]. An example on how genetic research may be useful in the treatment of DDD is the work of Sudo et al[56]. They showed how knowledge of the regulatory mechanism of the molecular response of NP cells to nutrient deprivation might reveal a new strategy for treating disc degeneration. It is important to remember the fact that it is not a single gene but interaction between genes that contributes to DDD[57].

Because genetic factors may interact with environmental factors, several authors emphasised the point that studies on genetic factors should include analysis of the interaction between genetic, behavioural, and environmental factors[49,58,59]. Battie[60], one of the authors of many Finnish twin studies, concluded in a relatively recent study that genetic and environmental influences on disc degeneration seemed to be of similar importance.

1.3 LBP and DDD

1.3.1 Possible pathophysiological mechanism for LBP from DDD

The sensory pathways of the IVD follow a dual pattern[61]. One route enters the adjacent dorsal root segmentally, whereas the other supply is non-segmental and ascends through the paravertebral sympathetic chain with re-entry through the thoracolumbar white rami communicantes. In the healthy IVD, only the outer third of the AF is innervated. Coppes[62] described a more extensive disc innervation in the severely degenerated human lumbar disc compared with the normal discs and it has been postulated that this neural ingrowth into the IVD is an important factor in discogenic LBP[63,64]. Freemont et al[65] were able to show that nociceptive nerve fibres grew into the inner third of the annulus. Some of these fibres contained neuropeptides, which are associated with nociception[62,64,66,67]. Degenerative discs are known to produce high levels of pro-inflammatory mediators like interleukin-6 (IL-6), interleukin-8 (IL-8), and prostaglandin E2 (PGE2)[68]. These inflammatory and pro- inflammatory mediators from the

diseased degenerative tissue are thought to sensitise nerve endings of the nociceptive nerve fibres, and the inflammatory response is thought by many to be the main pathophysiological cause of CLBP from DDD[64,68-72].

During the degenerative process and break down of disc structure, the IVD segment becomes unstable [43]. The excessive motion following this process is thought to cause pain because of the stretching and compressing of structures like ligaments, joint capsules, annular fibres, or end plates, which are known to have a significant number of nociceptors[43].

It is now generally agreed that degenerated IVDs are a major tissue source in CLBP[73-75]. However, the phenomenon of central sensitisation has to be taken into consideration[74]. Nociceptive stimuli from the degenerated IVD are transmitted via the spinothalamic tract to the cerebral cortex and C-fibres fire repetitively to the dorsal horn[76]. Over time, this constant firing of neurons causes increased excitability in cell membranes of the central nervous system[77]. This generates pain hypersensitivity and the pain is no longer coupled solely to the peripheral nociceptive tissue source, the IVD in this case, but also to the hyper excitable neuron cells of the central nervous system[77].

1.3.2 The biopsychosocial model

Waddell[50,78] proposed a biopsychosocial model of LBP. In this model, pain is basically thought of as physiological. However, according to Waddell's model, the whole process of experiencing chronic pain may be modified by psychological factors such as the patient's personality and pre-existing psychological state. Factors like social environment, illness behaviour, psychological distress, attitudes, and beliefs can, in this model, at least partly explain the patient's current level of pain and disability. This can modulate the process of central sensitisation. Other authors have later reported on emotional or cognitive regulation of pain, confirming the biopsychosocial model[79,80]. The model is important for the rationale of treatment of LBP with a cognitive approach.

1.3.3 Environmental factors for LBP

LBP affects between 14% and 80% of working age people, depending on case definition[81]. Specific occupational physical activities that have been associated with LBP include heavy manual work, lifting and twisting, postural stress, and whole body vibration[50].

Occupational activities like bending/twisting, awkward postures, sitting, standing/walking, carrying, pushing/pulling, lifting and manual handling/assisting patients have an uncertain strength of relationship to LBP[81-85]. Harkness[86], in a study of risk factors for new-onset LBP amongst newly employed workers, concluded that several aspects of the workplace environment, other than mechanical factors, were important in predicting new-onset LBP.

Jørgensen et al[84] discussed some of the main critiques of existing studies on the relationship between physical work demands and musculoskeletal pain. They mentioned potentially confounding factors like individual and socioeconomic factors in addition to the predominant use of self-reported measurements of physical work demands, which have been shown to have poor validity.

1.3.4 Individual risk factors for LBP

LBP increases with age but the dose-response relationship between age and LBP is not linear, suggesting that multiple factors are involved[87,88]. Causes of severe back disorders have been found to be clustered around a subject's socioeconomic status, indicated by formal education[89,90]. Low job control or satisfaction can increase the risk of hospitalisation for back disorders[50,89]. Smoking was found to be associated with LBP, but the results could be difficult to interpret because of linkage to social class, education, and occupation[50,91,92]. Obesity and being overweight increase the risk of LBP[93]. An association of atherosclerosis with LBP and the degree of disc degeneration was found in some studies[94-97]. Studies of Linton[98,99] indicated that psychological factors like psychological distress (odds ratio=13.2) and poor function (odds ratio=6.4) were associated with a greater risk of developing back pain than perceived workload, gender, and foreign birth.

In conclusion, studies have shown a correlation between LBP and DDD, although the strength of this association remains unclear [81,100-102]. Studies that have attempted to identify possible risk factors for LBP found genetic, vascular, work-related, and lifestyle-related causes. At present, the common explanation is that the cause of the problem of CLBP is multifactorial (Fig. 1) and that evaluation of patients with DDD and LBP should aim at identifying underlying psychosocial factors as well as biological factors [57,74,103-105].

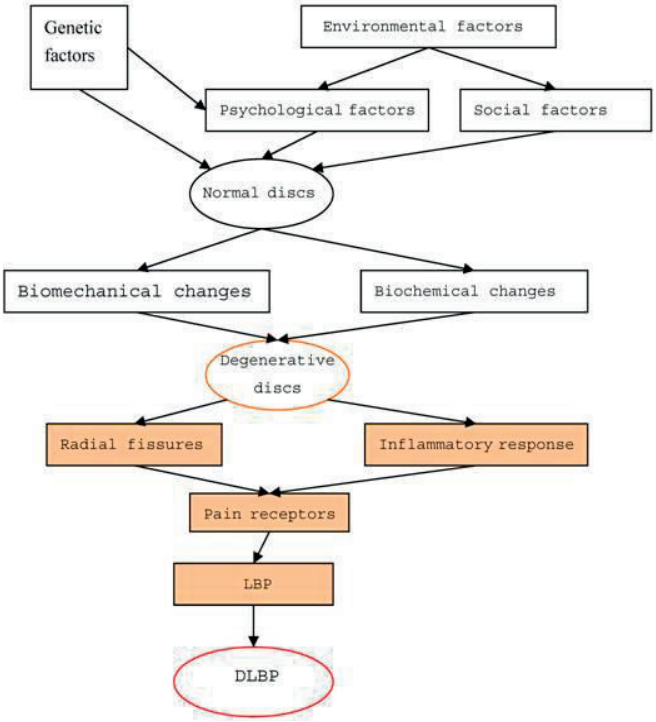

Figure 1. Possible pathogenesis of discogenic low back pain. Adapted from Zhang YG, Guo TM, Guo X, Wu SX. Clinical diagnosis for discogenic low back pain. International journal of biological sciences 2009;5:647-58.

1.4 Classification and clinical presentation

It is common to distinguish between specific and nonspecific back pain. Approximately 15% of LBP incidences can be related to a specific cause such as fracture, malignant disease, or rheumatic disorders[106]. The Norwegian guidelines for the diagnosis and treatment of LBP[1] recommend an internationally recognised tripartite division of clinical symptoms: a) nonspecific LBP; b) LBP with affection of nerve root/nerve root pain; and c) conditions of LBP with possibility of malignant disease or cauda equina affection. The concept of 'flags' was introduced to categorise seriousness of the disease and prognostic aspects[50]. Red flags indicate that there might be spinal pathology or referred pain to the spinal column, which should prompt the clinician to take action immediately. Yellow flags indicate that there might be psychosocial causes for LBP and that there is a risk of the development of chronicity. Green flags are factors that indicate a good prognosis for rapid spontaneous recovery[1,11]. The duration of LBP is usually divided into acute (0–4 weeks), sub-acute (4–12 weeks), and chronic (more than 12 weeks)[107,108]. The term "long lasting" is preferred over "chronic" by some authors[1] focusing on the dynamic aspect of treatment.

1.5 Imaging

Degenerative changes in the spine can be seen in different X-ray modalities like Computed Tomography (CT) and plain radiographs. The morphological appearance and biochemical matrix composition can be visualised on MRI and several radiological classification systems for DDD have been described[109,110]. Degenerative changes of the IVD in patients eligible for our study were classified based on four MRI findings:

1.5.1 Disc height reduction

Some authors reported a correlation between disc height and LBP[111-114]. Several methods on how to measure this have been proposed, using different radiological modalities. Andersson[115] noted that accurate measurements could not be obtained from routine roentgenographs. Raininko[116] reported fair to excellent intra-observer agreement on MRI scans. Frobin[117,118] compared the height of lumbar discs measured from radiographs with the height classified from MRI images in a cross-sectional study using the DCRA method when assessing radiographs. He found that loss of disc height on MRI images was compatible with radiographs on average,

although imprecise in the assessment of individual discs[117]. In conclusion, disc height is a common finding in DDD but shows a weak-to-moderate association with LBP[119].

1.5.2 Modic changes

In 1988, Modic et al[120] reviewed 474 patients referred for lumbar spine MRI. On the basis of this study, they proposed a classification of degenerative changes in the lumbar vertebral bone marrow. In comparison with histopathological findings, they observed that the signal intensity changes appeared to reflect a spectrum of vertebral body marrow changes associated with DDD. Their classification has become the most common system for the classification of degenerative changes in the lumbar spine.

Table 1. From Modic MT, Steinberg PM, Ross JS, Masaryk TJ, Carter JR. Degenerative disk disease: assessment of changes in vertebral body marrow with MR imaging. Radiology 1988;166:193-9.

Modic type	MRI findings	Histopathological findings
Type 1	Reduced signal intensity on T1-weighted spin-echo images and increased signal intensity on T2-weigthed images	Disruption and fissuring of the end plates, vascularised fibrous tissue
Type 2	Increased signal intensity on T1-weighted images and isointense or slightly increased signal intensity on T2-weighted images	Yellow marrow (fat) replacement of red bone marrow
Type 3	Reduced signal activity on both T1- and T2-weighted images	Osteosclerosis

Later, some works have tried to assess the correlation between clinical findings, i.e., LBP and Modic changes on MRI. Braithwaite et al[121] evaluated pain correlation between lumbar

discography and Modic changes. They found that Modic changes appeared to be a relatively specific but insensitive sign of a painful lumbar disc in patients with LBP. Luoma et al[122] found an increased risk of LBP (including all types) in relation to all signs of disc degeneration. Kjaer et al[123] found that patients with degenerative discs and at the same time Modic changes in the vertebra had a distinct clinical profile different from that of patients with degenerative disc only. They concluded that LBP and Modic changes were strongly correlated and that people with both DDD and Modic changes might deserve to be diagnosed as having *specific* LBP. Cheung[124] concluded that disc degeneration was strongly associated with back pain in a dose-related manner based on MRI findings.

1.5.3 High-intensity zone

Disruption of the inner AF in the form of radial fissures as a result of the degenerative process can be visualised on MRI scans. Because of their appearance on MRI, such degenerative changes are called high-intensity zone(HIZ) [125,126]. Such changes have been found to correlate with LBP in patients with DDD and it has been suggested that the outer annular disruption is painful[124-128]. In a study published in 1991, Aprill and Bogduk[125] assessed the prevalence, reliability, and validity of this sign in 500 patients undergoing MRI for back pain. HIZ occurred in 28% of the patients with back pain and the positive predictive value was 86%. Later, several authors reported on the correlation between lumbar discography, pain reproduction, and HIZ findings on MRI[126,128-130]. While Lam et al[129] found that the sensitivity, specificity, and positive predictive value for pain reproduction were high (81%, 79%, and 87%, respectively), Carragee[131,132], in a study of 2000 patients, concluded that although the prevalence of HIZ was slightly higher in symptomatic patients, the prevalence in asymptomatic individuals with DDD (25%) was too high for meaningful clinical use of this sign.

In conclusion, HIZ on MRI occur frequently in patients with LBP. The presence of HIZ is thought to be an indicator for LBP with a high sensitivity and low specificity[69].

1.5.4 Morphological changes in the disc

Morphological changes in the disc can be assessed by signal intensity of the NP in addition to anterior and posterior bulge of the IVD, and have been associated with nonspecific LBP[133,134]. The decrease in proteoglycan content and subsequent loss of water content result in reduced

signal intensity on T2-weighted MRI[135,136]. Relative signal intensity of the NP in the relation to the signal intensity of the cerebrospinal fluid can be calculated to assess the degree of disc degeneration[137]. Luoma[134] classified the signal intensity as: 1=Bright; 2=Grey; 3=Dark; and 4=Black, with an inter-observer agreement (weighted kappa) in the range of 0.59–0.83 and an intra-observer agreement rate of 57% to 81%.

In summary, although MRI can act as a tool for basic research into disc physiology and the aetiology of disc degeneration, studies on MRI have only been able to establish a weak correlation between progressive disc degeneration and LBP development[24]. Degenerative changes have been found in symptomatic as well as asymptomatic people[119,133,138,139]. In our study, we assessed the reliability of MRI findings in candidates for TDR[140]. Inter-observer agreement here was generally moderate or good.

1.6 Treatment

1.6.1 Surgery for LBP

Hibbs and Albee [141] (in 1911) and Chandler [142] (in 1929) were the first to report a surgical technique with the intention to produce a fusion between the posterior aspects of the spinal vertebrae. The indication for surgery at that time was Pott's disease (tuberculosis of the spine). Howorth[143] (in 1937) was the first to use fusion surgery in conjunction with a ruptured NP. Barr[142,143] (in 1947) wrote that the unpredictable results of LBP after disc excision operation was the result of underlying structural weakness of the disc and recommended spinal fusion together with disc excision, the "combined operation"[144,145].

During the 70s and 80s, fusion surgery gained increasing recognition and, between 1996 and 2001, the annual number of spinal fusion operations in the United States rose by 77%[146]. Implants for spinal fusion with pedicle screws became almost supreme in the market. In 1998, the annual rate for spinal lumbar fusion in the United States reached 77 628 which increased to 210 407 in 2008[147]. Deyo[148] mentioned as possible explanations for this changes in the population, technological advances, uncertainty regarding indications, as well as the financial incentives for surgeons, hospitals, and the device industry, which may have had synergistic effects.

Several randomised studies in recent years have shown that fusion surgery has a certain effect on the patients' pain and function[149] and it has been compared with non-operative treatment. In a Swedish study of 294 patients, Fritzell et al[150] concluded that fusion surgery compared significantly better than care as usual (mainly physical therapy) with respect to health-related quality of life (HRQoL; assessed by the ODI), pain, and net back to work. Brox et al[151,152], in two randomised controlled trials, found no significant difference in ODI between fusion surgery and a multi-disciplinary treatment regime similar to that used in our study. In an English randomised controlled trial from 2005, no significant difference in ODI between fusion surgery and an intensive rehabilitation programme similar to Brox was found[153].

Fusion surgery requires healing and stabilising of the spinal musculature postoperatively as well as healing of bone and there is a certain percentage of non-union [154].

TDR surgery – 'the long quest for mobility'

Arthrodesis – the process of surgical fusion of a joint – is generally not considered an optimal solution, due to the increased stresses and subsequent degeneration in the adjacent joints[155]. This fact led to the introduction of arthroplasty in the hip as an alternative to arthrodesis early in the 60s, which has had a great impact on quality of life for patients with coxarthrosis [155-158].

In the spine, it has been claimed that the fusion of segments leads to biomechanical changes that bring about increased degeneration by superimposing stress on neighbouring segments, a phenomenon called adjacent level disease (ALD)[159-161]. Alteration of the biomechanics in the adjacent segments has been demonstrated by several authors [162,163]. It has been claimed that this occurs in addition to the painful excessive motion of the degenerate IVD mentioned before[43].

The incidence of symptomatic ALD was reported to be in the range of 5.2 to 18.5%, while the incidence of radiographic ALD was in the range of 8 to 100%, suggesting that it is difficult to rule out the effect of the age-related natural course of DDD[54,160].

In 1966, the Swedish spinal surgeon Fernström[164] presented 191 patients in which he had implanted steel balls as a replacement for degenerated discs in order to preserve motion. Patients were operated for up to eight segments at a time, but the clinical results were poor. The Fernström steel ball illustrated two important aspects of disc prostheses. First, 88% of the

patients developed subsidence after 4 to 7 years, a phenomenon in which the prosthesis, due to mechanical wear and tear, drops down into the end plates of the vertebra and eventually into the vertebral body[142,165,166]. Second, adverse biomechanical conditions with instability were encountered in the operated spine. A ball that rests against an end plate will have virtually infinite degrees of freedom and movement will only be limited by soft tissue. This is non-physiological and the problem is amplified the more the number of segments operated. The failure of the Fernström ball prosthesis brought discredit to the concept of implantation of a mobile device after removing the degenerative disc among spinal surgeons. However, the idea of a motion-preserving device was not totally left.

Some devices were patented but did not become a commercial success (Table 2, Fig. 2). Since 1973, there has been an almost yearly acquisition of a new disc replacement patent, of which only a small number have reached clinical use[142,167]. In the late 80s and early 90s, the French spine surgeon Marnay[168] began to implant a new type of disc prosthesis with three components, called Pro Disk 1. At about the same time, Bütner-Janz et al[169] in eastern Germany began to implant the Charité prosthesis on a commercial basis. Later, the Maverick prosthesis was developed. These are the three most common types of lumbar disc prosthesis today. In the lumbar spine, the rationale for introducing a motion-sparing device became the avoidance of the junctional degeneration seen after arthrodesis (i.e., fusion surgery), by the preservation of segmental motion[170].

Table 2. Development of disc prostheses and disc substitutes. From Errico TJ. Lumbar disc arthroplasty. Clin Orthop Relat Res 2005:106-17.

Year	Inventor(s)	Device
1955	Cleveland, Hamby, and Glaser	Acryl substance in the disc space
1966	Fernström	Steel ball after removing the disc
1974	Froming	Fluid-filled elastic chamber sandwiched between two metal cup end plates
1975	Stubstad	Dacron mesh containing a silicon disc
1978	Fassio	Silicon prosthesis
1980s	Heller	Posteriorly hinged metal prosthesis with interposed titanium springs
1980s	Steffee	Polyolefin rubber contained between two titanium plates (the Acroflex disc)
1989	Bütner-Janz	Charité prosthesis. Metal on plastic
1989	Marnay	ProDisc I prosthesis. Metal on plastic
2002	Mathews et al	Maverick. Metal on metal

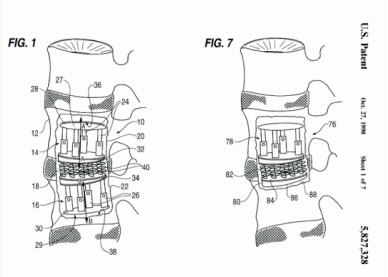

Figure 2. The Buttermann intervertebral prosthetic device from 1998. Patented but never implanted in humans.

Thus, the theoretical advantages of modern disc arthroplasty are the preservation of motion (hence the expression 'motion-preserving device'), avoidance of trauma of the back muscles (the procedure takes place transabdominally), and the lack of need for bone healing[171,172]. A biomechanical classification system for modern TDRs has been proposed (Table 3)[171].

Table 3. Classification of biomechanical properties of modern lumbar intervertebral prostheses. Adapted from Errico TJ. Lumbar disc arthroplasty. Clin Orthop Relat Res 2005:106-17.

1	Constrained	The prosthesis has mechanical limitations in the physiological ROM
2	Semi-constrained	The prosthesis has mechanical limitations in certain directions of motion but may have freedom of movement beyond the physiological ROM in other directions
3	Unconstrained	The prosthesis is mobile beyond the physiological ROM in all directions

To prevent subsidence, modern prostheses have a 'footprint', a contact surface against the end plate that distributes the axial pressure to avoid subsidence, in addition to restrictions on movement in one or more directions [171]. Although these motion-sparing devices have different designs, two crucial components seem to be common in the expected theoretical mechanism of pain relief after TDR[142]: (1) complete excision of the nucleus and (2) the restoration or improvement of normal intervertebral biomechanics[171].

1.6.2 Complications in TDR surgery

The complication and reoperation rates after implanting a lumbar disc prosthesis are about the same magnitude as those reported for fusion surgery[173]. Common types of complications in the two surgical methods include postoperative wound infection, sepsis, and subsidence of implant. Specific complications for the anterior access to the spine in order to reach the lower segments of the spine are perioperative vascular damage, perioperative intestinal damage, postoperative ileus, postoperative retrograde ejaculation, and complications related to the implant[174,175]. The

risk of injury to the great vessels and retroperitoneal structures is greater during revision than primary procedures[176].

1.6.3 Outcome after TDR

Outcomes after TDR are summarised in a Cochrane paper from 2012[177]. The study included 40 publications including seven randomised controlled trials. Although statistically significant, the difference between TDR and standard fusion surgery was not beyond generally accepted clinically important difference. However, there are indications for less radiological degeneration of the adjacent levels with disc prosthesis compared with fusion surgery[178-180] although this is controversial[181].

1.6.4 Non-surgical treatment

A variety of non-surgical treatments are described in the literature. Ostelo et al[182], in a Cochrane review from 2005, referred to three behavioural treatment approaches: operative, cognitive, and respondent. Middelkoop et al[103] reviewed the effectiveness of physical and rehabilitation interventions including exercise therapy, back school, transcutaneous electrical nerve stimulation (TENS), low-level laser therapy, education, massage, behavioural treatment, traction, multi-disciplinary treatment, lumbar supports, and heat/cold therapy.

1.6.5 Outcome after non-surgical treatment

In the review by Middelkoop[103], MDR was found to be more effective in reducing pain than no treatment and patients receiving behavioural therapy had reduced pain compared with waiting list controls. However, none of the studies in this review reached a difference that was defined as clinically important and the evidence for effectiveness was low.

The Cochrane review by Ostelo et al[182] concluded that adding behavioural components to usual treatment programmes for CLBP (i.e., physiotherapy, back education, or various forms of medical treatment) had no significant effect on pain relief either in the short-term or long-term. However, patients receiving combined respondent-cognitive therapy and progressive relaxation therapy had better short-term pain relief than waiting list controls[182].

In a meta-analysis of psychological interventions for CLBP, Hoffman et al[183] concluded that multi-disciplinary approaches with a psychological component could have positive short-term effects on pain interference and positive long-term effects on return to work compared with active control conditions.

In a randomised controlled trial, Brox et al[151] showed equal improvement in back pain, use of analgesics, emotional distress, and life satisfaction after treatment with cognitive intervention and exercises compared with fusion surgery. In a later study, they reported that lumbar fusion failed to show any benefit over cognitive intervention and exercises in patients with CLBP after previous surgery for disc herniation[152]. The non-surgical treatment used in these two studies resembled the one used in our study[184].

Overall, a moderate effect of multi-disciplinary treatment is reported in the literature[103,182,185]. Several authors of review studies reported that firm conclusions about the effectiveness of different non-surgical treatment options for patients with LBP were hard to draw due to the heterogeneity of the populations, interventions, and comparison groups[103,183,186]. No complications have been described for non-surgical treatment of LBP.

1.7 Health economy and LBP

1.7.1 Health economic consequences of LBP

From the 1890s when LBP was first reported in the context of compensation until today, there has been a dramatic increase in spending on sick leave and disability pension due to back pain [14]. Frymoyer[187] reported in 1991 that although low back disorders were extremely prevalent in all societies, it was the rate of disability that had increased and not the frequency of LBP per se for the last decades. This was later confirmed by Norwegian authors[188].

The direct and indirect costs associated with LBP in Norway was between 13 and 15 billion Norwegian kroner in 2008[188] and patients with CLBP are known to have a higher consumption of health services than most other groups of patients[189,190]. Gore et al[191] compared comorbidities, pain-related pharmacotherapy, and health care service use/costs (pharmacy, outpatient, inpatient, and total) between patients with CLBP and patients without a CLBP diagnosis in a cost of illness study of 101 294 patients from a life insurance database. They found that patients with

the CLBP diagnosis had a significantly higher level of health resource utilisation and health care costs.

Back problems were ranked sixth of the 15 most costly conditions in America, with national costs of $12.2 billion[192]. In the same study, back problems were ranked fourth of the 20 most costly health conditions for employers.

1.7.2 Cost-effectiveness studies

The following definitions are from the textbook of Drummond et al[193]:

1. Economic evaluation always involves a comparative analysis of alternative causes of action.

2. Analyses, in which costs are related to a single, common effect that may differ in magnitude between alternative treatment options, are usually referred to as cost-effectiveness analyses (CEAs).

The rationale behind CEAs is that the result of such analyses can act as a source of information for decision makers[194]. Two quantities are typically assessed: The additional costs of a new treatment compared with the existing alternative (or "standard treatment") and the additional health benefits [195]. Mean costs and mean effect for each treatment group are obtained and the difference in costs between the "new" and standard treatment are divided by the difference in effects to present what is called the Incremental Cost Effectiveness Ratio or ICER[194,196]:

$$\textbf{ICER} = C_t - C_c / E_t - E_c = \Delta C/\Delta E = \text{Costs pr unit of health gained}$$

The ICER reflects the cost per unit of health benefit obtained from switching from one intervention to another [196].

Variation in the numerator (incremental cost) and denominator (incremental effect) introduces sampling uncertainty [196,197] . Claxton[197] find it useful to distinguish between variability, heterogeneity and uncertainty in economical trials. While variability refers to the natural variation between patients in their response to treatment and the costs they incur, heterogeneity refers to differences between patients who have different characteristics. The uncertainty then refers to the fact that we can never know the true mean of costs and effects as it will vary from study population to study population.

The uncertainness of the ICER estimate can be presented in a cost-effectiveness plane [198] (fig 3). Here, the incremental effect is plotted on the X-axis and the incremental costs on the Y-axis. The slope of a ray from the origin to any cost-effect combination represents the cost-effectiveness ratio [194]. In our study we also make use of Cost Effectiveness Acceptability Curves (CEAC) to graphically visualize sampling uncertainty in relation to a willingness to pay limit. Such curves are constructed by calculating the probability that the estimated cost-effectiveness ratio falls below specified values of willingness to pay [199].

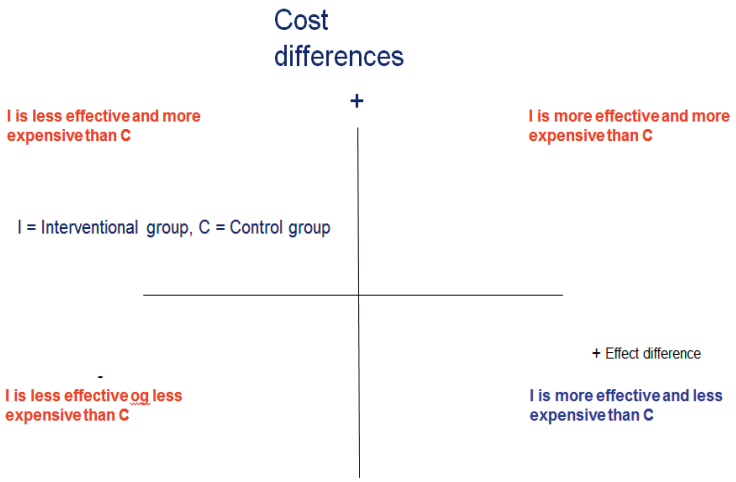

Figure 3. The cost-effectiveness plane.

As an expression for the effectiveness (or health gained), the outcome scores of general HRQoL questionnaires can be stratified into different health states [200,201]. Using different techniques (see later), a random sample of people from the normal population is asked to evaluate different health states[202,203]. Treatment benefit is thus expressed in a way that allows health states that are considered less preferable (0) to full health (1) to be given quantitative values. Because these quantitative values represent a valuation or preference of health states for the patients, they are called utility indexes (more utility for the patient with increasing value) or preference-based measures (some health states are preferred over others)[204]. When combined with a follow-up

period, health utility indexes are used to calculate QALYs[194]. Two of the most used indexes are the EQ-5D and the SF-6D[203,205,206]. Health economic studies that use QALYs as a measure of effectiveness are called cost-utility studies[193,207,208]. QALYs have the advantage of combining multiple dimensions of outcome (survival and quality of life) into a single measure that allows comparisons to be made across therapeutic areas and illnesses[194].

Various methods are used to measure health state preferences[193,194,200,209]. At least two techniques have been described [205,210]: the Time Trade Off (TTO) and the Standard Gamble (SG) methods. In the TTO method, the subject is offered two alternatives: either to live with a chronic disease/health state for a certain time *t* followed by death or to be healthy for a time $x < t$. Time x is then varied until the subject is indifferent between the two alternatives[193]. In the Standard Gamble (SG) method[210], the subject is offered two alternatives: alternative one is to live with the chronic disease until death; alternative two is to either return to perfect health and live for an additional *t* years with a probability *p* or immediate death with a probability of *1-p*. *p* is then varied until the subject is indifferent between the two alternatives[193]. While the TTO method was used to construct EQ-5D utilities, the SG method was used to construct the SF-6D utility index.[203,205,206].

1.7.3 Cost-effectiveness studies on the treatment of CLBP

Rivero-Arias et al[211] compared fusion surgery with an intensive programme of rehabilitation in patients with CLBP in a cost-utility study alongside a clinical randomised trial. They found that surgical stabilisation with fusion surgery might not be cost-effective. However, there was a question of selection bias in this study as patients were eligible if the clinician and patient were uncertain which of the study treatment strategies was best[153].

In 2004, Fritzell et al [212] reported that fusion was more expensive than non-specific rehabilitation (treatment as usual) after 2 years but that cost-effectiveness was dependent on the threshold of willingness to pay (WTP) set by the society.

Fritzell[213] also assessed the concept of TDR compared with instrumented lumbar fusion (FUS) in a cost-utility study alongside a randomised controlled trial. They found that TDR was significantly less costly from a healthcare perspective but, because of a non-difference in clinical outcome, the cost/QALY was not meaningful to calculate. In this study, it was not possible to state whether TDR or FUS was most cost effective after 2 years.

Schweikert et al concluded that the addition of cognitive-behavioural treatment to standard therapy compared with standard 3-week inpatient rehabilitation for patients with CLBP might be cost effective by reducing work days lost [214]. This finding was later confirmed[215]. Skouen et al concluded that the light multi-disciplinary treatment model was a cost-effective treatment for men with CLBP, using a form of MDR similar to that used in the present study [216].

In a RCT setting, Søgaard [217] et al compared simple behavioral extension of setting up group meetings for patients to a strict physiotherapeutic focus. They found that the behavioral model was cost effective over strict physiotherapy and that increasing frequency and guidance of a traditional physiotherapeutic regimen was unlikely to be cost-effective. Jensen et al [218] compared brief intervention to multidisciplinary intervention in a recent cost-effectiveness study. They found that the brief intervention resulted in fewer sick leave weeks and was less expensive than the multidisciplinary intervention. The multidisciplinary intervention only outperformed the brief intervention in terms of costs in a subgroup of sick-listed employees who thought they were at risk of losing their job or had little influence on their work situation.

Van der Roer[219] compared intensive group training with physiotherapy as usual in a cost-effectiveness study and found no difference in total costs and a non-significant difference in effect.

Van der Roer et al[186] also tried to assess the cost effectiveness of various interventions for LBP. Due to the heterogeneity of interventions, controls, and study populations, they concluded that more studies were needed before any conclusions could be made. In conclusion, the variable reporting quality of the few existing cost-utility studies of LBP makes direct comparison of the cost and effectiveness of different surgical and non-surgical treatment options difficult[103,186,220-222].

1.8 Biomechanical aspects

The intervertebral segment, acting as a unit, provides flexibility in movement by allowing bending, flexion, and torsion of the spinal column [18,42]. Fujiwara et al [54,223] found increasing segmental motion with an increasing grade of degenerative changes up to a certain degree and a certain decrease with further degeneration. Segmental instability is frequently considered a cause of LBP[43]. It is thought that the excessive motion of the degenerative spine beyond normal

constraints can cause the compression or stretching of structures known to have nociceptors like ligaments, joint capsules, annular fibres, or end plates. This has partly been the rationale behind fusion surgery[144,224].

A few studies have assessed the range of motion (ROM) in segments with implanted disc prosthesis. Huang et al[170] observed patients 7 to 10 years after implantation of the ProDisc II L prosthesis and found a segmental range of flexion-extension motion in the order of 3.5°. In contrast, Bertagnoli et al[212] and Tropiano et al [225,226] observed a ROM in the order of 10° less than 2 years after implantation of the ProDisc II prosthesis. Reviewing studies with findings up to 4 years after implantation of the SB Charité III prosthesis, de Kleuver et al[155] reported an average ROM of 5°–12°. In a noticeable percentage of cases, the ROM deteriorated with time and some arthroplasties eventually resulted in fusion. Differences in cohort, the length of follow-up, prosthesis model, and surgical technique might be responsible for these discrepancies. In addition, measurement protocols differ among studies. As segmental motion is small, and the range of sagittal plane motion is given by the difference of two angles to be measured in flexion and extension, inferior measurement precision could have confounded the results of some past studies.

In a review study early in the commercialised era of TDR (2002), de Kleuver[155] noted that few studies evaluated the mobility of the prosthesis and those who did failed to describe the method used to measure the motion properties of the device. In a Cochrane review paper 10 years later of 40 randomised controlled studies on disc prosthesis, Jacobs et al[177] noted that "The primary goal of prevention of adjacent level disease and facet joint degeneration by using total disc replacement, as noted by the manufacturers and distributors, was not properly assessed and not a research question at all". Huang et al[227] found a weak correlation between segmental ROM and clinical outcome.

1.9 Outcome measures

The ability to accurately assess severity and change in symptoms in a reliable and valid way is essential in clinical studies where outcomes reported by the patient are an issue[228,229]. Blazeby[208] defined PROs as "Outcomes that assess any aspect of a patient's health that come directly from the patient without the interpretation of the patient's responses by a physician or anyone else".

The use of validated outcome measures or patient-related outcome measure (PROM) is essential[224]. A consensus for the taxonomy of measurement properties relevant for evaluating such health instruments has been reached by the COnsensus based Standards for the selection of health Measurement Instruments (COSMIN) group[230-232].

A variety of both generic and disease-specific instruments for measuring clinical outcome in low back patients exist[233-236]. Grotle[229] reviewed 36 disease-specific outcome measures for back-specific outcome questionnaires and found that only a few of them could be considered acceptably validated.

The field of psychometrics is concerned with the construction and validation of measurement instruments that are used to measure clinical outcome[237]. Central concepts are Classical Test Theory (CTT) and Item Response Theory (IRT). CTT comprises a set of principles that seek to clarify how successful a questionnaire or instrument is in estimating clinical outcome[238]. IRT assumes that patients with a particular level of quality of life (or other traits) will have a certain probability of responding positively on questions (items) according to their level of ability (e.g., more or less pain)[239]. Central to IRT is whether an instrument constitutes a single composite scale, i.e., taps a single underlying construct. If a certain criterion is fulfilled then the questionnaire is said to be *unidimensional*; the underlying trait is measured along a continuous scale[240]. In our study of measurement properties of the ODI, EQ-5D, and SF-6D, we used both CTT and IRT.

2 AIMS

The aim of this randomized controlled study was to provide an improved foundation for the choice of treatment for patients with degenerative disc disease and chronic low back pain. More specifically, we wished to find out whether surgery with total disc replacement or multidisciplinary rehabilitation would provide the best clinical results and health economical efficacy two year after treatment start. In addition, we wished to evaluate some outcome measures used in health economic analysis of patients with low back pain.

2.1 Specific aims:

- To assess patient related outcome after treatment with TDR compared to MDR.
- To assess the cost-effectiveness of TDR compared to MDR.
- To assess the biomechanical properties of intervertebral disc prosthesis and its relation to clinical outcome.
- To assess the impact of using different utility measures for estimation of efficacy in cost utility studies.

3 METHODS

3.1 Design

This was a multicenter study conducted at five university hospitals in Norway. The study was designed as a randomized controlled trial (RCT)[241]. The trial was designed to have 80% power to detect a significant difference of at least 10 points in change in the mean Oswestry disability index score between the intervention groups at two year follow-up [152]. Baseline standard deviation was estimated at 18 [242]. Considering these assumptions and adding 25% for a multicenter study design and 30% for possible drop-outs, we estimated we required 180 patients. We included patients with low back pain and degenerative discs in the period between April 2004 and May 2007. All patients were treated within three months after randomization. Patients were randomized in blocks with a website hosted by the medical faculty. Allocation was concealed for all people involved in the trial. A coordinating secretary not involved in the treatment could access randomization details on the internet. The patient and the treating unit were informed about the allocation shortly after randomization. At each center (the five university hospitals), patients were stratified on whether they had had previous surgery (microsurgical decompression) or not (before randomization). Independent observers collected and entered data.

3.2 Participants

Patients from all health regions in Norway were included. Participants were recruited from patients submitted to the five university clinics for low back pain and included or excluded according to certain exclusion and inclusion criteria. No supplemental recruitment attempts were done. Mean age (SD) was 41.1 (7.1) years in the surgery group and 40.8 (7.1) years in the rehabilitation group. There were 40 women (47%) in the surgery group and 51 (59%) in the rehabilitation group. Mean (SD) duration of back pain (months) was 76 (72) in the surgery group and 85 (74) in the rehabilitation group. Most baseline characteristics were similar in the two treatment groups. Low back pain score and SF-36 mental health subscores, however, were significantly worse in the rehabilitation group than in the surgery group. Of the 605 patients screened for eligibility, 173 were included in the study and treated between April 2004 and

September 2007 (86 with surgery and 87 with rehabilitation). The drop-out rate from inclusion to two year follow-up was 15% (n=13) in the surgical arm and 24% (n=21) in the rehabilitation arm. Five patients (6%) crossed over from rehabilitation to surgery, but none crossed from surgery to rehabilitation.

In the health economic analysis, 144 of the 173 patients included in the clinical outcome study provided enough data on cost and resource use and were considered eligible for the health-economic analysis; 68 of these patients were in the rehabilitation group and 74 were in the surgery group.

In the DCRA study, 120 patients (74 in the prosthesis cohort, 46 in the intensive rehabilitation cohort) provided x-rays of acceptable quality and were included in the final analysis. The mean age was 42 (SD 7.35) years with 65 (54%) women, and the mean duration of back pain was 81 (SD 77) months.

In the methodological study comparing the utility indices, 133 out of 173 patients had completely filled out the ODI, the EQ-5D, and the SF-6D (SF-36) at baseline so values for each of the instruments could be calculated. At 2-year follow up, 113 patients had values for all three instruments, so change scores could be calculated.

3.3 Treatment

3.3.1 Surgery

The Prodisc-L prosthesis (ProDisc II, Synthes Spine) is a semiconstrained type prosthesis (table 3). It is unconstrained in axial rotation and semi-constrained with respect to flexion, extension and lateral bending. It consists of 3 parts: A lower and an upper endplate of Co-Cr alloy with a keel and between these an insert of polyethylene (UHMWPE – ultra high molecular weight polyethylene) which is attached to the lower endplate [171]. The endplates comes in large or medium size, the angle between them 6 or 11 degrees and the height between them 10, 12 or 14 mm. Access was made through a Pfannenstiel or a para-median incision with a retroperitoneal approach[243]. Nearly whole of the diseased disc was removed and the ProDisc II prosthesis was placed using instruments from the

Figure 4. The ProDisc II

manufacturer with the guide of a fluoroscope. The operation technique was standardized and the surgeons who performed the operation should have inserted at least 6 disc prostheses before operating on the patients in the study. Hospital stay postoperatively varied from 2 to 21 days (mean 7 days) postoperatively. Routinely they got a sick leave for 6 weeks on departure from hospital. They were not allowed to receive any kind of physiotherapy training until the first control 6 weeks postoperatively.

3.3.2 Multidisciplinary Rehabilitation

This was an outpatient programme with emphasize on exercises and cognitive intervention. The model was similar to that used in the study of Brox et al [151,152]. Before the study started three consensus conferences were held with physiotherapists and specialists in physical medicine and rehabilitation from the five study centers in order to standardize the treatment. The treatment was interdisciplinary and directed by a team of physiotherapists and specialists in physical medicine and rehabilitation. Patients were given exercises and education 5-6 hours a day 5 days a week for 3 weeks. In the exercise part, emphasis was put on general mobility and included water exercises, general strength work outs and endurance training outdoor and indoor. In the educational part lectures about basic physiology and anatomy of the back and pain physiology were given along with motivational lectures. In the motivational lectures it was stressed that taking part in physical activities was not dangerous and did not aggravate problems with low back pain. This message was later reinforced both during treatment and on follow up. Cognitive coping strategies were discussed with the patient individually or in groups. With local adjustments, services from other specialties like psychologists, nurses and social workers were also offered. As a part of the treatment plan patients were invited to a control at the outpatient clinic three, six and twelve months after the end of the programme.

3.4 Imaging

In our study, inclusion criteria included several degenerative changes of the intervertebral disc visualized on MRI that were thought to have relation to the experience of low back pain [119,244,245]. The MR series was evaluated by two independent observers [140,246,247].

3.4.1 Disc height

The midsagittal diameter of the intervertebral disc was measured in mm [109]. Interobserver reliability was moderate to good (k = 0, 62 at L4/L5 and 0, 58 at L5/S1) [246].

3.4.2 Modic changes (MC)

The primary and secondary signal intensity changes in the vertebral bone marrow adjacent to the endplate were rated as: 0 = no changes, type I = hypointense T1 signal and hyperintense T2 signal, type II = hyperintense T1 signal and iso- or slightly hyperintense T2 signal, and type III = hypointense T1 signal and hypointense T2 signal [120].

The maximal craniocaudal (CC) extension of MC at each separate endplate were rated based on the Nordic Modic Consensus Group criteria: 0 = no signal changes, 1 = located to the endplate only (minimal or small dots), 2 = less than 25% of vertebral body height, 3 = 25 – 50% of vertebral body height, 4 = more than 50% of vertebral body height [248]. Based on our data, the interobserver reliability was found to be moderate or good for type and extent (k = 0,55 – 0,77) [246].

3.4.3 Posterior HIZ

The area of high-signal intensity in the posterior annulus fibrosus that is brighter than nucleus pulpous on T2-weighted images and is surrounded superiorly, inferiorly and anteriorly by the low-intensity (black) signal of the annulus fibrosus was rated as present or non-present [125].

Interobserver reliability was found to be moderate, but better at L4/L5 than L5/S1 (k = 0,58 vs. 0,46) [246].

3.4.4 Nucleus pulposus signal

The signal was visually rated as bright (1), grey (2), dark (3) or black (4) on sagittal T2-weighted images, using cerebrospinal fluid as intensity reference [134]. Interobserver reliability was moderate to good (k = 0, 69 at L4/L5 and 0, 58 at L5/S1) [246].

3.5 DCRA method

3.5.1 Measurement protocol

When measuring the rotational and translational segmental motion as well as disc height and dorsoventral displacement of lumbar columna segments, the method of Distortion Compensated Roentgen Analysis (DCRA) compensates for distortion caused by axial rotation, lateral tilt and off-centre positioning of the spine [118,249,250]. This permits to process radiographs taken in normal clinical settings. Knowledge of the exposure geometry is not required. All motion segments imaged on a lateral radiograph can be evaluated. Sagittal plane rotational motion is obtained in degrees. Dorsoventral displacement and translational motion, disc height and vertebral height are determined in relative units, i.e. divided by the individual, mean vertebral depth. This is done in order to compensate for variations in radiographic magnification and stature.

Measurement errors for sagittal plane rotational motion amount to 1.0° - 2.3° and for translational motion to 1.6% - 3.4% of vertebral depth, in both cases the largest error occurring at L5/S1. Disc height, dorsoventral displacement and vertebral height are determined with errors in the order of 1.8%, 1.5% and 1.2% of mean vertebral depth, respectively. To perform DCRA, the contours of the vertebrae are mapped and digitised. Series of computer programs check geometric properties of the contours, objectively locate vertebral ‘corners’ and calculate the parameters.

3.5.2 Data collection and analysis

Sagittal plane motion of the segments TH12/L1 to L5/S1 is determined from the pre- and postoperative pairs of flexion-extension radiographs. In the operated segment, disc height (or postoperatively: height of the intervertebral space) and dorsoventral displacement is determined from the pre- and postoperative radiographs taken in extension. As disc height and

displacement (as defined here) both depend on the angle of lordosis, and in order to permit comparison with normal data, disc height and displacement are corrected to standard angles of lordosis. For the purpose of quality control, height of the cranial and caudal vertebrae of the operated segments, measured from the pre- and postoperative pairs of radiographs, are compared: For each vertebra the four height values determined pre- and postoperatively in extension and flexion should coincide within the limits of the measurement error. Rotational and translational motion as well as disc height and dorsoventral displacement are compared with previously determined normal data [118,249]. As the magnitude of translational motion depends linearly on the magnitude of rotational motion, actual translational motion is also compared with that motion predicted for a normal subject under the individual magnitude of rotational motion. Thus, the comparison between actual and predicted translational motion is independent of the magnitude of rotational motion performed by the patient. For the segments instrumented with disc prostheses, the deviation of disc height and displacement from the norm is expressed in standard deviations S of the appertaining distribution in the normal population. This permits to pool disc height and displacement data from all patients studied.

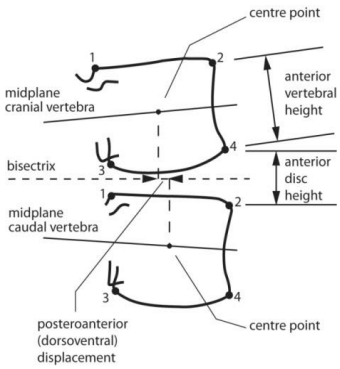

Figure 5. Parameters measured by DCRA

Table 4. Definition of DCRA parameters

DCRA parameter	Definition (figure 5)
Mean vertebral depth	Mean distances between corners 1 and 2 and corners 3 and 4
Sagittal plane angle	Angle between vertebral midplanes. The vertebral midplane is defined as the horizontal line that runs through the midpoints between the upper and lower corners of a vertebra (i.e. midway between corners 1 and 3 and corners 2 and 4).
Disc height	The sum of the distances from the corners of the adjacent vertebrae (e.g. corner 2 of the caudal disc and corner 4 of the cranial disc) to the bisectrix between the midplanes. This sum is divided by the mean depth of the cranial vertebra in order to express the value in units of vertebral depth. In this study, the term “disc height” is used synonymously with “intervertebral space”. Disc height can be compared with age- gender appropriate data. As the given sagittal plane angle is typically different from the reference angle of the normative database. The deviation of the corrected height of the intervertebral space from the norm is then dependent of the sagittal plane angle adopted when the radiograph was taken.
Posteroanterior (dorsoventral) displacement	Distance between the centre points (geometric centre of corners 1 through 4) of the vertebrae to the bisectrix, divided by the mean depth of the cranial vertebra. Displacement is counted as positive when the cranial vertebra is displaced in the anterior direction with respect to the caudal vertebra. Displacement can be compared with age- and gender appropriate normal data. As the given sagittal plane angle is typically different from the reference angle of the normative database, a correction is applied prior to the comparison. The correction depends linearly on the difference between the given sagittal plane angle and the corresponding reference angle of the norm. The deviation of the corrected displacement from the norm is then dependent of the sagittal plane angle adopted when the radiograph was taken. This holds for both mobile and fused segments.
Sagittal plane range of motion	The midplane angle during low back extension minus the midplane angle during low back flexion
Sagittal plane translational motion	The posterior displacement during low back extension minus the posteroanterior displacement during low back flexion

3.6 Instruments (Patient Related Outcomes)

3.6.1 ODI

As the primary outcome measure, we used the Oswestry Disability Index**.** The ODI is a back-specific questionnaire [251,252]. Patients rate physical disability in activities of daily living due to low back pain in 10 questions, each of which has verbal response alternatives. Ratings are summed to yield a score ranging from 0 (not disabled at all) to 100 (completely disabled). The ODI has been found to be a responsive and valid measure for patients with LBP [242,251,253]. We used the Norwegian translation of the validated questionnaire (version 2.0)[253].

3.6.2 SF-36

The Short Form 36 (SF36) is a generic health related quality of life questionnaire[254]. It measures quality of life along eight dimensions: Physical function, role physical, bodily pain, general health, vitality, social function, role emotional and mental health. Each dimension has 4-6 questions or items. Scores range from 0 to 100, higher scores corresponds to better health status. We used the Norwegian version of the SF-36v2. SF-36 are found to be a good measure of health status and patient function when compared to a disease specific instrument for low back pain patients[236,255,256].

3.6.3 SF-6D

The SF-6D utility index is comprised of 11 items from the SF-36 [254] that were revised into a six-dimensional health state classification system. The six dimensions are physical functioning, role limitations, social functioning, pain, mental health, and vitality. It reflects a continuous outcome scored on a 0.29–1.00 scale, with 1.00 indicating full health [202]. SF-6D health states were evaluated against a normal population using the Standard Gamble (SG) method. We used the United Kingdom (UK) tariff [202]. The SF-6D was calculated based on the Norwegian SF-36 (version 2) with the use of syntax files in SPSS 15(SPSS, New York, US).

3.6.4 EQ-5D

For the EQ-5D utility index, responses on a questionnaire with five dimensions, each comprised of three levels, are revised into an index with a range from -0.59 to 1, with 1.00 indicating full

health. The 243 possible health states on the EQ-5D are evaluated against a normal population using the time trade off method (TTO)[205 257]. We used the Norwegian version of the EQ-5D and syntax files obtained from the EQ-5D society using the UK tariff to calculate the index.

3.6.5 Other instruments

For psychological variables we included emotional distress (Hopkins symptom check list (HSCL-25), scores range from 1 to 4, with lower scores indicating less severe symptoms) and the fear avoidance belief questionnaire (FABQ) for work and physical activity (scores range from 0 to 42 (work) and from 0 to 24 (physical), with lower scores indicating less severe symptoms)[258,259]. Self-efficacy beliefs for pain were registered by a subscale of the arthritis self-efficacy scale (scores range from 1 to 10 and are summarized and divided by 5; lower scores indicate uncertainty in managing the pain)[260].

At the two year follow up control, two independent observers blinded to treatment evaluated patients using the back performance scale which consists of five tests with a score ranging from 0 to 15, worst possible[261] and the Prolo scale which consists of functional and economic parts, summed to a worst score of 2 and a best score of 10[262,263].

We also recorded satisfaction with the result of the treatment on a seven point Likert scale, and satisfaction with care on a five point Likert scale [235].

The visual analog scale (VAS) were also used [264]. It is one of the most used outcome instruments to measure pain [256]. Patients were asked to rate the intensity of low back pain on a visual analogue scale, ranging from 0 (no pain) to 100 (worst pain imaginable).

3.7 Health economic analysis

We performed a full health economic evaluation using a cost-utility analysis [193,194,198].

3.7.1 Treatment effects and health utilities

The EQ-5D utility index [265] was used in the main analysis and the SF-6D utility index [266] was used for comparison [267]. Both costs and effects were measured at baseline; 6 weeks (not SF-6D); and 3, 6, 12, and 24 months post-treatment. Combining utility indexes and time, the quality-adjusted

life years gained (QALYs) were estimated as area-under-the-curve (AUC) using the trapezoidal method [194,268].

3.7.2 Costs and resource use

All relevant costs were identified, measured, and valued. Resource use was assessed, and the analyses were performed from a societal perspective, including index treatment, other hospital care, primary care, patients' private costs, and costs due to loss of production both for the patient and their relatives. The Norwegian krone (NOK), with 2006 as a base year was used, and costs were inflation-adjusted into 2012 prices and converted into Euros using the rate 1 $€_{2012}$ = 6.7 NOK_{2006}. Actual costs were assigned to patients regardless of their randomized group, so patients who were randomized to receive MDR but crossed over and received operation after having had MDR were assigned costs for both treatments.

All relevant costs and resources were identified, measured, and valued as follows:

1. Index treatment. For TDR, the resource use multiplied by unit costs, and incorporating spare capacity when appropriate, summarized the cost for each Index treatment. Cost components included were: prosthesis, operation room time, wake-up services, post-operative stay in hospital, and post-operative x-ray.

For MDR, we used a top-down approach, i.e., the total cost of a spine clinic was estimated, and then how much of the clinic's costs were associated with MDR was determined [269]. A consequence of this approach is that the costs are the same for all patients. Spare capacity was included. A premium of 12% was added to common costs based on data from previous estimates of the cost weights for the Norwegian DRG system (ISF) [270].

2. Hospital costs during follow-up. The number of planned and unplanned re-admissions, including outpatient visits and re-operations were registered in electronic patient administrative systems. Patients who underwent surgery received one mandatory consultation with an X-ray 6 weeks after surgery. Patients in the MDR group were offered four follow-up consultations at 6 weeks and 3, 6, and 12 months, and costs were assigned if accepted.

3. Primary care costs during follow-up. Unplanned visits to general practitioners, physiotherapists, or other practitioners in the public health service were recorded in a cost diary kept by the patient as described in a previous cost-effectiveness study [212].

4. Patients' private resource use. The use of medication (both prescribed and over-the-counter), contact with practitioners outside the public health service, and other costs were reported by the patient in a cost diary. Costs for relatives were included.

5. Loss of production. The human capital approach was used to estimate the costs related to days each patient spent out of work due to low back pain. Costs related to production losses were calculated as the number of days out of work multiplied by the average wage adjusted for part-time sick leave. Income before taxes was used for patients and after taxes for relatives when calculating costs related to work loss [271].

3.8 Psychometric evaluation of outcome instruments

We used criteria from the COSMIN checklist[272] for assessing five central measurement properties when comparing the EQ-5D, the SF-6D and the Oswestry Disability Index:

1. Measurement error concerns the systematic and random error of a patient`s score that is not attributed to true changes in the construct to be measured [231].

2. Structural validity concerns the degree to which the scores of an instrument are an adequate reflection of the dimensionality of the construct to be measured [272].

3. Criterion validity concerns the degree to which the scores of an instrument are an adequate reflection of a "gold standard" when this is present [272].

4. Responsiveness is defined as the ability of an instrument to detect change over time in the construct to be measured [272].

5. Interpretability concerns the qualitative meaning of quantitative scores or change in scores. A core question is: "What is the smallest change in score in the construct to be measured which patients consider important?" This is expressed as the Minimal Important Change (MIC) value [272].

Table 5 Taxonomy, definitions, study questions and statistical methods used in study 4. [273]

COSMIN taxonomy	Definition	Study question	Statistical method(s) used
1. Measurement error	The systematic and random error of a patient's score that is not attributed to *true* changes in the construct to be measured	How much of a change score value of EQ-5D, SF-6D and ODI is actually "noise"? (Preciseness of the instruments).	• Standard Error of Measurement, • Minimal Detectable Change score • Bland Altman plot (agreement) and Limits of Agreement
2. Structural validity	The degree to which the scores of an instrument are an adequate reflection of the construct to be measured	Do EQ-5D and SF-6D measure HRQOL a)On a scale? b)Along the same scale? Are items in EQ-5D and SF-6D adequate?	Rasch Analysis with a) Unrestricted (Partial-Credit) polytomous model b) Principal Component Analysis
3. Criterion validity	The degree to which the scores of an instrument are an adequate reflection of a "gold standard"	To what degree do EQ-5D and SF-6D show similarity with ODI?	Spearman rank correlation coefficient with 1000 bootstrap replications of *baseline* scores of ODI, EQ-5D and SF6-D
4. Responsiveness	The ability of an instrument to detect change over time in the construct to be measured	If there is a change in health status – to what degree will EQ-5D, SF-6D and ODI detect it?	• Spearman rank correlation between *change* scores of ODI, EQ-5D and SF-6D • Receiver Operating Curve • Area Under Receiver Operating Curve
5. Interpretability	The qualitative meaning of quantitative scores or change in scores.	What is the lowest possible change score on the EQ-5D, SF-6D and ODI scale that patients would detect as clinical meaningful?	Instrument change score value with optimum sensitivity and specificity on the Receiver Operating Curve; The Minimal Important Change Score value (MIC value)

3.9 Statistical methods

3.9.1 Study 1

The main statistical analysis was in the intention to treat population at one and two year follow-up. According to our protocol the analysis was performed with the assumption that patients who dropped out had no improvement after drop-out (last value carried forward). We also determined if different centres had different outcomes. We used $\chi 2$ test or Fisher's exact test to analyze categorical variables and independent two sided t test or analysis of variance to analyze continuous variables. A significance level of 5% was used throughout. We did not adjust for significantly different baseline scores. We conducted a per protocol analysis for the primary outcome variable (score on Oswestry Disability Index). Consistent with criteria from the Food and Drug Administration, we considered an individual change in score of at least 15 points from baseline to two year follow-up as a minimal important change. A deterioration of 6 points in the score was considered a "change for the worse"[274]. We calculated the number needed to treat with confidence intervals[275]. A mixed model analysis was used to evaluate the effect of each efficacy variable over time and between groups. In the mixed model patients were not excluded from the analysis of an efficacy variable if the variable was missing at some, but not all, time points after baseline. In the additional analysis (categorical or ordinal data at two year follow up), missing data were not replaced. Significantly different baseline scores were not adjusted for in the longitudinal model. Each outcome variable was adjusted for the baseline values of the variable.

3.9.2 Study 2

Time-weighted averages of the preference scores that were measured at the beginning and end of each measurement period were used to calculate QALYs from the EQ-5D and SF-6D utilities.[194] The accumulated QALYs for TDR and multidisciplinary rehabilitation over 24 months were calculated for the following periods: baseline to 6 weeks (not SF-6D), 6 weeks to 3 months 3 to 6 months, 6 to 12 months, and 12 to 24 months[213].

An incremental cost-effectiveness analysis of the index treatment was carried out. Cost effectiveness was calculated as the difference between the costs of the surgery group and the

rehabilitation group divided by the difference in QALYs gained between the two groups. The results are presented as the incremental cost-effectiveness ratio (ICER).

In order to derive a 95% confidence interval (CI) for the ICER, we used a nonparametric bootstrap method using 10,000 replications [85]. The bootstrap replications were plotted in a cost-effectiveness plane to illustrate the uncertainty of the ICER estimate [276]. The concept of net monetary benefit was used to construct a cost-effectiveness acceptability curve (CEAC).[277] CEAC shows the probability that surgery is cost effective at 2 years in terms of the desired QALY.[277]

Patients who dropped out of this study during the first 6 months of follow-up were removed from further analysis because it was believed that information regarding resource use and utility values would be too scarce for further analysis. Patients who crossed over from one treatment group to the other were analyzed according to the intention to treat principle: After cross-over, further resource use and utility values were analyzed according to the group the patient was randomly assigned to at baseline.

In this study, 13% of the resource use items on follow-up and 8.3% of the utility scores were missing between baseline and 24 months. Multiple imputations were used to address these missing data.[278-281] Each missing value was replaced with *m* plausible values, where *m* is the number of imputations performed. The missing values were replaced using a multiple linear regression model. The covariates included the intervention group, age, and sex. This imputation method was used to determine both the costs of the resources and the utility scores used to produce $m = 5$ data sets. Arithmetic means and 95% CIs are presented for the costs and QALYs of each trial group and the average of the means of the five datasets.

The Student *t* test and corresponding 95% CIs were used to analyze differences in cost and utility. The Chi-square test was used to analyze differences in the times required to return to work. Two-sided p values < 0.05 were regarded as significant.

3.9.3 Study 3

The independent t-test was used to examine the differences in ROM, disc height, and alignment between the disc prosthesis cohort and the non-operated cohort. The independent t-test was also used for comparing inter-individual changes in the prosthesis group. When comparing intra-individual changes from the start of the study to the two-year follow up, a paired t-test was used in the non-operated segments of the prosthesis group and in all segments of the intensive

rehabilitation group. The dependent t-test was used for intra-individual changes in operated segments. Spearman's rank correlation coefficient was used to examine the association between disc height and segmental motion against the clinical postoperative variables ODI, EQ-5D, VAS, and the PF of SF-36. In patients with disc prostheses at two levels, segmental range of motion and disc heights were averaged when assessing the correlation to clinical outcome. Therefore, one and two level disc prosthesis patients were analysed as one cohort [227].

3.9.4 Study 4

A combination of statistical methods is generally recommended for interpreting changes of patient related outcomes [282,283].

1. Measurement error

We used the standard error of measurement (SEM) to express instrument imprecision [235,284-286]. The advantage of using SEM is that it is considered to be an attribute of the measure and not a characteristic of the sample itself[287]. The SEM value could be calculated from a test-retest study or in a group of stable patients. The SEM in this study was calculated as:

$$S_w = SEM = \sqrt{\frac{1}{2n}\sum d_t^2}$$

where s_w is the within-subject standard deviation, d is the difference between two observations in patients i who reported "unchanged" on a four-point scale between 3 and 6 months follow up and n is the number of subjects[288]. The s_w statistics is also called the $SEM_{consistency}$ [289].

The lowest change that exceeds measurement error and noise at a 95% confidence level is defined as:

$$MDC_{95} = 1.96 * \sqrt{2} * SEM = 2.77 * SEM$$

Here, the $*\sqrt{2}$ is introduced because there are two measurements for each patient. The minimum detectable change (MDC) at a 95% confidence level, is denoted MDC_{95} [290]. With a scale value

≥MDC_{95}, we can be 95% certain that a change in the measured underlying construct has really occurred[274].

To assess the agreement between EQ-5D and SF-6D, a Bland Altman plot was constructed. [291]. The average EQ-5D and SF-6D change score values were plotted against the mean difference in change score values of both instruments. Limits of Agreement (LoA) based on a +/- 196*$SD_{difference}$ interval for the differences were also constructed.

2. Structural validity

Both EQ-5D and SF-6D are constructed to measure the dimension of general health related quality of life (HRQoL) alongside a continuous scale (from low to high). Using Item Response Theory (IRT), the unidimensionality of the two utility indexes was tested. The category ordering of the questionnaire items (the probability of moving from an easier to a harder accomplished category of item answers in parallel with being increasingly disabled) was also tested.

We employed the unrestricted (Partial-Credit) polytomous model of the Rasch model (for general information about fit to the Rasch model, see appendix A) and the test proposed by Smith to reveal unidimensionality [292]. The SF-6D and EQ-5D were tested for unidimensionality in a principal component analysis (PCA) [293]. We performed a test equating procedure with baseline values from the SF-6D and the EQ-5D. The response of each patient to a question was tested against what was predicted by the Rasch model. Deviation from the model is expressed in residuals. Independent t-tests were used to test if the magnitude of the residuals represents a significant deviation. The CI calculated for this was 95%. We carried out a binominal test for the proportion of t-tests outside the range of -1.96–1.96.

3. Criterion validity

In this analysis we compared the scores of the EQ-5D and SF-6D to the disease specific instrument ODI. The rationale was that the ODI has been found to be a responsive and valid measure for patients with LBP [242,251,253]and that an improvement assessed by the ODI should be correlated with an improvement assessed by the two utility indexes.

Spearman rank correlation coefficient (r) with 1000 bootstrap replications of the *baseline* scores was calculated to assess the correlation between the scores of the EQ-5D and ODI and SF-6D and ODI.

4. Responsiveness

Responsiveness was assessed by using the ODI and the seven-point global scores at 2-year follow-up as "gold standard". First, we calculated the Spearman rank correlation coefficient (r) with 1000 bootstrap replications for the correlation between *change* scores from baseline to 2 year FU for the EQ-5D, SF-6D and ODI. Second, we analyzed the area under the Receiver Operator Curve (ROC) for the change scores of the EQ-5D, SF-6D and ODI by using a dichotomization of the patient global scores as follows: Categories 1 to 3 was considered "improved" and categories 4 to 7 were "non-improved". Sensitivity was defined as the proportion of patients who were correctly classified as "improved" and specificity was defined as the proportion of patients who were correctly classified as "non-improved". A receiver operating characteristic (ROC) curve was then calculated by plotting every possible change score from baseline to 2 year FU for EQ-5D, SF-6D and ODI using the global score as an anchor[294,295]. The area under the ROC curve (AUC) was then calculated. This value corresponds to the possibility of correctly diagnosing a patient as having improved when this is really the case[295] and reflects how responsive the instruments are to detect a change in the underlying construct.

5. Interpretability

Interpretability was calculated based on the sensitivity and specificity results from the ROC analysis described above. The cut-off value for differentiating between patients with or without improvement at optimum sensitivity and specificity was determined using ROC analysis [295]. This corresponds to the upper left point on the ROC curve and it can be interpreted as the point or value that yields the lowest overall misclassification [284,296].

4 RESULTS

Paper I: Surgery with disc prosthesis versus rehabilitation in patients with low back pain and degenerative disc: two year follow-up of randomized study

In the planned analysis, the mean change from baseline to two-year follow-up was 20.8 ODI points (95% CI, 16.4 to 25.2) in the surgery group and 12.4 points (95% CI, 8.5 to 16.3) in the rehabilitation group. The mean treatment effect (between-group difference) at two-year follow-up was 8.4 ODI points (95% CI, 3.6 to 13.2). In the mixed model analysis (unplanned analysis according to the original protocol), ODI improved significantly more in the surgical group than the rehabilitation group at all time points. The mean change from baseline to two-year follow-up was 22.5 (95% CI 18.5 to 26.4) in the surgery group and 15.6 (95% CI, 11.7 to 19.5) in the rehabilitation group. The mean treatment effect (between-group difference) at two year follow-up was 6.9 (95% CI, 2.1 to 11.7). 70% (n=51) of patients in the surgical group and 47% (n= 31) in the rehabilitation group had an improvement in ODI of at least 15 points ($p< 0.006$) in the unplanned analysis. The number needed to treat was 4.4 (95% CI, 2.6 to 14.5) (intention to treat).

Paper II: Cost-effectiveness of total disc replacement vs. multidisciplinary rehabilitation in patients with chronic low back pain – a Norwegian multicenter RCT

The mean QALYs gained (SD) using EQ-5D was 1.29 (0.53) in the TDR group and 0.95 (0.52) in the MDR group, a significant difference of 0.34 (95% CI, 0.18 to 0.50). The mean total cost per patient (SD) in the TDR group was €87,622 (58,351) compared with €74,116 (58,237) in the MDR group, which was not significantly different (95% CI: -4,041 to 31,755). The Incremental Cost Effectiveness Ratio (ICER) for the TDR procedure varied from €39,748/QALY (95% CI: €15990 to €65 645) using EQ-5D (TDR cost- effective) to €128,328/QALY (95% CI: €51 329 to €219 907) using SF-6D (TDR not cost-effective). Using per-protocol instead of ITT indicated that TDR was not cost-effective, irrespective of the use of EQ-5D or SF-6D. Not using multiple imputations for missing data resulted in a considerable loss of observations and higher ICER, rendering TDR not cost - effective.

Paper III: Segmental mobility, disc height and patient-reported outcomes after surgery for degenerative disc disease: a prospective randomized trial comparing disc replacement and multidisciplinary rehabilitation

No significant change in movement in the sagittal plane from baseline (pre-surgery) to two years (post-surgery) was found in segments with TDR. At the L4/L5 segment, a significant difference of 2,88° (p = 0.041, 95%CI: 0.11° to 5.25°) was found between the two treatment groups at two year follow up with greater mobility in the TDR group. At the L5/S1 segment, a non-significant difference of 1.64° (p = 0.077, 95%CI: -0.18° to 3.47°) was found. It remained the same or increased slightly (significant at the L4/L5 level only) in untreated segments in the TDR group. The disc height was significantly increased in the TDR group compared to the MDR group. There was no correlation between segmental movement or disc height and patient-reported outcomes in either group two years after treatment.

Paper IV: Comparison of the SF-6D, the EQ-5D, and the Oswestry Disability Index in patients with chronic low back pain and degenerative disc disease

The SF-6D had better similarity with the disease-specific instrument (ODI) regarding sensitivity, specificity, and responsiveness. Measurement error was lower for the SF-6D (0.056) compared to the EQ-5D (0.155). The minimal important change score value was 0.031 for SF-6D and 0.173 for EQ-5D. The minimal detectable change score value at a 95% confidence level was 0.157 for SF-6D and 0.429 for EQ-5D. The difference in mean change score values (SD) between them was 0.23 (0.29) exceeding the minimal important change score values of both measures. Analysis of psychometric properties indicated that the indexes are unidimensional when considered separately, but that they do not exactly measure the same underlying construct.

5 DISCUSSION

The principal findings from our study were:

- A statistical significant difference in the main outcome variable (ODI) in favor of disc prosthesis surgery when compared to non-surgical treatment was found.

- Total disc replacement was cost-effective compared to multidisciplinary rehabilitation over two years. However, this result was dependent on the choice of clinical outcome measure for effectiveness (utility index).

- No significant difference in range of segmental motion from pre- treatment to two year post-treatment was found in segments with prosthesis. No correlation between segmental movement and clinical outcome was found in either treatment group.

- The difference in important measurement properties between EQ - 5D and SF - 6D, two of the most used patient related outcome variables in cost-effectiveness studies, was too large to consider them interchangeable.

A general discussion about some aspects of the trial design is given before discussing the main findings in the trial.

5.1 Randomized controlled Trials and bias

By randomly allocating patients to two or more treatments, the goal in randomized controlled trials is to eliminate bias [208,241,297]. Jadad [241] defines bias as "…any factor or process that tends to deviate results or conclusions of a trial systematically away from the truth". Several possibilities of bias exists [241].

5.1.1 Selection bias

Selection bias occurs when the way in which individuals are selected for a trial or the way in which the interventions are administered after people are accepted into the study affects the outcomes [241].

Prior to randomization, patients in our study were interviewed and received information about the treatment options by both a physical internist and a surgeon. In that way, we ensured a balanced information process and that all patients were provided with the best available information concerning the treatment options. Participants found eligible for the study according to exclusion and inclusion criteria were then randomized in blocks of four at each of the study centers (university hospitals) with the help of a website hosted centrally by the medical faculty of the University Hospital of Trondheim, St Olav. This randomization process ensured that treatment group allocation could not be influenced by the investigators or the study participants.

5.1.2 Ascertainment bias.

Ascertainment bias occurs when the results or conclusions of a trial are systematically distorted by knowledge of which intervention each participant is receiving[241,298]. It can be introduced by the person administering the interventions, the participants, the investigators analyzing and assessing the outcomes and by people that write the describing report from the trial [241].

Because of the nature of this trial, it was not possible to keep participants or personnel involved in the treatment unaware of the intervention identity after randomization had taken place. However, independent observers collected and entered data. In addition, two independent observers blinded to treatment [299] evaluated patients at the final two year follow up consultation. Patients were informed before this session not to reveal the treatment received and had a tape placed on their abdominal wall to hide scarring from the operation.

Patients may have had expectations about which treatment they could be allocated to. This in turn could influence on their motivation for treatment and the self-reporting of outcome after treatment. For example, one of the inclusion criteria was that patients should have had structured physical rehabilitation therapy with the duration of at least one year before attending the study. Patients randomized to the MDR could have felt that "this is more of the same" and because of that have lost their motivation. This could have influenced their answers when evaluated on the follow up controls.

5.1.3 Bias introduced by inappropriate handling of withdrawals, drop outs and protocol violations.

This type of bias could occur for example if patients that do not benefit from the treatment withdraw more often than those who stay in the study [241]. To prevent this, we performed an Intention To Treat analysis with last value carried forward for patients who dropped out. This ensures that treatment groups are equal apart from random variations and at the same time, it allows for non-compliance and deviations from treatment approach as it would appear in routine practice outside a trial setting [300].

It could be argued that patients who withdrew after randomization or dropped out during or after treatment had a superior or inferior outcome. In order to assess this, we sent a questionnaire to such patients. The nine patients who withdrew after surgery experienced a reduction in Oswestry score of 30.2 (SD 4.5) points. The six who withdrew after rehabilitation had a reduction of 11.8 (SD 3.0), and the 11 patients who withdrew without treatment had no change (1.0 (SD 4.5) points).This might support the assumption of no improvement in outcome after drop-out, justifying use of the last value carried forward analysis.

5.2 Interpretation of main findings

5.2.1 Surgery with disc prosthesis versus rehabilitation

MDR

The MDR used in this study was published in 2003 and 2006. It was a single center study and the multicenter intervention in the present study may differ slightly because it involves several centers with different treatment cultures. The mean ODI was reduced by 29% (12.4 points) in

the rehabilitation group of our study (ITT analysis). Two former single center studies have used a form of MDR similar to the one used in our study: Brox[152,301] et al found an analogue reduction of 29% (12.0 points) at one year follow up while Fairbank[153] et al found a smaller reduction of 19% (8.7 points) at two year follow up.

Such comprehensive regimens like the MDR in our study have not been well compared with more simple interventions like the brief intervention described by Indahl in the trial in the -90-ties [302,303] and by Storheim et al in 2003 [304,305]. In the latter study the authors found no effect on return to work, but for most other variables a simple cognitive intervention (1-2 consultations) was better or as good as a more comprehensive group exercise program. In our study, there was no difference between treatment groups in return to work after two years and the "net back to work"(subtracting patients who went back to work from patients who stopped working[150]) was 31% (n=21) in the surgical group and 23% (n=15) in the rehabilitation group (P=0.31).

TDR

Previous studies have reported on the effectiveness and safeness of TDR surgery compared to fusion surgery[306-313]. Due to different prosthesis designs and different outcome measures, it is not straight forward to compare these studies to our study. In addition, many of these studies has been assessed as of low quality either because of high risk of bias or conflict of interest[177,314].

In a Cochrane study, Jacobs[315] et al looked at five studies with 1301 patients comparing TDR (865 patients) to fusion surgery (436 patients). They conclude that although surgery with TDR compared to fusion surgery showed a statistical significant improvement in favor of TDR, the differences on the primary outcome parameters, including their 95% CIs, were small and did not exceed pre-defined clinical relevance. This was confirmed in a meta study by Eerenbemt[314] et al which looked at 116 studies including prospective cohort studies, prospective controlled studies, cost-efficacy studies and studies reporting complication rates.

In a study including 7172 TDR procedures and 62 731 anterior fusion procedures, Kurtz[316] et al found that the revision burden for TDR (11,2%) was significantly higher than fusion surgery (5,8%). Furthermore, the revision rate of TDR fell within the revision burden range of hip and knee replacement, which are generally considered successful and cost-effective procedures. In our study, there was a complication rate of 8% resulting in impairment at two year follow-up and

a revision rate of 6% including a patient that lost a leg as a direct consequence of revision surgery. This revision rate was within acceptable limit compared to other reports. However, it must be taken into account that we (similar to the aforementioned studies comparing TDR to fusion surgery) were not able to show a clearly clinical significant difference in ODI score at the two year follow up between treatment groups.

Shear stress forces cause wear particles from the polyethylene inlay of the prosthesis to accumulate[317-319]. Punt[317] et al showed that the wear and tear of the polyethylene in TDR is compatible with total hip or total knee replacement. These particles cause an inflammatory response in the periprosthetic tissue and there is evidence of increasing wear with implantation time and a potential for osteolysis in the spine[319,320]. Potential long-term revision rate with adhering complication rate on revisions needs to be considered [321]. Several authors therefore recommend regular long-term follow-up for patients undergoing TDR[318,319]. The need for longer follow up in the prosthesis group and the expenditure of this should therefore be taken into consideration when assessing and comparing clinical and health economic consequences of choosing disc prosthesis surgery.

Serious consequences of vascular injury in anterior lumbar surgery seldom occur[322] but could nevertheless be critical[323] and should be accounted for when comparing treatment methods. The risk of vascular injury is greater during revision than primary procedures[176].

ODI

The rationale for choosing ODI as the primary outcome variable was that it has been found to be a responsive and valid measure for patients with LBP [242,251,253]. The ODI is by far the most used and validated disease specific instrument for patients with chronic low back pain [242]. The validity and reliability of the Norwegian translation of the Oswestry Disability index was documented by Grotle et al [253]. Even small changes can be detected with this instrument [256,324-326]. The high responsiveness of ODI was confirmed in study IV were the area under the ROC curve, the possibility of correctly discriminating between “improved” or “non-improved” patients with a 95% CI was 94% (87.5–97.6). The effect size of ODI in our study was quite similar to values reported in other studies of patients with low back pain [150-153]

Mean difference in improvement in Oswestry Disability Index after 24 months in the Cochrane study of Jacobs et al[315] comparing TDR to fusion surgery was 4.27 (95%CI 1.85 to 6.88) well

below the pre specified clinical significant score value[308] of the difference suggesting that there is no clinical relevant differences between total disc replacement and fusion techniques. While statistically significant effects are those that occur beyond some level of chance, clinical significance refers to the benefits derived from treatments, its impact upon the patient and its implication for clinical management of the patient [327,328]. To assess the clinically significance in our study we used the concept of Minimally Important Change (MIC) in study IV [289]. The MIC value was defined as the smallest change in ODI score from baseline to two year follow-up which patients considered important [237,289]. The power analysis of study 1 was based on detecting a clinical significant difference between groups of 10 ODI points [151,152]. Unfortunately, there is no agreement on the clinically important difference between two treatment groups so the value of 10 ODI points in our study was a somewhat arbitrarily chosen limit. As an alternative, in study I, we assessed the proportion of patients that on an individually basis achieved a clinically meaningful improvement using a value of 15 ODI points [308,329]. Using this value, 70% of the patients in the surgical group and 47% in the rehabilitation group achieved this clinical important difference. However, it is generally recommended to calculate the MIC value specific to each study which we did in study IV [330,331]. The MIC value for the ODI on an individually basis found in study IV was 12,88 (sensitivity 88%, specificity 85%), which was almost exactly the same value found in a study by Copay[296] et al. Using this value, the numbers increased to 75% (n= 55) in the surgery group and 57% (n=37) in the rehabilitation group. The minimal detectable change score value at a 95% confidence level ($MDC_{95,}$ the limit for changes in score value that occur beyond measurement error or "noise" with a probability of 95%) calculated in study IV was 11.75 points for ODI. The number of patients in study I that experienced a change in ODI beyond this value (intention to treat) was 77% (n= 56) in the surgery group and 61% (n= 40) in the rehab group. Although recommended, these parameters (MIC and MDC_{95}) are not usually presented in clinical studies with the ODI as an outcome variable but we think this is strength of the study because it gives a more precise estimate of the effect after treatment.

The clinically important change score value of 12.88 ODI points found in study IV could be further compared to results from study I. In the primary (planned) analysis of the change in ODI score from baseline to two years follow-up, there was an improvement of 20.8 (95% CI 16.4 to 25.2) in the surgery group and 12.4 (95% CI 8.5 to 16.3) in the rehabilitation group. In the mixed model analysis (intention to treat, unplanned analysis according to the protocol) the numbers were 22,5 (95%CI 16,4 to 25,2) for surgery and 15,6 (95%CI 11,7 to 19,5) in the rehabilitation

group. The method of last value carried forward was used in the first study in case of drop out with the assumption that patients who dropped out had no improvement after leaving the study. In the mixed model analysis, patients were not excluded from the analysis of an efficacy variable if the variable was missing at some, but not all, time points after baseline. The interpretation of this is that patients in the surgery group had a clinical important change in ODI score regardless of analysis method used while this was true for the patients in the rehabilitation group only when missing values were adjusted for.

In conclusion, revision surgery after TDR surgery is complicated and potentially more dangerous than revision after fusion surgery. The need for longer follow-up because of wear and tear of the polyethylene inlay is recommended indicating increased resource use in this group. Surgery with TDR do not show a clearly clinically significant difference in improvement compared to fusion surgery in most studies. This is also the case when comparing TDR to MDR in our study.

Finally, we based our study inclusion criteria on an anticipated causal relationship between clinical presentation of low back pain and morphological changes of the intervertebral disc on MRI scans. This correlation has not yet been fully established[332]. In a separate study we found that the combined MRI findings from our inclusion criteria were not related to the degree of disability or the intensity of LBP [247]. Other contributing causes of low back pain should also be considered.

5.2.2 Cost-effectiveness of total disc replacement

As mentioned by Drummond et al, it is useful to contrast two different activities in health care evaluation: Measurement and decision analysis [193].

We present results of a cost-utility analysis (measurement) based on a single randomized controlled trial as opposed to a modeled (decision) analysis. The latter would have permitted us to extend the follow up period beyond two years and to include more precise estimates of rare events. Also with a modeling approach, other relevant treatment options (e.g. different kinds of non-surgical treatment or no treatment) could be compared to TDR [193,194]. A modeling approach may be a better tool to inform decision makers and perhaps make a better fundament for a decision analysis because the information is on a more general basis.

Due to heterogeneity of the populations, interventions, and comparison groups, several authors of review studies report that firm conclusions about the effectiveness of different non-surgical

treatment options are hard to draw [103,186]. This heterogeneity in addition to variable reporting quality of the few existing cost utility studies of low back pain makes direct comparison of both effectiveness and cost of different non-surgical treatment options difficult [220]. Thus, we cannot exclude that a much simpler intervention had been as effective and therefore influenced the cost-effectiveness (less costly, same effect).

In our study, we found that cost per quality-adjusted life year (QALY) gained using TDR was €39,748/QALY (95% CI, €51 330 to €219 907) or approximately $53,627 (95% CI, $64 933, $278 18) using EQ-5D. Tosteson et al compared surgery with nonoperative care for three common diagnoses: spinal stenosis (SPS), degenerative spondylolisthesis (DS), and intervertebral disc herniation (IDH) [333]. Using EQ-5D as outcome measure for clinical efficacy, the costs per QALY gained decreased for SPS from $77,600 at 2 years to $59,400 (95% CI: $37,059, $125,162) at 4 years, for DS from $115,600 to $64,300 per QALY (95% CI: $32,864, $83,117), and for IDH from $34,355 to $20,600 per QALY (95% CI: $4,539, $33,088). Although the indication for surgery is less controversial in the Tosteson study than ours, this illustrates how the cost-effectiveness for spine surgery may change in longer follow up studies perhaps because the improvement in quality of life is sustained while the cost pr QALY decreases with time. The results from the Tosteson study also show that our estimate of cost per QALY gained was in the same range as other common spinal surgery procedures when compared to non-operative treatment. We conclude in our study that longer follow-up is needed to assess how the cost-effectiveness of TDR will change.

During follow-up, 13% of resource use data and 8.3% of utility scores were missing between baseline and 24 months. Although loss to follow-up could be regarded as being within acceptable limits [334], it is a limitation. Missing values and cross-overs were also within acceptable limits for a RCT, but are nevertheless problematic in all studies because it introduces additional uncertainty. Uncertainty in this cost-effectiveness study was handled formally in the sensitivity analysis. Three of the five sensitivity analyses performed changed the conclusions about cost-effectiveness: Using SF-6D instead of EQ-5D, using per protocol instead of ITT, and not using multiple imputations. Five patients crossed over from MDR and underwent TDR but no one the other way. Surprisingly, the per-protocol analysis rendered surgery not cost-effective. The reason for this could be found in the 5 patients who crossed over. They were extremely costly compared to other patients and had a low gain in QALY's. For this reason, the ICER was changed from € 39 772/QALY to € 86 712/QALY. Arguably, patients who do not follow their randomized

treatment are in one way or another different from the rest of their treatment group. These differences can be observed, for example if patients exhibit observable or unobservable comorbidities. For this reason, ITT analysis is considered the gold standard in health economic evaluations [335].

Multiple imputation (MI) was used to avoid excluding patients with a few missing values and to avoid bias [278,279,281]. The considerable loss of data not using MI could partly be explained by the nature of the outcome effect variable: EQ-5D is a scale variable comprised of 5 items. If an item is missing, the scale cannot be computed and the utility score for this patient will be missing. Standard regression analysis were used to provide estimates of the missing data conditional on complete variables in the analysis[278,336]. The complete variables included were intervention group, age, and sex. A "richer" model including more complete variables may have increased the precision of the imputated estimates.

5.2.3 Biomechanical changes after total disc replacement

No significant change in movement in the sagittal plane from baseline (pre-surgery) to two years (post-surgery) was found in segments with TDR. The clinical importance of the significant difference of 2,88° (p = 0.041, 95%CI: 0.11° to 5.25°) at the L4/L5 level between treatment groups are questionable. At the L5/S1 level , segmental motion in spinal levels with implanted disc prosthesis did not differ significantly from same levels with a natural course. Mean range of motion (SD) in our study in segments with implanted prosthesis was 7.73 ° (6.46°) at the L4/L5 level and 6.20° (4.82°) at the L5/S1 level. Ziegler et al report that ROM for ProDisc L© averaged 7.7° two years after surgery but did not report the measurement method or measurement error [306]. Siepe et al reported a decrease in ROM from 8.1° preoperative to 5.4° postoperative with ProDisc II [337]. Using the DCRA method , Leivseth et al in a study from 2006 reported a range of motion less than 45% of the normal range two years postoperatively in the ProDisc II prosthesis [338]. Here, the postoperative ROM was 8.0° and 3.5° at the L4/L5 and L5/S1 level respectively. Although a different prosthesis design, Guyer et al reported no significant change in mean ROM two years after surgery with the Charité prosthesis and a postoperative ROM in instrumented segments of 6.0°[339]. In a review study of 9 studies reporting outcome after TDR mainly with the use of the SB Charitè prosthesis from 2003, de Kleuver[155] noted that for those studies that reported mobility, the motion seemed to move with a reported average range of motion of 5°–12°. However, mobility of the motion segment was frequently lost in these studies either

because of surgical arthrodesis performed as a result of clinical failure or because of spontaneous fusion with a bone ridge over the prosthesis.

Although segmental range of motion after total disc replacement is a central characteristic of disc prosthesis, there seems to be a lack of consensus when reporting this in clinical studies. None of the aforementioned studies used the same method for measurement of ROM and disc heights. Therefore, direct comparisons across studies are difficult.

Although clinical trials that involve total disc replacement report segmental range of motion as outcome criteria[306,339,340], the association between postoperative mobility and clinical outcome has not been unequivocally established. Huang et al. reported a positive but weak correlation between range of motion and several clinical outcomes [227]. Using the same statistical methods for comparing clinical outcome with range of motion, we were unable to reproduce the results of Huang et al in study III. We conclude that total disc replacement did not lead to increased lumbar motion compared to the natural history of degenerative disc disease and while segmental motion increased marginally but non-significantly and disc height increased significantly in the prosthesis cohort, these two parameters remained virtually unchanged in the rehabilitation cohort. Nevertheless, clinical outcomes improved in the rehabilitation as well as the prosthesis cohort, though to a lesser extent. This observation suggests that segmental motion and disc height are not major determinants of clinical outcome.

We suggest four possible explanations for the low mobility of the segments instrumented with disc prostheses found in study III:

Surgical procedure. When implanting disc prosthesis in lumbar segments, release of fibrotic and contracted anterior structures (ligaments, annulus fibrosus) occurs. The extent of the release of the posterior structures is more open to variation. While motion of the prosthesis is semiconstrained, the stiffness of the posterior tissues might hinder flexion of the segment-prosthesis complex.

Location of the axis of rotation. Restriction of motion would be expected if the axis of rotation of the artificial joint did not coincide with the physiological axis of rotation. In situ, the centre of rotation of the ProDisc prosthesis is located approximately 5 mm below the cranial endplate of the caudal vertebra. Thus there is a small misalignment of the prosthesis axis with respect to the physiological axes of rotation of the L4/L5 and L5/S1 joints as determined by Penning et al. [341]

Disc height. In L5/S1 segments instrumented with disc prostheses, the change in disc space height from pre- to post-treatment amounted to 3.2 SD on average. For L5/S1 one SD equals approximately 10% of disc height or approximately 1.2 mm.[11] Thus a 3.2 SD amounts to approximately 4 mm, equivalent to 1/3 of the normal disc height. It is conceivable that the considerable increase of the disc space after implantation could cause abnormal strain of the ligaments of the zygapophyseal joints and thus impede motion. However, there was no clear relationship found between post-treatment disc height and the range of segmental motion as documented in this study.

Natural course. A forth reason for the low mobility of instrumented segments could be the natural course of the degenerative disease itself. Several studies report decreased segmental ROM at the most severe stages of disc degeneration[146].

What remains unanswered in our study is if the small range of motion in the prosthesis after two years could prevent adjacent level disease (if this is an issue). In a study of segmental motion in 155 implanted ProDisc L prostheses, Auerbach[342] et al showed that patients with TDR lost slight relative contribution to total lumbar motion from the operative level which was mostly compensated for by the caudal adjacent level. In a separate study, we showed that ALD was observed at similar frequencies in the two treatment groups in this study at the 2-year follow-up. However, the surgery group had increased facet arthropathy at the implant level [343].

Although the development of adjacent level disease is a central issue in fusion and TDR surgery, it is likely that degenerative disc disease is a multisegmental, progressive disease caused by a genetic predisposition or a tissue response to an insult or altered mechanical environment[17,18]. The primary rationale for inserting a mobility preserving device, the prevention of development of adjacent level disease[171], thus remains controversial. The phenomenon called Adjacent Level Disease could be dependent on which model is used to analyse range of motion [344,345]. A recent study using non-linear finite element model suggests that fusion surgery with stand-alone interbody cage could even reduce stress on adjacent levels[181] .

5.2.4 Difference in efficacy measures of health in economical trials

To assess if surgery with total disc replacement could be cost effective, we used several statistical and graphical methods that are well recognized, up to date and recommended by the Norwegian authorities [193,198,346]. Using these methods, it was difficult to assess if surgery with

total disc replacement could be cost effective because this varied with the use of clinical effectiveness outcome measurement. The discrepancy between the two utility indexes revealed in study II and IV could partly be explained by the method of which utilities are estimated. Two important differences between these two techniques are often mentioned: First, using the SG (Standard Gamble - SG) method, people are asked to take a risk while the TTO (Time Trade Off – TTO) are riskless. Second, while the TTO method assumes that utility is linear in duration, the SG method has no restriction on the utility function for the duration of the health state [347,348]. Partly as a result of this, the health states utilities of EQ5-D (TTO) and SF6-D (SG) show systematically different utilities [193,198,348-350]. The TTO method tends to produce lower utility values than the SG method)[351]. This could be observed looking at the baseline utility values of the present study. Also, a floor effect of SF-6D and a ceiling effect of EQ5-D has been observed [352]. This could also be observed in the Rasch analysis in study 4.

No standardized recommendation for the use of utility indices exists. However, several studies report that assessment of cost-effectiveness could be dependent on the index used as confirmed in study II and IV by our study[267,353-355]. The problem with choice of utility index and outcome of a cost utility analysis exists also in other clinical settings[356]. In a paper from 2011, Whitehurst et al discuss the differences of EQ-5D and SF-6D on a group mean score level [357]. They found that the same pattern of difference between the two utility indexes can be found on a group level as well as on an individual level: The SF-6D provides higher score values for poorer health states while EQ-5D provides higher score values for milder health states and that the two indexes cannot be used interchangeably. The practical consequences of the latter is that the effect of treatments could be overestimated and new treatments falsely accepted as cost effective when they are not [355,358].

One of the main issues of reporting health economic outcomes in clinical trials is to inform decision makers so that priorities of scarce health resources can be made [194]. In order to compare and prioritize across disease conditions, it is important to use a generic as opposed to a disease specific outcome measure [207,237]. A central question becomes this:

What is the pay-off for using generic as opposed to disease specific instruments?

The EQ-5D is one of the most widely used generic instruments for assessing efficacy in cost-effectiveness studies and specifically, it is thought to be adequate when reporting efficacy in health economic trials of CLBP patients [255,257,359]. The validity and reliability of EQ5-D generally

and for back pain patients specifically is documented in several papers [360-362]. For this reason, the EQ-5D was chosen as the main outcome variable in study II. The SF6-D utility index is based on the SF-36 questionnaire for generic health related quality of life [206]. The SF-36 is known to have relatively good responsiveness properties (ability to detect change in health status after treatment) when compared to disease specific questionnaires like the ODI [236]. Specifically, the combined pain and function scale from SF-36 have been reported to have a high level of sensitivity and specificity in identifying improvement after treatment [236,255,256]. Djurasovic[363] in a recent study of outcome after fusion surgery showed that the health dimensions pain intensity, walking, and social life measured by the ODI, best predicted improvement in overall health-related quality of life, as measured by using the physical component score of the SF-36. Generic instruments in general are known to be less sensitive to change of health conditions after treatment compared to disease specific instruments [236,237,364]. The more specific the outcomes tool, the more sensitive the response [255]. In study IV, we found that the responsiveness was higher for the SF-6D than the EQ-5D. However, the criterion validity for the indices (their similarity with the gold standard, the ODI) were lower for the SF-6D than the EQ- 5D. In comparative studies like the present, the ability to detect if the treatment has caused a change may be the most important property of an instrument. This might favor the SF-6D in such studies.

6 CONCLUSIONS

We found a statistical significant difference in the main outcome variable (ODI) in favor of disc prosthesis surgery when compared to non-surgical treatment. However, we were not able to verify that this difference was clinically meaningful. A clinical significant improvement in quality of life related to low back pain was found in both treatment groups in the per protocol analysis using a mixed model analysis. Disc prosthesis surgery is subject to potentially severe complications and our results indicate that non-surgical treatment should be the first treatment option.

Total disc replacement surgery was cost effective compared non-surgical treatment using EQ-5D as efficacy outcome. However, cost-effectiveness in this study was dependent on the utility index used to evaluate effectiveness. When SF-6D was used as an efficacy outcome variable, disc prosthesis surgery was not cost-effective. This introduces an element of arbitrariness when assessing cost-effectiveness in cost-utility studies and underscores the need for a clear statement from health authorities to guide further health economic trials.

In segments with inserted prosthesis, we were not able to show any significant change in segmental mobility from preoperative to two year postoperative. There was no significant difference in segmental motion comparing operated segments with segments at the same levels in the non-surgical treatment group. Furthermore, no correlation was found between movement in the prosthesis and clinical outcome.

The difference in important measurement properties between two of the most used measures of effectiveness (utility indices) in health economic trials were too large to consider them interchangeable. Because it could have a great impact on the probability of acceptance of a new treatment as cost-effective or not, this underscores the importance of choosing the right clinical outcome measurement when reporting results in cost-effectiveness studies.

Future research

In a recent double-blind RCT Albert et al [365] tested the efficacy of antibiotic treatment in patients with chronic low back pain ([6 months) and Modic type 1 changes in the IVD. Patients were randomized to either 100 days of antibiotic treatment (Bioclavid) or placebo and were blindly evaluated at baseline; end of treatment and at 1-year follow-up. The antibiotic group improved highly statistically significant on all outcome measures and improvement continued from 100 days follow-up until 1-year follow-up. However, further studies are needed to confirm this. A Norwegian study is being planned randomizing patients either to antibiotics or to palcebo (Bjørn Skogstad, personal communication). If confirmed, this opens up for a radically new perspective in the treatment of degenerative disc disease and low back pain.

Biological intervertebral disc replacement in the field of tissue engineering is another approach to the problem [366]. This includes the introduction of functional cells and supporting biomaterials with cells, cells plus biomaterial or biomaterial alone into the degenerate disc[366-369]. Inkjet printing of 3-D hydrogel structures used as scaffolds for various cell types are also described[370-372]. Here, a soft scaffold is fabricated according to a computer-aided design template using a single device in what is called "computer aided scaffold topology design"[373]. It should be noted that studies in this field is partly motivated by the notion that total disc replacement surgery is "…very traumatic and that the non-biological prosthesis wears with time"[366] and also the believe that fusion surgery can cause adjacent level disease[160,366,374].

Novel MRI techniques, such as quantitative MRI, T1ρ MRI, sodium MRI and nuclear magnetic resonance spectroscopy, are more sensitive in quantifying the biochemical changes of disc degeneration, as measured by alteration in collagen structure, as well as water and proteoglycan loss [24,375]. These methods may be helpful in the struggle to establish further the connection between degenerative disc disease and low back pain.

Finally, in order to evaluate the long term effect of TDR, we are now conducting an 8-year follow up of the patients in our study [376] in terms of clinical results, costs, reoperation- and revision rate, degenerative changes and prognostic factors. Hopefully, this will provide further valuable information to the choice of treatment for patients with degenerative disc disease and low back pain.

REFERENCES

1. Lærum E, Brox JI, Storheim K, Espeland A. Korsryggsmerter: med og uten nerverotaffeksjon. Oslo: Formi; 2007.
2. van Tulder M, Becker A, Bekkering T, et al. Chapter 3. European guidelines for the management of acute nonspecific low back pain in primary care. Eur Spine J 2006;15 Suppl 2:S169-91.
3. Airaksinen O, Brox JI, Cedraschi C, et al. Chapter 4. European guidelines for the management of chronic nonspecific low back pain. Eur Spine J 2006;15 Suppl 2:S192-300.
4. Balague F, Mannion AF, Pellise F, Cedraschi C. Non-specific low back pain. Lancet 2012;379:482-91.
5. Hoy D, Brooks P, Blyth F, Buchbinder R. The Epidemiology of low back pain. Best Pract Res Clin Rheumatol 2010;24:769-81.
6. Vos T, Flaxman AD, Naghavi M, et al. Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet 2012;380:2163-96.
7. Ihlebaek C, Hansson TH, Laerum E, et al. Prevalence of low back pain and sickness absence: a "borderline" study in Norway and Sweden. Scandinavian journal of public health 2006;34:555-8.
8. Hagen KB, Thune O. Work incapacity from low back pain in the general population. Spine (Phila Pa 1976) 1998;23:2091-5.
9. Walsh K, Cruddas M, Coggon D. Low back pain in eight areas of Britain. J Epidemiol Community Health 1992;46:227-30.
10. Freburger JK, Holmes GM, Agans RP, et al. The rising prevalence of chronic low back pain. Arch Intern Med 2009;169:251-8.
11. Waddell G. 1987 Volvo award in clinical sciences. A new clinical model for the treatment of low-back pain. Spine (Phila Pa 1976) 1987;12:632-44.
12. Meucci RD, Fassa AG, Paniz VM, Silva MC, Wegman DH. Increase of chronic low back pain prevalence in a medium-sized city of southern Brazil. BMC Musculoskelet Disord 2013;14:1471-2474.
13. van Tulder MW, Koes B, Seitsalo S, Malmivaara A. Outcome of invasive treatment modalities on back pain and sciatica: an evidence-based review. Eur Spine J 2006;15 Suppl 1:S82-92.
14. Allan DB, Waddell G. An historical perspective on low back pain and disability. Acta Orthop Scand Suppl 1989;234:1-23.
15. Ehrlich GE. Low back pain. Bull World Health Organ 2003;81:671-6.
16. Harkness EF, Macfarlane GJ, Silman AJ, McBeth J. Is musculoskeletal pain more common now than 40 years ago?: Two population-based cross-sectional studies. Rheumatology (Oxford) 2005;44:890-5.
17. Roberts S, Evans H, Trivedi J, Menage J. Histology and pathology of the human intervertebral disc. J Bone Joint Surg Am 2006;88 Suppl 2:10-4.
18. Urban J, Roberts S. Degeneration of the intervertebral disc. Arthritis research & therapy 2003;5:120 - 30.
19. Marchand F, Ahmed A. Investigation of the laminate structure of lumbar disc anulus fibrosus. Spine 1990;15:402 - 10.
20. Yu J. Elastic tissues of the intervertebral disc. Biochem Soc Trans 2002;30:848-52.
21. Maroudas A, Stockwell RA, Nachemson A, Urban J. Factors involved in the nutrition of the human lumbar intervertebral disc: cellularity and diffusion of glucose in vitro. J Anat 1975;120:113-30.
22. Roberts S, Menage J, Urban J. Biochemical and structural properties of the cartilage end-plate and its relation to the intervertebral disc. Spine 1989;14:166 - 74.
23. Urban JP, Smith S, Fairbank JC. Nutrition of the intervertebral disc. Spine (Phila Pa 1976) 2004;29:2700-9.

24. Urban JP, Winlove CP. Pathophysiology of the intervertebral disc and the challenges for MRI. J Magn Reson Imaging 2007;25:419-32.
25. Roberts S, Menage J, Eisenstein S. The cartilage end-plate and intervertebral disc in scoliosis: calcification and other sequelae. J Orthop Res 1993;11:747 - 57.
26. Roberts S, Urban J, Evans H, Eisenstein S. Transport properties of the human cartilage end-plate in relation to its composition and calcification. Spine 1996;21:415 - 20.
27. Urban J, Holm S, Maroudas A. Diffusion of small solutes into the intervertebral disc: as in vivo study. Biorheology 1978;15:203 - 21.
28. Urban M, Fairbank J, Etherington P, Loh F, Winlove C, Urban J. Electrochemical measurement of transport into scoliotic intervertebral discs in vivo using nitrous oxide as a tracer. Spine 2001;26:984 - 90.
29. Chan SC, Ferguson SJ, Gantenbein-Ritter B. The effects of dynamic loading on the intervertebral disc. Eur Spine J 2011;20:1796-812.
30. Roughley PJ, Melching LI, Heathfield TF, Pearce RH, Mort JS. The structure and degradation of aggrecan in human intervertebral disc. Eur Spine J 2006;15 Suppl 3:S326-32.
31. Johnstone B, Bayliss M. The large proteoglycans of the human intervertebral disc. Changes in their biosynthesis and structure with age, topography, and pathology. Spine 1995;20:674 - 84.
32. Wilkins RJ, Browning JA, Urban JP. Chondrocyte regulation by mechanical load. Biorheology 2000;37:67-74.
33. Boos N, Weissbach S, Rohrbach H, Weiler C, Spratt K, Nerlich A. Classification of age-related changes in lumbar intervertebral discs: 2002 Volvo Award in basic science. Spine 2002;27:2631 - 44.
34. Crean J, Roberts S, Jaffray D, Eisenstein S, Duance V. Matrix metalloproteinases in the human intervertebral disc: role in disc degeneration and scoliosis. Spine 1997;22:2877 - 84.
35. Goupille P, Jayson MI, Valat JP, Freemont AJ. Matrix metalloproteinases: the clue to intervertebral disc degeneration? Spine (Phila Pa 1976) 1998;23:1612-26.
36. Roughley PJ. Biology of intervertebral disc aging and degeneration: involvement of the extracellular matrix. Spine (Phila Pa 1976) 2004;29:2691-9.
37. Kang JD, Stefanovic-Racic M, McIntyre LA, Georgescu HI, Evans CH. Toward a biochemical understanding of human intervertebral disc degeneration and herniation. Contributions of nitric oxide, interleukins, prostaglandin E2, and matrix metalloproteinases. Spine (Phila Pa 1976) 1997;22:1065-73.
38. Weiler C, Nerlich AG, Zipperer J, Bachmeier BE, Boos N. 2002 SSE Award Competition in Basic Science: expression of major matrix metalloproteinases is associated with intervertebral disc degradation and resorption. Eur Spine J 2002;11:308-20.
39. Roberts S, Caterson B, Menage J, Evans E, Jaffray D, Eisenstein S. Matrix metalloproteinases and aggrecanase: their role in disorders of the human intervertebral disc. Spine 2000;25:3005 - 13.
40. Antoniou J, Steffen T, Nelson F, et al. The human lumbar intervertebral disc: evidence for changes in the biosynthesis and denaturation of the extracellular matrix with growth, maturation, ageing, and degeneration. J Clin Invest 1996;98:996-1003.
41. Adams MA, Stefanakis M, Dolan P. Healing of a painful intervertebral disc should not be confused with reversing disc degeneration: implications for physical therapies for discogenic back pain. Clin Biomech (Bristol, Avon) 2010;25:961-71.
42. Roberts S. Disc morphology in health and disease. Biochem Soc Trans 2002;30:864-9.
43. Inoue N, Espinoza Orias AA. Biomechanics of intervertebral disk degeneration. Orthop Clin North Am 2011;42:487-99, vii.
44. Frino J, McCarthy RE, Sparks CY, McCullough FL. Trends in adolescent lumbar disk herniation. J Pediatr Orthop 2006;26:579-81.
45. Simmons ED, Jr., Guntupalli M, Kowalski JM, Braun F, Seidel T. Familial predisposition for degenerative disc disease. A case-control study. Spine (Phila Pa 1976) 1996;21:1527-9.
46. Varlotta G, Brown M, Kelsey J, Golden A. Familial predisposition for herniation of a lumbar disc in patients who are less than twenty-one years old. J Bone Joint Surg Am 1991;73:124 - 8.

47. Matsui H, Kanamori M, Ishihara H, Yudoh K, Naruse Y, Tsuji H. Familial predisposition for lumbar degenerative disc disease. A case-control study. Spine 1998;23:1029 - 34.
48. Matsui H, Terahata N, Tsuji H, Hirano N, Naruse Y. Familial predisposition and clustering for juvenile lumbar disc herniation. Spine (Phila Pa 1976) 1992;17:1323-8.
49. Kao PY, Chan D, Samartzis D, Sham PC, Song YQ. Genetics of lumbar disk degeneration: technology, study designs, and risk factors. Orthop Clin North Am 2011;42:479-86, vii.
50. Waddell G. The back pain revolution. Edinburgh: Churchill Livingstone; 1998.
51. Sambrook PN, MacGregor AJ, Spector TD. Genetic influences on cervical and lumbar disc degeneration: a magnetic resonance imaging study in twins. Arthritis Rheum 1999;42:366-72.
52. Livshits G, Popham M, Malkin I, et al. Lumbar disc degeneration and genetic factors are the main risk factors for low back pain in women: the UK Twin Spine Study. Ann Rheum Dis 2011;70:1740-5.
53. MacGregor AJ, Andrew T, Sambrook PN, Spector TD. Structural, psychological, and genetic influences on low back and neck pain: a study of adult female twins. Arthritis Rheum 2004;51:160-7.
54. Podichetty VK. The aging spine: the role of inflammatory mediators in intervertebral disc degeneration. Cell Mol Biol (Noisy-le-grand) 2007;53:4-18.
55. Beynon R, Sterne JA, Wilcock G, et al. Is MRI better than CT for detecting a vascular component to dementia? A systematic review and meta-analysis. BMC neurology 2012;12:33.
56. Sudo H, Yamada K, Iwasaki K, et al. Global identification of genes related to nutrient deficiency in intervertebral disc cells in an experimental nutrient deprivation model. PloS one 2013;8:e58806.
57. Taher F, Essig D, Lebl DR, et al. Lumbar degenerative disc disease: current and future concepts of diagnosis and management. Advances in orthopedics 2012;2012:970752.
58. Kalb S, Martirosyan NL, Kalani MY, Broc GG, Theodore N. Genetics of the degenerated intervertebral disc. World neurosurgery 2012;77:491-501.
59. Battie MC, Levalahti E, Videman T, Burton K, Kaprio J. Heritability of lumbar flexibility and the role of disc degeneration and body weight. J Appl Physiol 2008;104:379-85.
60. Battie MC, Videman T, Levalahti E, Gill K, Kaprio J. Genetic and environmental effects on disc degeneration by phenotype and spinal level: a multivariate twin study. Spine (Phila Pa 1976) 2008;33:2801-8.
61. Edgar MA. The nerve supply of the lumbar intervertebral disc. J Bone Joint Surg Br 2007;89:1135-9.
62. Coppes MH, Marani E, Thomeer RT, Groen GJ. Innervation of "painful" lumbar discs. Spine (Phila Pa 1976) 1997;22:2342-9; discussion 9-50.
63. Takahashi K, Aoki Y, Ohtori S. Resolving discogenic pain. Eur Spine J 2008;17 Suppl 4:428-31.
64. Ozawa T, Ohtori S, Inoue G, Aoki Y, Moriya H, Takahashi K. The degenerated lumbar intervertebral disc is innervated primarily by peptide-containing sensory nerve fibers in humans. Spine (Phila Pa 1976) 2006;31:2418-22.
65. Freemont AJ, Peacock TE, Goupille P, Hoyland JA, O'Brien J, Jayson MI. Nerve ingrowth into diseased intervertebral disc in chronic back pain. Lancet 1997;350:178-81.
66. Palmgren T, Gronblad M, Virri J, Kaapa E, Karaharju E. An immunohistochemical study of nerve structures in the anulus fibrosus of human normal lumbar intervertebral discs. Spine (Phila Pa 1976) 1999;24:2075-9.
67. Lawson SN, Crepps BA, Perl ER. Relationship of substance P to afferent characteristics of dorsal root ganglion neurones in guinea-pig. The Journal of physiology 1997;505 (Pt 1):177-91.
68. Burke JG, Watson RW, McCormack D, Dowling FE, Walsh MG, Fitzpatrick JM. Intervertebral discs which cause low back pain secrete high levels of proinflammatory mediators. J Bone Joint Surg Br 2002;84:196-201.
69. Zhang YG, Guo TM, Guo X, Wu SX. Clinical diagnosis for discogenic low back pain. International journal of biological sciences 2009;5:647-58.

70. Hayashi S, Taira A, Inoue G, et al. TNF-alpha in nucleus pulposus induces sensory nerve growth: a study of the mechanism of discogenic low back pain using TNF-alpha-deficient mice. Spine (Phila Pa 1976) 2008;33:1542-6.
71. Schaible HG, Schmidt RF. Effects of an experimental arthritis on the sensory properties of fine articular afferent units. J Neurophysiol 1985;54:1109-22.
72. Weinstein J, Claverie W, Gibson S. The pain of discography. Spine (Phila Pa 1976) 1988;13:1344-8.
73. Hurri H, Karppinen J. Discogenic pain. Pain 2004;112:225-8.
74. Karppinen J, Shen FH, Luk KD, Andersson GB, Cheung KM, Samartzis D. Management of degenerative disk disease and chronic low back pain. Orthop Clin North Am 2011;42:513-28, viii.
75. Manchikanti L, Glaser SE, Wolfer L, Derby R, Cohen SP. Systematic review of lumbar discography as a diagnostic test for chronic low back pain. Pain Physician 2009;12:541-59.
76. DeLeo JA. Basic science of pain. J Bone Joint Surg Am 2006;88 Suppl 2:58-62.
77. Latremoliere A, Woolf CJ. Central sensitization: a generator of pain hypersensitivity by central neural plasticity. The journal of pain : official journal of the American Pain Society 2009;10:895-926.
78. Waddell G, Main CJ, Morris EW, Di Paola M, Gray IC. Chronic low-back pain, psychologic distress, and illness behavior. Spine (Phila Pa 1976) 1984;9:209-13.
79. Villemure C, Bushnell MC. Cognitive modulation of pain: how do attention and emotion influence pain processing? Pain 2002;95:195-9.
80. Wiech K, Tracey I. The influence of negative emotions on pain: Behavioral effects and neural mechanisms. Neuroimage 2009;47:987-94.
81. Williams FM, Sambrook PN. Neck and back pain and intervertebral disc degeneration: role of occupational factors. Best Pract Res Clin Rheumatol 2011;25:69-79.
82. Battie M, Haynor D, Fisher L, Gill K, Gibbons L, Videman T. Similarities in degenerative findings on magnetic resonance images of the lumbar spines of identical twins. J Bone Joint Surg Am 1995;77:1662 - 70.
83. Battie MC, Videman T, Gibbons LE, et al. Occupational driving and lumbar disc degeneration: a case-control study. Lancet 2002;360:1369-74.
84. Jorgensen MB, Korshoj M, Lagersted-Olsen J, et al. Physical activities at work and risk of musculoskeletal pain and its consequences: protocol for a study with objective field measures among blue-collar workers. BMC Musculoskelet Disord 2013;14:213.
85. Kwon BK, Roffey DM, Bishop PB, Dagenais S, Wai EK. Systematic review: occupational physical activity and low back pain. Occup Med (Lond) 2011;61:541-8.
86. Harkness EF, Macfarlane GJ, Nahit ES, Silman AJ, McBeth J. Risk factors for new-onset low back pain amongst cohorts of newly employed workers. Rheumatology (Oxford) 2003;42:959-68.
87. Manek NJ, MacGregor AJ. Epidemiology of back disorders: prevalence, risk factors, and prognosis. Curr Opin Rheumatol 2005;17:134-40.
88. de Schepper EI, Damen J, van Meurs JB, et al. The association between lumbar disc degeneration and low back pain: the influence of age, gender, and individual radiographic features. Spine (Phila Pa 1976) 2010;35:531-6.
89. Kaila-Kangas L, Kivimaki M, Riihimaki H, Luukkonen R, Kirjonen J, Leino-Arjas P. Psychosocial factors at work as predictors of hospitalization for back disorders: a 28-year follow-up of industrial employees. Spine (Phila Pa 1976) 2004;29:1823-30.
90. Dionne C, Koepsell TD, Von Korff M, Deyo RA, Barlow WI, Checkoway H. Formal education and back-related disability. In search of an explanation. Spine (Phila Pa 1976) 1995;20:2721-30.
91. Shiri R, Karppinen J, Leino-Arjas P, Solovieva S, Viikari-Juntura E. The association between smoking and low back pain: a meta-analysis. Am J Med 2010;123:87 e7-35.
92. Holm S, Nachemson A. Nutrition of the intervertebral disc: acute effects of cigarette smoking. An experimental animal study. Ups J Med Sci 1988;93:91 - 9.
93. Shiri R, Karppinen J, Leino-Arjas P, Solovieva S, Viikari-Juntura E. The association between obesity and low back pain: a meta-analysis. Am J Epidemiol 2010;171:135-54.

94. Kurunlahti M, Tervonen O, Vanharanta H, Ilkko E, Suramo I. Association of atherosclerosis with low back pain and the degree of disc degeneration. Spine (Phila Pa 1976) 1999;24:2080-4.
95. Kauppila L. Prevalence of stenotic changes in arteries supplying the lumbar spine. A postmortem angiographic study on 140 subjects. Ann Rheum Dis 1997;56:591 - 5.
96. Kauppila L, McAlindon T, Evans S, Wilson P, Kiel D, Felson D. Disc degeneration/back pain and calcification of the abdominal aorta. A 25-year follow-up study in Framingham. Spine 1997;22:1642 - 7.
97. Kauppila LI. Atherosclerosis and disc degeneration/low-back pain--a systematic review. Eur J Vasc Endovasc Surg 2009;37:661-70.
98. Linton SJ. A review of psychological risk factors in back and neck pain. Spine (Phila Pa 1976) 2000;25:1148-56.
99. Linton SJ. Do psychological factors increase the risk for back pain in the general population in both a cross-sectional and prospective analysis? Eur J Pain 2005;9:355-61.
100. Andersson GB. Epidemiological features of chronic low-back pain. Lancet 1999;354:581-5.
101. Borenstein DG, O'Mara JW, Jr., Boden SD, et al. The value of magnetic resonance imaging of the lumbar spine to predict low-back pain in asymptomatic subjects : a seven-year follow-up study. J Bone Joint Surg Am 2001;83-A:1306-11.
102. Jensen MC, Brant-Zawadzki MN, Obuchowski N, Modic MT, Malkasian D, Ross JS. Magnetic resonance imaging of the lumbar spine in people without back pain. N Engl J Med 1994;331:69-73.
103. van Middelkoop M, Rubinstein SM, Kuijpers T, et al. A systematic review on the effectiveness of physical and rehabilitation interventions for chronic non-specific low back pain. Eur Spine J 2011;20:19-39.
104. Willems P, de Bie R, Oner C, Castelein R, de Kleuver M. Clinical decision making in spinal fusion for chronic low back pain. Results of a nationwide survey among spine surgeons. BMJ open 2011;1:e000391.
105. Willems P. Decision making in surgical treatment of chronic low back pain: the performance of prognostic tests to select patients for lumbar spinal fusion. Acta orthopaedica Supplementum 2013;84:1-35.
106. Waddell G. Subgroups within "nonspecific" low back pain. J Rheumatol 2005;32:395-6.
107. Spitzer. Scientific approach to the assessment and management of activity-related spinal disorders. A monograph for clinicians. Report of the Quebec Task Force on Spinal Disorders. Spine (Phila Pa 1976) 1987;12:S1-59.
108. Hill JC, Whitehurst DGT, Lewis M, et al. Comparison of stratified primary care management for low back pain with current best practice (STarT Back): a randomised controlled trial. The Lancet;378:1560-71.
109. Masharawi Y, Kjaer P, Bendix T, et al. The reproducibility of quantitative measurements in lumbar magnetic resonance imaging of children from the general population. Spine (Phila Pa 1976) 2008;33:2094-100.
110. Benneker LM, Heini PF, Anderson SE, Alini M, Ito K. Correlation of radiographic and MRI parameters to morphological and biochemical assessment of intervertebral disc degeneration. Eur Spine J 2005;14:27-35.
111. Frymoyer JW, Newberg A, Pope MH, Wilder DG, Clements J, MacPherson B. Spine radiographs in patients with low-back pain. An epidemiological study in men. J Bone Joint Surg Am 1984;66:1048-55.
112. Pye SR, Reid DM, Smith R, et al. Radiographic features of lumbar disc degeneration and self-reported back pain. J Rheumatol 2004;31:753-8.
113. Beattie PF, Meyers SP, Stratford P, Millard RW, Hollenberg GM. Associations between patient report of symptoms and anatomic impairment visible on lumbar magnetic resonance imaging. Spine (Phila Pa 1976) 2000;25:819-28.

114. Kjaer P, Leboeuf-Yde C, Korsholm L, Sorensen JS, Bendix T. Magnetic resonance imaging and low back pain in adults: a diagnostic imaging study of 40-year-old men and women. Spine (Phila Pa 1976) 2005;30:1173-80.
115. Andersson GB, Schultz A, Nathan A, Irstam L. Roentgenographic measurement of lumbar intervertebral disc height. Spine (Phila Pa 1976) 1981;6:154-8.
116. Raininko R, Manninen H, Battie MC, Gibbons LE, Gill K, Fisher LD. Observer variability in the assessment of disc degeneration on magnetic resonance images of the lumbar and thoracic spine. Spine (Phila Pa 1976) 1995;20:1029-35.
117. Frobin W, Brinckmann P, Kramer M, Hartwig E. Height of lumbar discs measured from radiographs compared with degeneration and height classified from MR images. Eur Radiol 2001;11:263-9.
118. Frobin W, Brinckmann P, Leivseth G, Biggemann M, Reikeras O. Precision measurement of segmental motion from flexion-extension radiographs of the lumbar spine. Clin Biomech (Bristol, Avon) 1996;11:457-65.
119. Endean A, Palmer KT, Coggon D. Potential of magnetic resonance imaging findings to refine case definition for mechanical low back pain in epidemiological studies: a systematic review. Spine (Phila Pa 1976) 2011;36:160-9.
120. Modic MT, Steinberg PM, Ross JS, Masaryk TJ, Carter JR. Degenerative disk disease: assessment of changes in vertebral body marrow with MR imaging. Radiology 1988;166:193-9.
121. Braithwaite I, White J, Saifuddin A, Renton P, Taylor BA. Vertebral end-plate (Modic) changes on lumbar spine MRI: correlation with pain reproduction at lumbar discography. Eur Spine J 1998;7:363-8.
122. Lipson SJ. Spinal-fusion surgery -- advances and concerns. N Engl J Med 2004;350:643-4.
123. Kjaer P, Korsholm L, Bendix T, Sorensen JS, Leboeuf-Yde C. Modic changes and their associations with clinical findings. Eur Spine J 2006;15:1312-9.
124. Cheung KM, Karppinen J, Chan D, et al. Prevalence and pattern of lumbar magnetic resonance imaging changes in a population study of one thousand forty-three individuals. Spine (Phila Pa 1976) 2009;34:934-40.
125. Aprill C, Bogduk N. High-intensity zone: a diagnostic sign of painful lumbar disc on magnetic resonance imaging. Br J Radiol 1992;65:361-9.
126. Schellhas KP, Pollei SR, Gundry CR, Heithoff KB. Lumbar disc high-intensity zone. Correlation of magnetic resonance imaging and discography. Spine (Phila Pa 1976) 1996;21:79-86.
127. Bohm B, Meinig H, Eckardt A, Schadmand-Fischer S, Heine J. [Correlation of degenerative intervertebral disk displacement using MRI with discography findings in patients with back pain]. Orthopade 2005;34:1144-9.
128. Peng B, Hou S, Wu W, Zhang C, Yang Y. The pathogenesis and clinical significance of a high-intensity zone (HIZ) of lumbar intervertebral disc on MR imaging in the patient with discogenic low back pain. Eur Spine J 2006;15:583-7.
129. Lam KS, Carlin D, Mulholland RC. Lumbar disc high-intensity zone: the value and significance of provocative discography in the determination of the discogenic pain source. Eur Spine J 2000;9:36-41.
130. Ito M, Incorvaia KM, Yu SF, Fredrickson BE, Yuan HA, Rosenbaum AE. Predictive signs of discogenic lumbar pain on magnetic resonance imaging with discography correlation. Spine (Phila Pa 1976) 1998;23:1252-8; discussion 9-60.
131. Carragee EJ. Is lumbar discography a determinate of discogenic low back pain: provocative discography reconsidered. Current review of pain 2000;4:301-8.
132. Carragee EJ, Paragioudakis SJ, Khurana S. 2000 Volvo Award winner in clinical studies: Lumbar high-intensity zone and discography in subjects without low back problems. Spine (Phila Pa 1976) 2000;25:2987-92.

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 (Phila Pa 1976) 1997;22:427-34.
134. Luoma K, Riihimaki H, Luukkonen R, Raininko R, Viikari-Juntura E, Lamminen A. Low back pain in relation to lumbar disc degeneration. Spine (Phila Pa 1976) 2000;25:487-92.
135. Tertti M, Paajanen H, Laato M, Aho H, Komu M, Kormano M. Disc degeneration in magnetic resonance imaging. A comparative biochemical, histologic, and radiologic study in cadaver spines. Spine (Phila Pa 1976) 1991;16:629-34.
136. Tertti M. Low field MRI in evaluation of intervertebral discs. Acta Radiol Suppl 1991;377:54-5.
137. Luoma K, Vehmas T, Riihimaki H, Raininko R. Disc height and signal intensity of the nucleus pulposus on magnetic resonance imaging as indicators of lumbar disc degeneration. Spine (Phila Pa 1976) 2001;26:680-6.
138. Elfering A, Semmer N, Birkhofer D, Zanetti M, Hodler J, Boos N. Risk factors for lumbar disc degeneration: a 5-year prospective MRI study in asymptomatic individuals. Spine (Phila Pa 1976) 2002;27:125-34.
139. Powell MC, Wilson M, Szypryt P, Symonds EM, Worthington BS. Prevalence of lumbar disc degeneration observed by magnetic resonance in symptomless women. Lancet 1986;2:1366-7.
140. Berg L, Neckelmann G, Gjertsen O, et al. Reliability of MRI findings in candidates for lumbar disc prosthesis. Neuroradiology 2012;54:699-707.
141. Hibbs RA. A Further Consideration of an Operation for Pott's Disease of the Spine: With Report of Cases from the Service of the New York Orthopaedic Hospital. Ann Surg 1912;55:682-8.
142. Bono CM, Garfin SR. History and evolution of disc replacement. Spine J 2004;4:145S-50S.
143. Howorth MB. Evolution of Spinal Fusion. Ann Surg 1943;117:278-89.
144. Barr JS. Ruptured intervertebral disc and sciatic pain. J Bone Joint Surg Am 1947;29:429-37.
145. Lee CK, Langrana NA. A review of spinal fusion for degenerative disc disease: need for alternative treatment approach of disc arthroplasty? Spine J 2004;4:173S-6S.
146. Deyo RA, Mirza SK. Trends and variations in the use of spine surgery. Clin Orthop Relat Res 2006;443:139-46.
147. Rajaee SS, Bae HW, Kanim LE, Delamarter RB. Spinal fusion in the United States: analysis of trends from 1998 to 2008. Spine (Phila Pa 1976) 2012;37:67-76.
148. Deyo RA, Nachemson A, Mirza SK. Spinal-fusion surgery - the case for restraint. N Engl J Med 2004;350:722-6.
149. 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 (Phila Pa 1976) 2013;38:E409-22.
150. Fritzell P, Hagg O, Wessberg P, Nordwall A. 2001 Volvo Award Winner in Clinical Studies: Lumbar fusion versus nonsurgical treatment for chronic low back pain: a multicenter randomized controlled trial from the Swedish Lumbar Spine Study Group. Spine 2001;26:2521-32; discussion 32-4.
151. Brox JI, Sorensen R, Friis A, et al. Randomized clinical trial of lumbar instrumented fusion and cognitive intervention and exercises in patients with chronic low back pain and disc degeneration. Spine (Phila Pa 1976) 2003;28:1913-21.
152. Brox JI, Reikeras O, Nygaard O, et al. Lumbar instrumented fusion compared with cognitive intervention and exercises in patients with chronic back pain after previous surgery for disc herniation: a prospective randomized controlled study. Pain 2006;122:145-55.
153. Fairbank J, Frost H, Wilson-MacDonald J, et al. Randomised controlled trial to compare surgical stabilisation of the lumbar spine with an intensive rehabilitation programme for patients with chronic low back pain: the MRC spine stabilisation trial. BMJ 2005;330:1233.
154. Martin BI, Mirza SK, Comstock BA, Gray DT, Kreuter W, Deyo RA. Reoperation rates following lumbar spine surgery and the influence of spinal fusion procedures. Spine (Phila Pa 1976) 2007;32:382-7.

155. de Kleuver M, Oner FC, Jacobs WC. Total disc replacement for chronic low back pain: background and a systematic review of the literature. Eur Spine J 2003;12:108-16.
156. Charnley J. Arthroplasty of the hip. A new operation. Lancet 1961;1:1129-32.
157. Charnley J. Surgery of the hip-joint: present and future developments. Br Med J 1960;1:821-6.
158. Ethgen O, Bruyere O, Richy F, Dardennes C, Reginster JY. Health-related quality of life in total hip and total knee arthroplasty. A qualitative and systematic review of the literature. J Bone Joint Surg Am 2004;86-A:963-74.
159. Van de Kelft E, Verguts L. Clinical outcome of mono segmental total disc replacement for lumbar disc disease with ball in socket prosthesis (Maverick(R)): Prospective study with 4 year follow-up. World neurosurgery 2011.
160. Park P, Garton HJ, Gala VC, Hoff JT, McGillicuddy JE. Adjacent segment disease after lumbar or lumbosacral fusion: review of the literature. Spine (Phila Pa 1976) 2004;29:1938-44.
161. Lee CK, Langrana NA. Lumbosacral spinal fusion. A biomechanical study. Spine (Phila Pa 1976) 1984;9:574-81.
162. Axelsson P, Johnsson R, Stromqvist B. The spondylolytic vertebra and its adjacent segment. Mobility measured before and after posterolateral fusion. Spine (Phila Pa 1976) 1997;22:414-7.
163. Weinhoffer SL, Guyer RD, Herbert M, Griffith SL. Intradiscal pressure measurements above an instrumented fusion. A cadaveric study. Spine (Phila Pa 1976) 1995;20:526-31.
164. Fernstrom U. Arthroplasty with intercorporal endoprothesis in herniated disc and in painful disc. Acta Chir Scand Suppl 1966;357:154-9.
165. Marshman LA, Friesem T, Rampersaud YR, Le Huec JC, Krishna M. Subsidence and malplacement with the Oblique Maverick Lumbar Disc Arthroplasty: technical note. Spine J 2008;8:650-5.
166. Siemionow KB, Hu X, Lieberman IH. The Fernstrom ball revisited. Eur Spine J 2012;21:443-8.
167. Szpalski M, Gunzburg R, Mayer M. Spine arthroplasty: a historical review. Eur Spine J 2002;11 Suppl 2:S65-84.
168. Tropiano P, Huang RC, Girardi FP, Cammisa FP, Jr., Marnay T. Lumbar total disc replacement. Seven to eleven-year follow-up. J Bone Joint Surg Am 2005;87:490-6.
169. Buttner-Janz K, Schellnack K. [Principle and initial results with the Charite Modular type SB cartilage disk endoprosthesis]. Magy Traumatol Orthop Helyreallito Seb 1988;31:136-40.
170. Huang RC, Girardi FP, Cammisa Jr FP, Tropiano P, Marnay T. Long-term flexion-extension range of motion of the prodisc total disc replacement. J Spinal Disord Tech 2003;16:435-40.
171. Errico TJ. Lumbar disc arthroplasty. Clin Orthop Relat Res 2005:106-17.
172. Mayer HM. Total lumbar disc replacement. J Bone Joint Surg Br 2005;87:1029-37.
173. Yajun W, Yue Z, Xiuxin H, Cui C. A meta-analysis of artificial total disc replacement versus fusion for lumbar degenerative disc disease. Eur Spine J 2010;19:1250-61.
174. Lindley EM, McBeth ZL, Henry SE, et al. Retrograde ejaculation after anterior lumbar spine surgery. Spine (Phila Pa 1976) 2012;37:1785-9.
175. Bertagnoli R, Zigler J, Karg A, Voigt S. Complications and strategies for revision surgery in total disc replacement. Orthop Clin North Am 2005;36:389-95.
176. Gerometta A, Rodriguez Olaverri JC, Bittan F. Infection and revision strategies in total disc arthroplasty. Int Orthop 2012;36:471-4.
177. Jacobs W, Van der Gaag NA, Tuschel A, et al. Total disc replacement for chronic back pain in the presence of disc degeneration. Cochrane Database Syst Rev 2012;9:CD008326.
178. Zander T, Bergmann G, Rohlmann A. Large sizes of vertebral body replacement do not reduce the contact pressure on adjacent vertebral bodies per se. Med Eng Phys 2009;31:1307-12.
179. Chen WM, Park C, Lee K, Lee S. In situ contact analysis of the prosthesis components of Prodisc-L in lumbar spine following total disc replacement. Spine (Phila Pa 1976) 2009;34:E716-23.
180. Harrop JS, Youssef JA, Maltenfort M, et al. Lumbar adjacent segment degeneration and disease after arthrodesis and total disc arthroplasty. Spine (Phila Pa 1976) 2008;33:1701-7.

181. Choi KC, Ryu KS, Lee SH, Kim YH, Lee SJ, Park CK. Biomechanical comparison of anterior lumbar interbody fusion: stand-alone interbody cage versus interbody cage with pedicle screw fixation -- a finite element analysis. BMC Musculoskelet Disord 2013;14:220.
182. Ostelo RW, van Tulder MW, Vlaeyen JW, Linton SJ, Morley SJ, Assendelft WJ. Behavioural treatment for chronic low-back pain. Cochrane Database Syst Rev 2005:CD002014.
183. Hoffman BM, Papas RK, Chatkoff DK, Kerns RD. Meta-analysis of psychological interventions for chronic low back pain. Health Psychol 2007;26:1-9.
184. Hellum C, Johnsen LG, Storheim K, et al. Surgery with disc prosthesis versus rehabilitation in patients with low back pain and degenerative disc: two year follow-up of randomised study. BMJ 2011;342:d2786.
185. van Tulder MW, Ostelo R, Vlaeyen JW, Linton SJ, Morley SJ, Assendelft WJ. Behavioral treatment for chronic low back pain: a systematic review within the framework of the Cochrane Back Review Group. Spine (Phila Pa 1976) 2000;25:2688-99.
186. van der Roer N, Goossens ME, Evers SM, van Tulder MW. What is the most cost-effective treatment for patients with low back pain? A systematic review. Best Pract Res Clin Rheumatol 2005;19:671-84.
187. Frymoyer JW, Cats-Baril WL. An overview of the incidences and costs of low back pain. Orthop Clin North Am 1991;22:263-71.
188. Soldal Dm. The epidemiology of low back pain. Nor J Epidemiol 2008;18:107-10.
189. Von Korff M, Wagner EH, Dworkin SF, Saunders KW. Chronic pain and use of ambulatory health care. Psychosom Med 1991;53:61-79.
190. Is spinal fusion surgery overused? OR Manager 2004;20:7.
191. Gore M, Sadosky A, Stacey BR, Tai KS, Leslie D. The Burden of Chronic Low Back Pain: Clinical Comorbidities, Treatment Patterns, and Healthcare Costs in Usual Care Settings. Spine (Phila Pa 1976) 2011.
192. Pai S, Sundaram LJ. Low back pain: an economic assessment in the United States. Orthop Clin North Am 2004;35:1-5.
193. Drummond MF. Methods for the economic evaluation of health care programmes. 3rd ed. Oxford ; New York: Oxford University Press; 2005.
194. Glick H. Economic evaluation in clinical trials. Oxford ; New York: Oxford University Press; 2007.
195. Petrou S, Gray A. Economic evaluation alongside randomised controlled trials: design, conduct, analysis, and reporting. BMJ 2011;342:d1548.
196. Bambha K, Kim WR. Cost-effectiveness analysis and incremental cost-effectiveness ratios: uses and pitfalls. Eur J Gastroenterol Hepatol 2004;16:519-26.
197. Claxton K. Exploring uncertainty in cost-effectiveness analysis. Pharmacoeconomics 2008;26:781-98.
198. Gray A, Clarke PM, Wolstenholme JL, Wordsworth S. Applied methods of cost-effectiveness analysis in health care. Oxford: Oxford University Press; 2011.
199. Barton GR, Briggs AH, Fenwick EA. Optimal cost-effectiveness decisions: the role of the cost-effectiveness acceptability curve (CEAC), the cost-effectiveness acceptability frontier (CEAF), and the expected value of perfection information (EVPI). Value Health 2008;11:886-97.
200. Brazier J. Measuring and valuing health benefits for economic evaluation. Oxford: Oxford University Press; 2007.
201. Nord E. Health state values from multiattribute utility instruments need correction. Ann Med 2001;33:371-4.
202. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. J Health Econ 2002;21:271-92.
203. Dolan P. Modeling valuations for EuroQol health states. Med Care 1997;35:1095-108.
204. Neumann PJ, Goldie SJ, Weinstein MC. Preference-based measures in economic evaluation in health care. Annu Rev Public Health 2000;21:587-611.

205. Dolan P, Gudex C, Kind P, Williams A. The time trade-off method: results from a general population study. Health Econ 1996;5:141-54.
206. Brazier J, Usherwood T, Harper R, Thomas K. Deriving a preference-based single index from the UK SF-36 Health Survey. J Clin Epidemiol 1998;51:1115-28.
207. Petitti DB. Meta-analysis, decision analysis, and cost-effectiveness analysis: methods for quantitative synthesis in medicine. New York: Oxford University Press; 2000.
208. Athanasiou T, Debas H, Darzi A. Key Topics in Surgical Research and Methodology. Berlin, Heidelberg: Springer Berlin Heidelberg; 2010.
209. Bang H, Zhao H. Median-Based Incremental Cost-Effectiveness Ratio (ICER). Journal of statistical theory and practice 2012;6:428-42.
210. Von Neumann J, Morgenstern O. Theory of games and economic behavior. Princeton, N.J.: Princeton University Press; 1944.
211. Rivero-Arias O, Campbell H, Gray A, Fairbank J, Frost H, Wilson-MacDonald J. Surgical stabilisation of the spine compared with a programme of intensive rehabilitation for the management of patients with chronic low back pain: cost utility analysis based on a randomised controlled trial. Bmj 2005;330:1239.
212. Fritzell P, Hagg O, Jonsson D, Nordwall A. Cost-effectiveness of lumbar fusion and nonsurgical treatment for chronic low back pain in the Swedish Lumbar Spine Study: a multicenter, randomized, controlled trial from the Swedish Lumbar Spine Study Group. Spine 2004;29:421-34; discussion Z3.
213. Fritzell P, Berg S, Borgstrom F, Tullberg T, Tropp H. Cost effectiveness of disc prosthesis versus lumbar fusion in patients with chronic low back pain: randomized controlled trial with 2-year follow-up. Eur Spine J 2010.
214. Schweikert B, Jacobi E, Seitz R, et al. Effectiveness and cost-effectiveness of adding a cognitive behavioral treatment to the rehabilitation of chronic low back pain. J Rheumatol 2006;33:2519-26.
215. van Geen JW, Edelaar MJ, Janssen M, van Eijk JT. The long-term effect of multidisciplinary back training: a systematic review. Spine (Phila Pa 1976) 2007;32:249-55.
216. Skouen JS, Grasdal AL, Haldorsen EM, Ursin H. Relative cost-effectiveness of extensive and light multidisciplinary treatment programs versus treatment as usual for patients with chronic low back pain on long-term sick leave: randomized controlled study. Spine (Phila Pa 1976) 2002;27:901-9; discussion 9-10.
217. Sogaard R, Bunger CE, Laurberg I, Christensen FB. Cost-effectiveness evaluation of an RCT in rehabilitation after lumbar spinal fusion: a low-cost, behavioural approach is cost-effective over individual exercise therapy. Eur Spine J 2008;17:262-71.
218. Jensen C, Nielsen CV, Jensen OK, Petersen KD. Cost-effectiveness and cost-benefit analyses of a multidisciplinary intervention compared with a brief intervention to facilitate return to work in sick-listed patients with low back pain. Spine (Phila Pa 1976) 2013;38:1059-67.
219. van der Roer N, van Tulder M, van Mechelen W, de Vet H. Economic evaluation of an intensive group training protocol compared with usual care physiotherapy in patients with chronic low back pain. Spine (Phila Pa 1976) 2008;33:445-51.
220. Dagenais S, Roffey DM, Wai EK, Haldeman S, Caro J. Can cost utility evaluations inform decision making about interventions for low back pain? The Spine Journal 2009;9:944-57.
221. Rolli Salathe C, Elfering A, Melloh M. [Efficacy, utility and cost-effectiveness of multidisciplinary treatment for chronic low back pain]. Schmerz 2012;26:131-49.
222. Soegaard R, Christensen FB. Health economic evaluation in lumbar spinal fusion: a systematic literature review anno 2005. Eur Spine J 2006;15:1165-73.
223. Fujiwara A, Tamai K, An HS, et al. The relationship between disc degeneration, facet joint osteoarthritis, and stability of the degenerative lumbar spine. J Spinal Disord 2000;13:444-50.
224. Allen RT, Rihn JA, Glassman SD, Currier B, Albert TJ, Phillips FM. An evidence-based approach to spine surgery. Am J Med Qual 2009;24:15S-24S.

225. Bertagnoli R, Kumar S. Indications for full prosthetic disc arthroplasty: a correlation of clinical outcome against a variety of indications. Eur Spine J 2002;11 Suppl 2:S131-6.
226. Tropiano P, Huang RC, Girardi FP, Marnay T. Lumbar disc replacement: preliminary results with ProDisc II after a minimum follow-up period of 1 year. J Spinal Disord Tech 2003;16:362-8.
227. Huang RC, Girardi FP, Cammisa FP, Jr., Lim MR, Tropiano P, Marnay T. Correlation between range of motion and outcome after lumbar total disc replacement: 8.6-year follow-up. Spine (Phila Pa 1976) 2005;30:1407-11.
228. Grotle M, Brox JI, Vollestad NK. Concurrent comparison of responsiveness in pain and functional status measurements used for patients with low back pain. Spine (Phila Pa 1976) 2004;29:E492-501.
229. Grotle M, Brox JI, Vollestad NK. Functional status and disability questionnaires: what do they assess? A systematic review of back-specific outcome questionnaires. Spine (Phila Pa 1976) 2005;30:130-40.
230. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol 2010;63:737-45.
231. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Qual Life Res 2010;19:539-49.
232. Terwee C. The COSMIN checklist manual. 2012.
233. Deyo RA, Battie M, Beurskens AJ, et al. Outcome measures for low back pain research. A proposal for standardized use. Spine (Phila Pa 1976) 1998;23:2003-13.
234. Hudak PL, Wright JG. The characteristics of patient satisfaction measures. Spine (Phila Pa 1976) 2000;25:3167-77.
235. Ostelo RW, de Vet HC. Clinically important outcomes in low back pain. Best Pract Res Clin Rheumatol 2005;19:593-607.
236. Walsh TL, Hanscom B, Lurie JD, Weinstein JN. Is a condition-specific instrument for patients with low back pain/leg symptoms really necessary? The responsiveness of the Oswestry Disability Index, MODEMS, and the SF-36. Spine (Phila Pa 1976) 2003;28:607-15.
237. Fayers PM, Machin D. Quality of life: the assessment, analysis and interpretation of patient-reported outcomes. Chichester: John Wiley; 2009.
238. DeVellis RF. Classical test theory. Med Care 2006;44:S50-9.
239. Walters SJ. Quality of life outcomes in clinical trials and health-care evaluation: a practical guide to analysis and interpretation. Chichester, West Sussex, U.K.: John Wiley & Sons; 2009.
240. Pallant JF, Tennant A. An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS). Br J Clin Psychol 2007;46:1-18.
241. Jadad AR. Randomised controlled trials: a user's guide. London: BMJ Books; 1998.
242. Fairbank JC, Pynsent PB. The Oswestry Disability Index. Spine (Phila Pa 1976) 2000;25:2940-52; discussion 52.
243. Bertagnoli R, Yue JJ, Shah RV, et al. The treatment of disabling multilevel lumbar discogenic low back pain with total disc arthroplasty utilizing the ProDisc prosthesis: a prospective study with 2-year minimum follow-up. Spine (Phila Pa 1976) 2005;30:2192-9.
244. Modic MT. Degenerative disc disease: genotyping, MR imaging and phenotyping. Skeletal Radiol 2007;36:91-3.
245. Esposito P, Pinheiro-Franco JL, Froelich S, Maitrot D. Predictive value of MRI vertebral end-plate signal changes (Modic) on outcome of surgically treated degenerative disc disease. Results of a cohort study including 60 patients. Neurochirurgie 2006;52:315-22.
246. Berg L, Gjertsen O, Hellum C, et al. Reliability of change in lumbar MRI findings over time in patients with and without disc prosthesis--comparing two different image evaluation methods. Skeletal Radiol 2012;41:1547-57.

247. Berg L, Hellum C, Gjertsen O, et al. Do more MRI findings imply worse disability or more intense low back pain? A cross-sectional study of candidates for lumbar disc prosthesis. Skeletal Radiol 2013.
248. Jensen TS, Sorensen JS, Kjaer P. Intra- and interobserver reproducibility of vertebral endplate signal (modic) changes in the lumbar spine: the Nordic Modic Consensus Group classification. Acta Radiol 2007;48:748-54.
249. Frobin W, Brinckmann P, Biggemann M, Tillotson M, Burton K. Precision measurement of disc height, vertebral height and sagittal plane displacement from lateral radiographic views of the lumbar spine. Clin Biomech (Bristol, Avon) 1997;12 Suppl 1:S1-S63.
250. Leivseth G, Brinckmann P, Frobin W, Johnsson R, Stromqvist B. Assessment of sagittal plane segmental motion in the lumbar spine. A comparison between distortion-compensated and stereophotogrammetric roentgen analysis. Spine (Phila Pa 1976) 1998;23:2648-55.
251. Fairbank JC, Couper J, Davies JB, O'Brien JP. The Oswestry low back pain disability questionnaire. Physiotherapy 1980;66:271-3.
252. Fairbank JC, Pynsent PB. The Oswestry Disability Index. Spine 2000;25:2940-52; discussion 52.
253. Grotle M, Brox JI, Vollestad NK. Cross-cultural adaptation of the Norwegian versions of the Roland-Morris Disability Questionnaire and the Oswestry Disability Index. J Rehabil Med 2003;35:241-7.
254. Ware JE, Jr., Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care 1992;30:473-83.
255. DeVine J, Norvell DC, Ecker E, et al. Evaluating the correlation and responsiveness of patient-reported pain with function and quality-of-life outcomes after spine surgery. Spine (Phila Pa 1976) 2011;36:S69-74.
256. Chapman JR, Norvell DC, Hermsmeyer JT, et al. Evaluating common outcomes for measuring treatment success for chronic low back pain. Spine (Phila Pa 1976) 2011;36:S54-68.
257. EuroQol--a new facility for the measurement of health-related quality of life. The EuroQol Group. Health Policy 1990;16:199-208.
258. Derogatis LR, Lipman RS, Rickels K, Uhlenhuth EH, Covi L. The Hopkins Symptom Checklist (HSCL). A measure of primary symptom dimensions. Mod Probl Pharmacopsychiatry 1974;7:79-110.
259. Waddell G, Newton M, Henderson I, Somerville D, Main CJ. A Fear-Avoidance Beliefs Questionnaire (FABQ) and the role of fear-avoidance beliefs in chronic low back pain and disability. Pain 1993;52:157-68.
260. Fritzell P, Hagg O, Nordwall A. Complications in lumbar fusion surgery for chronic low back pain: comparison of three surgical techniques used in a prospective randomized study. A report from the Swedish Lumbar Spine Study Group. Eur Spine J 2003;12:178-89.
261. Strand LI, Moe-Nilssen R, Ljunggren AE. Back Performance Scale for the assessment of mobility-related activities in people with back pain. Phys Ther 2002;82:1213-23.
262. Hagg O, Fritzell P, Ekselius L, Nordwall A. Predictors of outcome in fusion surgery for chronic low back pain. A report from the Swedish Lumbar Spine Study. Eur Spine J 2003;12:22-33.
263. Vanti C, Prosperi D, Boschi M. The Prolo Scale: history, evolution and psychometric properties. Journal of orthopaedics and traumatology : official journal of the Italian Society of Orthopaedics and Traumatology 2013.
264. Price DD, McGrath PA, Rafii A, Buckingham B. The validation of visual analogue scales as ratio scale measures for chronic and experimental pain. Pain 1983;17:45-56.
265. Dolan P. Modeling valuations for EuroQol health states. Medical Care 1997;35:1095-108.
266. Brazier J, Usherwood T, Harper R, Thomas K. Deriving a preference-based single index from the UK SF-36 Health Survey. Journal of Clinical Epidemiology 1998;51:1115-28.
267. Sogaard R, Christensen FB, Videbaek TS, Bunger C, Christiansen T. Interchangeability of the EQ-5D and the SF-6D in long-lasting low back pain. Value Health 2009;12:606-12.
268. Brooks R. EuroQol: the current state of play. Health Policy 1996;37:53.

269. Pedersen M. [Unit costs for hospital services related to hospital treatment of long-term low back pain. Multidisciplinary rehabilitation versus spinal surgery.] SINTEF; 2008.
270. Pedersen M, Sandvik AL. Benchmarking av kostnader ved regionsykehus, sentralsykehus og lokalsykehus2002.
271. Norwegian Medicines Agency. Guidelines on how to conduct pharmacoeconomic analyses2012.
272. Mokkink LB, Terwee CB, Knol DL, et al. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. BMC Med Res Methodol 2010;10:22.
273. Johnsen LG, Hellum C, Nygaard OP, 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.
274. Hagg O, Fritzell P, Nordwall A. The clinical importance of changes in outcome scores after treatment for chronic low back pain. Eur Spine J 2003;12:12-20.
275. Altman DG. Confidence intervals for the number needed to treat. BMJ 1998;317:1309-12.
276. Fenwick E, Claxton K, Sculpher M. Representing uncertainty: the role of cost-effectiveness acceptability curves. Health Econ 2001;10:779-87.
277. Lothgren M, Zethraeus N. Definition, interpretation and calculation of cost-effectiveness acceptability curves. Health Econ 2000;9:623-30.
278. Briggs A, Clark T, Wolstenholme J, Clarke P. Missing... presumed at random: cost-analysis of incomplete data. Health Econ 2003;12:377-92.
279. Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. J Clin Epidemiol 2006;59:1087-91.
280. van der Heijden GJ, Donders AR, Stijnen T, Moons KG. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example. J Clin Epidemiol 2006;59:1102-9.
281. Sterne JA, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. Bmj 2009;338:b2393.
282. Guyatt GH, Osoba D, Wu AW, Wyrwich KW, Norman GR. Methods to explain the clinical significance of health status measures. Mayo Clin Proc 2002;77:371-83.
283. Walters SJ, Brazier JE. What is the relationship between the minimally important difference and health state utility values? The case of the SF-6D. Health Qual Life Outcomes 2003;1:4.
284. 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;31:578-82.
285. Sell L, Bultmann U, Rugulies R, Villadsen E, Faber A, Sogaard K. Predicting long-term sickness absence and early retirement pension from self-reported work ability. Int Arch Occup Environ Health 2009;82:1133-8.
286. Sogaard R, Ebert B, Klaerke D, Werge T. Triton X-100 inhibits agonist-induced currents and suppresses benzodiazepine modulation of GABA(A) receptors in Xenopus oocytes. Biochim Biophys Acta 2009;1788:1073-80.
287. Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. J Clin Epidemiol 2003;56:395-407.
288. Bland JM, Altman DG. Measurement error. BMJ 1996;313:744.
289. de Vet HC, Ostelo RW, Terwee CB, et al. Minimally important change determined by a visual method integrating an anchor-based and a distribution-based approach. Qual Life Res 2007;16:131-42.
290. Beaton DE. Understanding the relevance of measured change through studies of responsiveness. Spine 2000;25:3192-9.
291. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1:307-10.

292. Smith EV, Jr. Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. J Appl Meas 2002;3:205-31.
293. Chou Y-T, Wang W-C. Checking Dimensionality in Item Response Models With Principal Component Analysis on Standardized Residuals. Educational and Psychological Measurement 2010;70:717-31.
294. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem 1993;39:561-77.
295. Olsen NT, Mogelvang R, Jons C, Fritz-Hansen T, Sogaard P. Predicting response to cardiac resynchronization therapy with cross-correlation analysis of myocardial systolic acceleration: a new approach to echocardiographic dyssynchrony evaluation. J Am Soc Echocardiogr 2009;22:657-64.
296. 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. Spine J 2008;8:968-74.
297. Pocock SJ. Clinical trials: a practical approach. Chichester: Wiley; 1983.
298. Chalmers TC, Celano P, Sacks HS, Smith H, Jr. Bias in treatment assignment in controlled clinical trials. N Engl J Med 1983;309:1358-61.
299. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA 1995;273:408-12.
300. Hollis S, Campbell F. What is meant by intention to treat analysis? Survey of published randomised controlled trials. BMJ 1999;319:670-4.
301. Brox JI, Nygaard OP, Holm I, Keller A, Ingebrigtsen T, Reikeras O. Four-year follow-up of surgical versus non-surgical therapy for chronic low back pain. Ann Rheum Dis 2010;69:1643-8.
302. Indahl A, Haldorsen EH, Holm S, Reikeras O, Ursin H. Five-year follow-up study of a controlled clinical trial using light mobilization and an informative approach to low back pain. Spine (Phila Pa 1976) 1998;23:2625-30.
303. Indahl A, Velund L, Reikeraas O. Good prognosis for low back pain when left untampered. A randomized clinical trial. Spine (Phila Pa 1976) 1995;20:473-7.
304. Storheim K, Brox JI, Holm I, Koller AK, Bo K. Intensive group training versus cognitive intervention in sub-acute low back pain: short-term results of a single-blind randomized controlled trial. J Rehabil Med 2003;35:132-40.
305. Storheim K, Holm I, Gunderson R, Brox JI, Bo K. The effect of comprehensive group training on cross-sectional area, density, and strength of paraspinal muscles in patients sick-listed for subacute low back pain. J Spinal Disord Tech 2003;16:271-9.
306. Zigler J, Delamarter R, Spivak JM, et al. Results of the prospective, randomized, multicenter Food and Drug Administration investigational device exemption study of the ProDisc-L total disc replacement versus circumferential fusion for the treatment of 1-level degenerative disc disease. Spine (Phila Pa 1976) 2007;32:1155-62; discussion 63.
307. Berg S, Tullberg T, Branth B, Olerud C, Tropp H. Total disc replacement compared to lumbar fusion: a randomised controlled trial with 2-year follow-up. Eur Spine J 2009;18:1512-9.
308. Blumenthal S, McAfee PC, Guyer RD, et al. A prospective, randomized, multicenter Food and Drug Administration investigational device exemptions study of lumbar total disc replacement with the CHARITE artificial disc versus lumbar fusion: part I: evaluation of clinical outcomes. Spine (Phila Pa 1976) 2005;30:1565-75; discussion E387-91.
309. Gornet MF, Burkus JK, Dryer RF, Peloza JH. Lumbar disc arthroplasty with Maverick disc versus stand-alone interbody fusion: a prospective, randomized, controlled, multicenter investigational device exemption trial. Spine (Phila Pa 1976) 2011;36:E1600-11.
310. Moreno P, Boulot J. [Comparative study of short-term results between total artificial disc prosthesis and anterior lumbar interbody fusion]. Rev Chir Orthop Reparatrice Appar Mot 2008;94:282-8.

311. Sasso RC, Foulk DM, Hahn M. Prospective, randomized trial of metal-on-metal artificial lumbar disc replacement: initial results for treatment of discogenic pain. Spine (Phila Pa 1976) 2008;33:123-31.
312. Guyer RD, McAfee PC, Hochschuler SH, et al. Prospective randomized study of the Charite artificial disc: data from two investigational centers. Spine J 2004;4:252S-9S.
313. McAfee PC, Cunningham B, Holsapple G, et al. A prospective, randomized, multicenter Food and Drug Administration investigational device exemption study of lumbar total disc replacement with the CHARITE artificial disc versus lumbar fusion: part II: evaluation of radiographic outcomes and correlation of surgical technique accuracy with clinical outcomes. Spine (Phila Pa 1976) 2005;30:1576-83; discussion E388-90.
314. van den Eerenbeemt KD, Ostelo RW, van Royen BJ, Peul WC, van Tulder MW. Total disc replacement surgery for symptomatic degenerative lumbar disc disease: a systematic review of the literature. Eur Spine J 2010;19:1262-80.
315. Jacobs WC, van der Gaag NA, Kruyt MC, et al. Total disc replacement for chronic discogenic low back pain: a cochrane review. Spine (Phila Pa 1976) 2013;38:24-36.
316. Kurtz SM, Lau E, Ianuzzi A, et al. National Revision Burden for Lumbar Total Disc Replacement in the United States: Epidemiologic and Economic Perspectives. Spine (Phila Pa 1976) 2010.
317. Punt IM, Austen S, Cleutjens JP, et al. Are periprosthetic tissue reactions observed after revision of total disc replacement comparable to the reactions observed after total hip or knee revision surgery? Spine (Phila Pa 1976) 2012;37:150-9.
318. Punt IM, Cleutjens JP, de Bruin T, et al. Periprosthetic tissue reactions observed at revision of total intervertebral disc arthroplasty. Biomaterials 2009;30:2079-84.
319. van Ooij A, Kurtz SM, Stessels F, Noten H, van Rhijn L. Polyethylene wear debris and long-term clinical failure of the Charite disc prosthesis: a study of 4 patients. Spine (Phila Pa 1976) 2007;32:223-9.
320. Kurtz SM, van Ooij A, Ross R, et al. Polyethylene wear and rim fracture in total disc arthroplasty. Spine J 2007;7:12-21.
321. McAfee PC, Geisler FH, Saiedy SS, et al. Revisability of the CHARITE artificial disc replacement: analysis of 688 patients enrolled in the U.S. IDE study of the CHARITE Artificial Disc. Spine (Phila Pa 1976) 2006;31:1217-26.
322. Inamasu J, Guiot BH. Vascular injury and complication in neurosurgical spine surgery. Acta Neurochir (Wien) 2006;148:375-87.
323. Wood KB, Devine J, Fischer D, Dettori JR, Janssen M. Vascular injury in elective anterior lumbosacral surgery. Spine (Phila Pa 1976) 2010;35:S66-75.
324. Frost H, Lamb SE, Stewart-Brown S. Responsiveness of a patient specific outcome measure compared with the Oswestry Disability Index v2.1 and Roland and Morris Disability Questionnaire for patients with subacute and chronic low back pain. Spine (Phila Pa 1976) 2008;33:2450-7; discussion 8.
325. Vianin M. Psychometric properties and clinical usefulness of the Oswestry Disability Index. Journal of chiropractic medicine 2008;7:161-3.
326. Changulani M, Shaju A. Evaluation of responsiveness of Oswestry low back pain disability index. Arch Orthop Trauma Surg 2009;129:691-4.
327. de Vet HC, Terwee CB, Ostelo RW, Beckerman H, Knol DL, Bouter LM. Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. Health Qual Life Outcomes 2006;4:54.
328. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. Control Clin Trials 1989;10:407-15.
329. Guyatt GH, Juniper EF, Walter SD, Griffith LE, Goldstein RS. Interpreting treatment effects in randomised trials. BMJ 1998;316:690-3.
330. Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. J Clin Epidemiol 2008;61:102-9.

331. Terwee CB, Roorda LD, Dekker J, et al. Mind the MIC: large variation among populations and methods. J Clin Epidemiol 2010;63:524-34.
332. Jensen RK, Leboeuf-Yde C. Is the presence of modic changes associated with the outcomes of different treatments? A systematic critical review. BMC Musculoskelet Disord 2011;12:183.
333. Tosteson AN, Tosteson TD, Lurie JD, et al. Comparative effectiveness evidence from the spine patient outcomes research trial: surgical versus nonoperative care for spinal stenosis, degenerative spondylolisthesis, and intervertebral disc herniation. Spine (Phila Pa 1976) 2011;36:2061-8.
334. 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 Orthop 2011;82:56-63.
335. Schulz KF, Altman DG, Moher D, Group C. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. Bmj 2010;340:c332.
336. Briggs AH. Statistical approaches to handling uncertainty in health economic evaluation. Eur J Gastroenterol Hepatol 2004;16:551-61.
337. Siepe CJ, Hitzl W, Meschede P, Sharma AK, Khattab MF, Mayer MH. Interdependence between disc space height, range of motion and clinical outcome in total lumbar disc replacement. Spine (Phila Pa 1976) 2009;34:904-16.
338. Leivseth G, Braaten S, Frobin W, Brinckmann P. Mobility of lumbar segments instrumented with a ProDisc II prosthesis: a two-year follow-up study. Spine (Phila Pa 1976) 2006;31:1726-33.
339. Guyer RD, McAfee PC, Banco RJ, et al. Prospective, randomized, multicenter Food and Drug Administration investigational device exemption study of lumbar total disc replacement with the CHARITE artificial disc versus lumbar fusion: five-year follow-up. Spine J 2009;9:374-86.
340. Freeman BJ, Davenport J. Total disc replacement in the lumbar spine: a systematic review of the literature. Eur Spine J 2006;15 Suppl 3:S439-47.
341. Penning L, Wilmink JT, van Woerden HH. Inability to prove instability. A critical appraisal of clinical-radiological flexion-extension studies in lumbar disc degeneration. Diagn Imaging Clin Med 1984;53:186-92.
342. Auerbach JD, Jones KJ, Milby AH, Anakwenze OA, Balderston RA. Segmental contribution toward total lumbar range of motion in disc replacement and fusions: a comparison of operative and adjacent levels. Spine (Phila Pa 1976) 2009;34:2510-7.
343. Hellum C, Berg L, Gjertsen O, et al. Adjacent level degeneration and facet arthropathy after disc prosthesis surgery or rehabilitation in patients with chronic low back pain and degenerative disc: second report of a randomized study. Spine (Phila Pa 1976) 2012;37:2063-73.
344. Yao J, Turteltaub SR, Ducheyne P. A three-dimensional nonlinear finite element analysis of the mechanical behavior of tissue engineered intervertebral discs under complex loads. Biomaterials 2006;27:377-87.
345. Li H, Wang Z. Intervertebral disc biomechanical analysis using the finite element modeling based on medical images. Comput Med Imaging Graph 2006;30:363-70.
346. Drummond MF, Jefferson TO. Guidelines for authors and peer reviewers of economic submissions to the BMJ. The BMJ Economic Evaluation Working Party. BMJ 1996;313:275-83.
347. Bleichrodt H. A new explanation for the difference between time trade-off utilities and standard gamble utilities. Health Econ 2002;11:447-56.
348. Read JL, Quinn RJ, Berwick DM, Fineberg HV, Weinstein MC. Preferences for health outcomes. Comparison of assessment methods. Med Decis Making 1984;4:315-29.
349. Doctor JN, Bleichrodt H, Lin HJ. Health utility bias: a systematic review and meta-analytic evaluation. Med Decis Making 2010;30:58-67.
350. Hallan S, Asberg A, Indredavik B, Wideroe TE. Quality of life after cerebrovascular stroke: a systematic study of patients' preferences for different functional outcomes. J Intern Med 1999;246:309-16.
351. Brazier J, Roberts J, Tsuchiya A, Busschbach J. A comparison of the EQ-5D and SF-6D across seven patient groups. Health Econ 2004;13:873-84.

352. Soegaard R. Interchangeability of the EQ-5D and the SF-6D in Long-Lasting Low Back Pain Source: Value in Health 12, no. 4 (2009): 606-612 Additional Info: Blackwell Publishing; 20090601 Standard No: ISSN: 1098-3015 DOI: 10.1111/j.1524-4733.2008.00466.x. Value in Health 2009;12:606-12.
353. Barton GR, Sach TH, Doherty M, Avery AJ, Jenkinson C, Muir KR. An assessment of the discriminative ability of the EQ-5Dindex, SF-6D, and EQ VAS, using sociodemographic factors and clinical conditions. Eur J Health Econ 2008;9:237-49.
354. Saal JS, Franson RC, Dobrow R, Saal JA, White AH, Goldthwaite N. High levels of inflammatory phospholipase A2 activity in lumbar disc herniations. Spine (Phila Pa 1976) 1990;15:674-8.
355. Joore M, Brunenberg D, Nelemans P, et al. The impact of differences in EQ-5D and SF-6D utility scores on the acceptability of cost-utility ratios: results across five trial-based cost-utility studies. Value Health 2009;13:222-9.
356. Vainiola T, Roine RP, Pettila V, Kantola T, Rasanen P, Sintonen H. Effect of health-related quality-of-life instrument and quality-adjusted life year calculation method on the number of life years gained in the critical care setting. Value Health 2011;14:1130-4.
357. Whitehurst DG, Bryan S, Lewis M. Systematic review and empirical comparison of contemporaneous EQ-5D and SF-6D group mean scores. Med Decis Making 2011;31:E34-44.
358. Sach TH, Barton GR, Jenkinson C, Doherty M, Avery AJ, Muir KR. Comparing cost-utility estimates: does the choice of EQ-5D or SF-6D matter? Med Care 2009;47:889-94.
359. Solberg TK, Olsen JA, Ingebrigtsen T, Hofoss D, Nygaard OP. Health-related quality of life assessment by the EuroQol-5D can provide cost-utility data in the field of low-back surgery. Eur Spine J 2005;14:1000-7.
360. Adobor RD, Rimeslatten S, Keller A, Brox JI. Repeatability, reliability, and concurrent validity of the scoliosis research society-22 questionnaire and EuroQol in patients with adolescent idiopathic scoliosis. Spine (Phila Pa 1976) 2010;35:206-9.
361. Rabin R, de Charro F. EQ-5D: a measure of health status from the EuroQol Group. AnnMed 2001;33:337.
362. Dyer MT, Goldsmith KA, Sharples LS, Buxton MJ. A review of health utilities using the EQ-5D in studies of cardiovascular disease. Health Qual Life Outcomes 2010;8:13.
363. Djurasovic M, Glassman SD, Dimar JR, 2nd, Crawford CH, 3rd, Bratcher KR, Carreon LY. Changes in the Oswestry Disability Index that predict improvement after lumbar fusion. Journal of neurosurgery Spine 2012;17:486-90.
364. Suarez-Almazor ME, Kendall C, Johnson JA, Skeith K, Vincent D. Use of health status measures in patients with low back pain in clinical settings. Comparison of specific, generic and preference-based instruments. Rheumatology (Oxford) 2000;39:783-90.
365. Albert HB, Sorensen JS, Christensen BS, Manniche C. Antibiotic treatment in patients with chronic low back pain and vertebral bone edema (Modic type 1 changes): a double-blind randomized clinical controlled trial of efficacy. Eur Spine J 2013;22:697-707.
366. Chan SC, Gantenbein-Ritter B. Intervertebral disc regeneration or repair with biomaterials and stem cell therapy--feasible or fiction? Swiss medical weekly 2012;142:w13598.
367. Mehrkens A, Muller AM, Valderrabano V, Scharen S, Vavken P. Tissue engineering approaches to degenerative disc disease--a meta-analysis of controlled animal trials. Osteoarthritis Cartilage 2012;20:1316-25.
368. Kalson NS, Richardson S, Hoyland JA. Strategies for regeneration of the intervertebral disc. Regenerative medicine 2008;3:717-29.
369. Hegewald AA, Ringe J, Sittinger M, Thome C. Regenerative treatment strategies in spinal surgery. Front Biosci 2008;13:1507-25.
370. Nishiyama Y, Nakamura M, Henmi C, et al. Development of a three-dimensional bioprinter: construction of cell supporting structures using hydrogel and state-of-the-art inkjet technology. J Biomech Eng 2009;131:035001.

371. Cui X, Boland T. Human microvasculature fabrication using thermal inkjet printing technology. Biomaterials 2009;30:6221-7.
372. Fedorovich NE, Swennen I, Girones J, et al. Evaluation of photocrosslinked Lutrol hydrogel for tissue printing applications. Biomacromolecules 2009;10:1689-96.
373. Boland T, Xu T, Damon B, Cui X. Application of inkjet printing to tissue engineering. Biotechnology journal 2006;1:910-7.
374. Etebar S, Cahill DW. Risk factors for adjacent-segment failure following lumbar fixation with rigid instrumentation for degenerative instability. J Neurosurg 1999;90:163-9.
375. Brayda-Bruno M, Tibiletti M, Ito K, et al. Advances in the diagnosis of degenerated lumbar discs and their possible clinical application. Eur Spine J 2013.
376. The Norwegian Spine Study Group: Lumbar Disc Prosthesis Versus Multidisciplinary Rehabilitation; 8-year Follow-up. 2012. at http://www.clinicaltrial.gov/ct2/show/NCT01704677?term=Total+disc+replacement&rank=7.)
377. Rasch G. Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago; 1960.
378. Davidson M. Rasch analysis of three versions of the Oswestry Disability Questionnaire. Man Ther 2007.
379. Rasch G. An item analysis which takes individual differences into account. Br J Math Stat Psychol 1966;19:49-57.
380. A rasch model for partial credit scoring. Psychometrika 1982;47:149-74.

Paper I

BMJ

RESEARCH

Surgery with disc prosthesis versus rehabilitation in patients with low back pain and degenerative disc: two year follow-up of randomised study

Christian Hellum, orthopaedic surgeon,[1] Lars Gunnar Johnsen, orthopaedic surgeon,[2,3] Kjersti Storheim, physiotherapist,[4,5,6] Øystein P Nygaard, neurosurgeon,[2] Jens Ivar Brox, consultant,[1] Ivar Rossvoll, orthopaedic surgeon,[2,3] Magne Rø, consultant,[7] Leiv Sandvik, professor,[8] Oliver Grundnes, orthopaedic surgeon[5] and the Norwegian Spine Study Group

[1]Department of Orthopaedics, Oslo University Hospital and University of Oslo, Kirkevn 166, 0407 Oslo, Norway
[2]National Centre for Diseases of the Spine, University Hospital of Trondheim, 7030 Trondheim
[3]Orthopaedic Department, University Hospital of Trondheim, 7030 Trondheim
[4]Norwegian Research Centre for Active Rehabilitation (NAR), Department of Orthopaedics, Oslo University Hospital, Kirkevn 166, 0407 Oslo
[5]Hjelp24, Nimi, Oslo Sognsveien 75 D, 0855 Oslo
[6]Faculty of Medicine, Norwegian University of Science and Technology, Trondheim, Høgskoleringen 1, 7491 Trondheim (MR), FORMI, Oslo University Hospital, Kirkevn 166, 0407 Oslo
[7]Multidiscipline Spinal Unit, Department of Physical Medicine and Rehabilitation, University Hospital of Trondheim, 7030 Trondheim
[8]Section for Biostatistics and Epidemiology, Oslo University Hospital, Kirkevn 166, 0407 Oslo
Correspondence to: C Hellum
christian.hellum@medisin.uio.no

Cite this as: *BMJ* 2011;342:d2786
doi:10.1136/bmj.d2786

ABSTRACT

Objective To compare the efficacy of surgery with disc prosthesis versus non-surgical treatment for patients with chronic low back pain.

Design A prospective randomised multicentre study.

Setting Five university hospitals in Norway.

Participants 173 patients with a history of low back pain for at least one year, Oswestry disability index of at least 30 points, and degenerative changes in one or two lower lumbar spine levels (86 patients randomised to surgery). Patients were treated from April 2004 to September 2007.

Interventions Surgery with disc prosthesis or outpatient multidisciplinary rehabilitation for 12-15 days.

Main outcome measures The primary outcome measure was the score on the Oswestry disability index after two years. Secondary outcome measures were low back pain, satisfaction with life (SF-36 and EuroQol EQ-5D), Hopkins symptom check list (HSCL-25), fear avoidance beliefs (FABQ), self efficacy beliefs for pain, work status, and patients' satisfaction and drug use. A blinded independent observer evaluated scores on the back performance scale and Prolo scale at two year follow-up.

Results The study was powered to detect a difference of 10 points on the Oswestry disability index between the groups at two years. At two years there was a mean difference of −8.4 points (95% confidence interval −13.2 to −3.6) in favour of surgery. In the analysis of prespecified secondary outcomes, there were significant differences in favour of surgery for low back pain (mean difference −12.2, −21.3 to −3.1), patients' satisfaction (63% (n=46) *v* 39% (n=26)), SF-36 physical component score (mean difference 5.8, 2.5 to 9.1), self efficacy for pain (mean difference 1.0, 0.2 to 1.9), and the Prolo scale (mean difference 0.9, 0.1 to 1.6). There were no significant differences in return to work, SF-36 mental component score, EQ-5D, fear avoidance beliefs, Hopkins symptom check list, drug use, and the back performance scale. One serious complication of leg amputation occurred during surgical revision of a polyethylene dislodgement. The drop-out rate was 20% (34) and the crossover rate was 6% (5).

Conclusions Surgical intervention with disc prosthesis for chronic low back pain resulted in a significantly greater improvement in the Oswestry score compared with rehabilitation, but this improvement did not clearly exceed the prespecified minimally important clinical difference between groups of 10 points, and the data are consistent with a wide range of differences between the groups, including values well below 10 points. The potential risks of surgery and the substantial amount of improvement experienced by a sizeable proportion of the rehabilitation group also have to be incorporated into overall decision making.

Trial registration www.clinicaltrial.gov NCT 00394732.

INTRODUCTION

Low back pain is common with a lifetime prevalence of about 59-84%.[1] Although relatively few patients develop chronic low back pain with disability, it represents extensive individual, societal, and financial problems. In patients who have had longstanding or serious disabling low back pain in the previous 12 months, a third will improve and have less serious problems during the following year.[2] Most patients who develop chronic low back pain, however, stay in this condition for years.

Fusion of assumed symptomatic segments in patients with chronic low back pain has been used widely, but randomised studies comparing fusion with non-surgical treatment indicate that a rehabilitation programme can be as effective as surgery. Four randomised studies have compared lumbar fusion with non-operative treatment.[3-7] Fritzell et al found that fusion significantly reduced pain and disability compared with usual care.[3] Brox et al and Fairbank et al compared fusion with a multidisciplinary rehabilitation programme focusing on cognitive intervention and supervised exercise.[4-7] They found similar improvement in pain and disability in the two intervention groups.

During the past 25 years, insertion of a disc prosthesis has become an option. In the four published

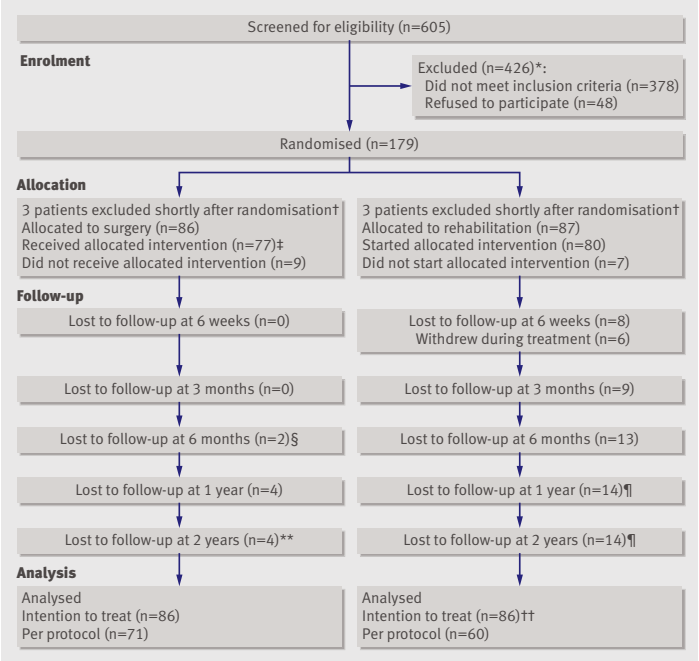

Fig 1 | Enrolment, randomisation, and follow-up of study patients, showing cumulative values at two years. *Not enough degenerative change to satisfy inclusion criteria (n=29), degenerative changes in more than two lower lumbar discs (n=80), Oswestry disability index score too low (n=88), did not want to undergo surgery (n=28), did not want to participate in rehabilitation (n=20), too much general pain (n=20), had previously been through similar training programme (n=26), and other reasons (n=135; deformity, psoriasis arthritis, language problems, coccygodynia, age, fracture, previous operation, tumour, spondylodiscitis, hip arthrosis). †Coronary heart disease and heart attack some days after randomisation (n=1); obvious exclusion criterion discovered some days after randomisation (n=50; earlier large abdominal operation (n=1), not enough degenerative change to satisfy inclusion criteria (n=2), degenerative changes in more than two lower lumbar discs (n=2). ‡One patient received one of two disc prostheses because of bleeding. §One patient with serious vascular complication underwent secondary leg amputation and was lost to follow-up. ¶One patient crossed over between 6 months and 1 year and five patients between 1 year and 2 years. Five patients underwent surgery with disc prosthesis and one patient with fusion. **Two patients underwent surgery with instrumented fusion before two year follow-up. ††One patient excluded because of missing baseline values and follow-up values

randomised studies comparing disc prosthesis with fusion, the clinical outcome of disc prosthesis was at least equivalent to that of fusion.[8-11] As surgical procedures should be evaluated against non-surgical methods,[12 13] we compared the efficacy of disc prosthesis and a multidisciplinary rehabilitation programme.

METHODS

Study design

A multicentre study conducted at five university hospitals in Norway included patients with low back pain and degenerative discs. Patients were included in the period between April 2004 and May 2007 and were treated within three months after randomisation. They were randomised in blocks with a website hosted by the medical faculty. Allocation was concealed for all people involved in the trial. A coordinating secretary not involved in the treatment could access randomisation details on the internet. The patient and the treating unit were informed about the allocation shortly after randomisation. Randomisation was stratified by centre (the five university hospitals) and whether the patient had had previous surgery (microsurgical decompression) or not. Independent observers collected and entered data. Storage of data was allowed by the Norwegian data inspectorate.

Participants

Patients were referred from all health regions in Norway. They were recruited from local hospitals or primary care to their nearest university hospital as usual without any supplemental recruitment attempt. An orthopaedic surgeon and a specialist in physical medicine and rehabilitation examined the patients before enrolment. All patients were informed about the procedures and told that neither of the treatment methods was documented as superior to the other. Eligible patients were aged 25-55 and had low back pain as the main symptom for at least a year, structured physiotherapy or chiropractic treatment for at least six months without sufficient effect, a score of at least 30 on the Oswestry disability index, and degenerative intervertebral disc changes in L4/L5 or L5/S1, or both. Degeneration had to be restricted to the two lower levels. We evaluated the following degenerative changes: at least 40% reduction of disc height,[14] Modic changes type I or II, or both,[15] high intensity zone in the disc,[16] and morphological changes classified as changes in signal intensity in the disc of grade 3 or 4.[17] The disc was classified as degenerative if the first criterion alone or at least two changes were found on magnetic resonance imaging. The discs were independently classified by two observers (orthopaedic surgeon/radiologist). When there was disagreement, a third observer classified the images and the outcome was decided by simple majority.

Degeneration of the facet joints was not an exclusion criterion, but symptoms of nerve root involvement were. Details of further inclusion and exclusion criteria, compliance with randomisation, and drop-outs are listed in the appendix 1 on bmj.com.

Study interventions

Rehabilitation—The rehabilitation was based on the treatment model described by Brox et al[4] and consisted of a cognitive approach and supervised physical exercise. A team of physiotherapists and specialists in physical medicine and rehabilitation directed the multidisciplinary treatment. Other specialists, such as psychologists, nurses, social workers, etc, could complete the team. The intervention was standardised through three seminars and videos and lecture sessions for the treatment providers before the study. The intervention was organised as an outpatient treatment in groups at the involved university hospitals and lasted for about 60 hours over three to five weeks. The treatment consisted of lectures and individual discussions focusing on relevant topics (such as anatomy and the

Table 1 | Baseline characteristics in patients with low back pain and degenerative disc randomised to disc prosthesis surgery or rehabilitation. Figures are numbers (percentage) unless stated otherwise

	Surgery (n=86)	Rehabilitation (n=86)
Mean (SD) age (years)	41.1 (7.1)	40.8 (7.1)
Women	40 (47)	51 (59)
Mean (SD) duration of back pain (months)	76 (72	85 (74)
Education:		
Primary school (9 years)	19 (22)	17 (20)
High school (12 years)	44 (51)	58 (67)
College	14 (16)	8 (9)
University	9 (11)	3 (4)
Mean (SD) body mass index (BMI)	25.6 (3.1)	25.5 (3.5)
Current smokers	42 (49)	37 (43)
Work status (working *v* not working):		
Working (includes part time sick leave)	24 (28)	22 (26)
On sick leave	25 (29)	34 (41)
Rehabilitation	29 (34)	25 (29)
Disability pension	3 (4)	0
Homemaker	0	2 (2)
Unemployed	1 (1)	0
Student	3 (4)	0
Unknown	1 (1)	3 (4)
Comorbidity	20 (23)	21 (24)
Daily consumption of narcotics	23 (27)	17 (20)
Previous surgery	23 (27)	25 (29)
Mean (SD) ODI score	41.8 (9.1)	42.8 (9.3)
Low back pain score*	64.9 (15.3)	73.6 (13.9)
Mean (SD) SF-36 score:		
Physical function	52.7 (17.6)	50.6 (17.7)
Role physical	25.3 (24.2	23.9 (18.7)
Bodily pain	24.9 (16.5)	24.4 (12.1)
General health	57.9 (19.7)	55.9 (19.9)
Vitality	37.8 (20.2)	33.1 (19.9)
Social function	53.0 (30.6)	57.6 (26.7)
Role emotion	72.5 (33.3)	67.6 (32.7)
Mental health	71.7 (18.0)	65.8 (18.9)
Physical component summary score	30.5 (7.1)	30.8 (6.5)
Mental component summary score	47.7 (13.0)	45.2 (13.2)
Mean (SD) HSCL-25	1.8 (0.5)	1.9 (0.5)
Mean (SD) FABQ work	25.9 (11.3)	27.4 (9.9)
Mean (SD) FABQ physical	14.1 (5.8)	12 (5.5)

ODI=Oswestry disability index (0 to 100, lower scores indicate less severe symptoms); SF-36=short form-36 (0 to 100, higher scores indicate better health status); HSCL-25=Hopkins symptom check list (for emotional distress, scores range from 1 to 4, lower scores indicate less severe symptoms); FABQ=fear avoidance belief questionnaire (scale ranges from 0 to 24 (physical) and from 0 to 42 (work), lower scores indicate less severe symptoms).
*Calculated with horizontal scale ranging from 0 (no pain) to 100 (worst pain imaginable), with word anchors at the beginning and end.

physiological aspects of the back, diagnostics, imaging, pain medicine, normal reactions, coping strategies, family and social life, and working conditions), daily workouts for increased physical capacity (endurance, strength, coordination, and specific training of the abdominal muscles and the lumbar multifidus muscles), and challenging patients' thoughts about, and participation in, physical activities previously labelled as not recommended (such as lifting, jumping, vacuum cleaning, dancing, and ball games). Follow-up consultations were conducted at six weeks, three months, six months, and one year after the intervention. See appendix 2 on bmj.com for detailed description of the rehabilitation intervention.

Surgery—The surgical intervention consisted of replacement of the degenerative intervertebral lumbar disc with an artificial lumbar disc (ProDisc II, Synthes Spine). The ProDisc consists of three pieces: two metal endplates of cobalt chromium molybdenum alloy and a core (made from ultrahigh molecular weight polyethylene) fixed to the inferior endplate after insertion. Surgeons used a Pfannenstiel or a para-median incision with a retroperitoneal approach. A nearly complete discectomy was performed with removal of the cartilaginous endplates and a sufficient release of the posterior longitudinal ligament to ensure disc space mobilisation. A fluoroscope was used to ensure that the prosthesis was placed in the midline and sufficiently towards the posterior edge of the vertebrae. All hospitals participating in the study used the same artificial lumbar disc device. One surgeon at each centre had main responsibility for the operation (five centres and five surgeons). Surgeons were required to have inserted at least six disc prostheses before performing surgery in the study. There were no major postoperative restrictions. Patients were not referred for postoperative physiotherapy, but at six weeks' follow-up they could be referred for physiotherapy if required, emphasising general mobilisation and non-specific exercises.

Outcome measures

The primary outcome measure was pain and disability measured with version 2.0 of the Oswestry disability index,[18] translated into Norwegian and tested for psychometric properties by Grotle et al.[19] (Scores range from 0 to 100, with lower score indicating less severe pain and disability.) Secondary outcomes included low back pain (measured with a visual analogue scale, ranging from 0 (no pain) to 100 (worst pain imaginable)) and general health status assessed with SF-36 (scores range from 0 to 100, higher scores correspond to better health status)[20 21] and EQ-5D (scores range from −0.59 to 1 (1 equals perfect health)).[22] For psychological variables we included emotional distress (Hopkins symptom check list (HSCL-25), scores range from 1 to 4, with lower scores indicating less severe symptoms) and the fear avoidance belief questionnaire (FABQ) for work and physical activity (scores range from 0 to 42 (work) and from 0 to 24 (physical), with lower scores indicating less severe symptoms).[23 24] Self efficacy beliefs for pain were registered by a subscale of the arthritis self efficacy scale (scores range from 1 to 10 and are summarised and divided by 5; lower scores indicate uncertainty in managing the pain).[25] Work status was evaluated as suggested by Fritzell et al.[3] (See table A in appendix 3 on bmj.com.) We calculated a net back to work rate, subtracting patients who went back to work from patients who stopped working, satisfaction with the result of the treatment on a seven point

Table 2 | Treatment and complications in 77 patients with low back pain and degenerative disc randomised to disc prosthesis surgery

Variable	Surgery group
No (%) by level of operation:	
L4/L5	17 (22)
L5/S1	35 (46)
L4/L5 and L5/S1	25 (33)
Median (range) operative time (min)	165 (72-570)
Median (range) blood loss (ml)	310 (50-6000)
Mean (SD) length of hospital stay (days)	7.2 (3.6)
No with complications:	
Intimal lesion in left common iliac artery*	1
Arterial thrombosis of dorsalis pedis artery†	1
Dural tear	0
Blood loss >1500 ml	4
Retrograde ejaculation (at one year)	1‡
Abdominal hernia	1
Superficial haematoma	1
Ileus	1
Temporary warm left foot	2
Temporary nausea at one year follow-up	1
Neurological deterioration:	
Motor deficit at two year follow-up	0
Temporary motor deficit	0
Sensory loss at two year follow-up	2
Temporary sensory loss	4
Radicular pain at two year follow-up	2
Temporary radicular pain	4
Infection:	
Superficial wound infection	0
Deep wound infection	0
Urinary tract infection	0
Total No (%) complications during two year follow-up	26 (34)
Additional spinal surgery within 2 years:	
Fusion	2§
Other	2¶

*Repeat surgery with insertion of new polyethylene inlay.
†Associated with temporary slightly colder foot at follow-up.
‡One patient reported retrograde ejaculation at baseline but not at one year follow-up, one at baseline and at follow-up, and one at follow-up but without baseline information.
§Fusion at level with disc prosthesis and level above.
¶Resection of spinous process because of possible painful contact between adjacent levels.

Likert scale, and satisfaction with care on a five point Likert scale.[26] Further daily consumption of drugs was registered. Patients attended for follow-up visits at six weeks, three and six months, and one and two years (the main end point of follow-up was at two years). At two years we sent a questionnaire including the most important outcome measures to 29 of the 34 patients who were lost to follow-up (see table B in appendix 3 on bmj.com).

At the two year follow-up, two independent observers blinded to treatment evaluated patients using the back performance scale (consists of five tests with a score ranging from 0 to 15, worst possible)[27] and the Prolo scale (consists of functional and economic parts, which are summed to a worst score of 2 and a best score of 10).[28] Patients were informed before this session not to reveal the treatment received, and had tape placed on their abdominal wall to hide the scarring from the operation. We also carried out a full health economic analysis, which will be reported elsewhere.

Statistical considerations

The trial was designed to have 80% power to detect a significant difference of at least 10 points in change in the mean Oswestry disability index score between the intervention groups at two year follow-up.[5] Baseline standard deviation was estimated at 18.[18] Considering these assumptions and adding 25% for a multicentre study design and 30% for possible drop-outs, we estimated we required 180 patients.

Planned analyses

The main statistical analysis was in the intention to treat population at one and two year follow-up. According to our protocol the analysis was performed with the assumption that patients who dropped out had no improvement after drop-out (last value carried forward). We also determined if different centres had different outcomes. We used χ^2 test or Fisher's exact test to analyse categorical variables and independent two sided t test or analysis of variance to analyse continuous variables. A significance level of 5% was used throughout. All statistical analyses were performed with SPSS version 16.0. We did not adjust for significantly different baseline scores.

Unplanned analyses (analyses not recorded in the original protocol)

We conducted a per protocol analysis for the primary outcome variable (score on Oswestry disability index). Consistent with criteria from the Food and Drug Administration,[8] we considered an individual change in score of at least 15 points from baseline to two year follow-up as a minimal important change. A deterioration of 6 points in the score was considered a "change for the worse."[29] We calculated the number needed to treat with confidence intervals.[30] A mixed model analysis was used to evaluate the effect of each efficacy

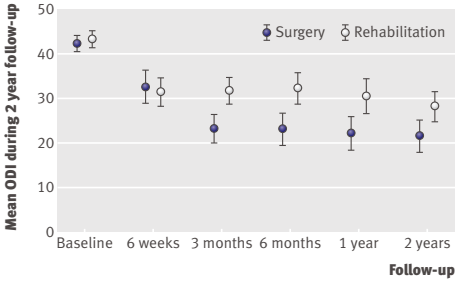

Fig 2 | Primary outcome variable within intention to treat mixed model analysis. Mean difference in Oswestry disability index (ODI) was 6.9 points at two year follow-up, P<0.001 (adjusted for baseline index)

Table 3 | Planned analysis of primary outcome in patients with low back pain and degenerative disc randomised to disc prosthesis surgery or rehabilitation. Mean (SD) outcome values on Oswestry disability index (ODI) at 12 and 24 months and treatment effect

	Mean outcome			
	Surgery	Rehabilitation	Treatment effect* (95% CI)	P value†
Baseline	41.8 (9.1)	42.8 (9.3)	—	—
1 year	22.3 (17.0)	33.0 (16.6)	−10.0 (−15.0 to −5.0)	‹0.001
2 years	21.2 (17.1)	30.0 (16.0)	−8.4 (−13.2 to −3.6)	0.001

ODI=see footnote for table 1 for scale details.
*Difference between groups in mean change from baseline.
†Two sided *t* test.

variable over time and between groups. In the mixed model patients were not excluded from the analysis of an efficacy variable if the variable was missing at some, but not all, time points after baseline. In the additional analysis (categorical or ordinal data at two year follow-up), missing data were not replaced. Significantly different baseline scores were not adjusted for in the longitudinal model. Each outcome variable was adjusted for the baseline values of the variable.

RESULTS

Of the 605 patients screened for eligibility, 173 were included in the study and treated between April 2004 and September 2007 (86 with surgery and 87 with rehabilitation) (fig 1). The drop-out rate from inclusion to two year follow-up was 20% (n=34) (15% (n=13) in the surgical arm and 24% (n=21) in the rehabilitation arm). Five patients (6%) crossed over from rehabilitation to surgery, but none crossed from surgery to rehabilitation. Of the 34 patients lost to follow-up, 26 answered a questionnaire two and a half to five years after treatment (see table B in appendix 3 on bmj.com).

Patients' characteristics

Most baseline characteristics were similar in the two treatment groups (table 1). Low back pain score and SF-36 mental health subscores, however, were significantly worse in the rehabilitation group than in the surgery group.

Surgical treatment and complications

Of the patients randomised to surgery, 25 (33%) underwent two level surgery. Median surgical time was 165 minutes (range 72-570 minutes) and median blood loss was 310 ml (range 50-6000 ml) (table 2). Four patients had bleeding of more than 1500 ml.

Six patients (8%) had complications resulting in impairment at two year follow-up, and the reoperation rate was 6.5% (n=5) (table 2). One patient had a serious complication: at the three month follow-up, the polyethylene inlay was found to be dislodged. During revision surgery, injury to the left common iliac artery led to compartment syndrome resulting in a lower leg amputation. One patient reported retrograde ejaculation at one year follow-up. At two year follow-up, two patients reported sensory loss in the thigh and two patients reported new radicular pain. In addition, one patient had an arterial thrombosis of the dorsalis pedis artery, which temporarily resulted in a slightly colder foot. Table 2 presents further complications. Two patients had an additional fusion and two patients had partial resection of the spinous processes because of persistent back pain.

Primary outcome

Planned analyses according to protocol

The mean change Oswestry disability index score from baseline to two year follow-up was 20.8 (95% confidence interval 16.4 to 25.2) in the surgery group and 12.4 (8.5 to 16.3) in the rehabilitation group (table 3). The mean treatment effect (difference between groups) at two year follow-up was −8.4 (−13.2 to −3.6) in the intention to treat analysis (last value carried forward). Subgroup analysis showed no differences in the main outcome variable between centres and level(s) operated on.

Unplanned analyses

In the mixed model analysis, the Oswestry score improved significantly more in the surgical group than in the rehabilitation group at all time points, in both the intention to treat (fig 2) and per protocol analyses (table 4). The mean change from baseline to two year follow-up was 22.5 (intention to treat) (95% confidence interval 18.5 to 26.4) in the surgery group and 15.6 (intention to treat) (11.7 to 19.5) in the rehabilitation group. The mean treatment effect (difference between groups) at two year follow-up was 6.9 (2.1 to 11.7) in the intention to treat analysis. In an analysis in

Table 4 | Unplanned analysis of primary outcome in patients with low back pain and degenerative disc randomised to disc prosthesis surgery or rehabilitation. Mean (SD) outcome values on Oswestry disability index (ODI) at follow-up and treatment effect (difference (95% confidence interval)), minus values indicating larger improvement in outcome with surgery

	Intention to treat analysis			Per protocol analysis		
	Surgery	Rehabilitation	Treatment effect	Surgery	Rehabilitation	Treatment effect*
Baseline	41.8 (9.1)	42.8 (9.3)	—	42.2 (9.2)	42.1 (8.3)	—
6 weeks	31.5 (17.2)	30.2 (13.6)	1.3 (−3.5 to 6.1)	31.1 (17.3)	29.6 (13.5)	1.7 (−3.1 to 6.6)
3 months	21.5 (14.1)	30.6 (13.1)	−9.1 (−13.9 to −4.3)	20.7 (13.5)	30.3 (12.7)	−9.5 (−14.4 to −4.6)
6 months	21.4 (16.3)	31.1 (14.9)	−9.7 (−14.6 to −4.8)	20.7 (15.9)	29.9 (14.6)	−9.2 (−14.2 to −4.2)
1 year	20.3 (17.2)	29.2 (16.1)	−8.9 (−13.8 to −4.0)	19.7 (16.4)	27.0 (15.0)	−7.3 (−12.3 to −2.3)
2 years	19.8 (16.7)	26.7 (14.5)	−6.9 (−11.7 to −2.1)	18.8 (15.8)	26.9 (13.9)	−8.1 (−12.9 to −3.2)

ODI=see footnote for table 1 for scale details.
*All P‹0.001 for trend in treatment effect over time. Two sided *t* test.

Table 5 | Planned analysis of secondary outcomes in patients with low back pain and degenerative disc randomised to disc prosthesis surgery or rehabilitation. Mean (SD) values at 12 and 24 months (unless stated otherwise) and treatment effect

Variable	Mean outcome		Treatment effect (95% CI)*	P value†
	Surgery	Rehabilitation		
Back pain score‡:				
Baseline	64.9 (15.3)	73.6 (13.9)		
1 year	35.6 (28.6)	53.2 (28.4)	−14.0 (−23.0 to −5.0)	0.003
2 years	35.4 (29.1)	49.7 (28.4)	−12.2 (−21.3 to −3.1)	0.009
SF-36 physical component summary:				
Baseline	30.5 (7.1)	30.8 (6.5)		
1 year	42.8 (12.2)	37.3 (11.0)	5.5 (1.9 to 9.1)	0.003
2 years	43.3 (11.7)	37.7 (10.1)	5.8 (2.5 to 9.1)	0.001
SF-36 mental component summary‡:				
Baseline	47.7 (13.0)	45.2 (13.2)		
1 year	50.2 (12.0)	49.2 (13.2)	0.2 (−3.5 to 3.8)	0.90
2 years	50.7 (11.6)	48.6 (12.8)	1.0 (−2.4 to 4.4)	0.50
EQ-5D:				
Baseline	0.30 (0.27)	0.27 (0.31)		
1 year	0.68 (0.34)	0.55 (0.32)	0.13 (0.01 to 0.25)	0.04
2 years	0.69 (0.33)	0.63 (0.28)	0.06 (−0.05 to 0.18)	0.26
HSCL-25				
Baseline	1.81 (0.50)	1.88 (0.51)		
1 year	1.51 (0.49)	1.67 (0.52)	−0.12 (−0.26 to 0.02)	0.10
2 years	1.50 (0.44)	1.63 (0.52)	−0.10 (−0.23 to 0.04)	0.20
FABQ work:				
Baseline	25.8 (11.2)	27.4 (27.4)		
1 year	19.2 (14.2)	23.1 (13.0)	−2.7 (−6.5 to 1.1)	0.20
2 years	18.1 (13.9)	21.2 (12.8)	−2.1 (−6.0 to 1.7)	0.30
FABQ physical:				
Baseline	14.0 (5.8)	12.5 (5.6)		
1 year	8.8 (6.7)	9.7 (5.8)	−1.3 (−3.2 to 0.6)	0.20
2 years	9.0 (6.8)	9.9 (6.0)	−1.5 (−3.4 to 0.5)	0.10
Self efficacy:				
Baseline	3.4 (1.5)	3.6 (1.6)		
1 year	6.3 (3.3)	5.2 (2.4)	1.2 (0.3 to 2.1)	0.01
2 years	6.1 (2.9)	5.3 (2.5)	1.0 (0.2 to 1.9)	0.02
No (%) returned to work at 2 years§:	21 (31)	15 (23)	—	0.31
No (%) satisfied with outcome at 2 years¶	46 (63)	26 (39)	—	0.005
No (%) satisfied with care at 1 year**	66 (90)	48 (73)	—	0.011
No (%) with drug consumption at 2 years††	16 (22)	14 (18)	—	0.30
Back performance scale at 2 years‡‡	3.2 (3.0)	4.0 (3.0)	−0.8 (−1.8 to 0.2)	0.10
Prolo scale at 2 years§§	7.0 (2.3)	6.1 (1.9)	0.9 (0.1 to 1.6)	0.019

ED-5Q score ranges from −0.59 to 1 (1 equals perfect health); self efficacy beliefs for pain scores ranges from 1 to 10 and are summarised and divided by 5. Lower scores indicate that he/she is very uncertain if he/she is able to manage pain. For other scores see footnote to table 1.
*Treatment effect is difference between groups in mean change from baseline. Positive value in SF-36, EQ-5D, self efficacy for pain, and Prolo scale and negative values in remaining variables indicate larger improvement in outcome with surgery.
†Two sided *t* test for continuous variables and Pearson's χ^2 test for categorical variables.
‡Values not adjusted for significantly different baseline scores.
§Net back to work rate calculated by subtracting patients who went back to work from patients who stopped working.
¶7 point Likert scale (1=completely recovered, 2=much recovered to 7=vastly worsened); slightly improved not included as satisfied with outcome.
**4 point global rating scale, not including slightly satisfied as satisfied with care.
††Use of drugs daily or not.
‡‡Scale comprises five tests with score ranging from 0 to 15 (worst possible).
§§Scale comprises functional and economic parts, summed to give worst score of 2 and best score of 10.

which the patient with lower leg amputation was given worst score in the group, the difference between the groups remained significant (P<0.001). Some 70% (n=51) of the patients in the surgery group and 47% (n=31) of the patients in the rehabilitation group had an improvement in Oswestry score of at least 15 points (P<0.006) (intention to treat). The number needed to treat was 4.4 (2.6 to 14.5). Worsening of low back pain was experienced by 11% (n=8) of the surgical group and 9% (n=6) of the rehabilitation group. Subgroup analysis showed no differences in the main outcome variable between centres and level(s) operated on.

Secondary outcomes

Planned analyses according to protocol

Low back pain, SF-36 physical summary, and patients' satisfaction improved significantly more in the surgical group than the rehabilitation group at two year follow-up (table 5). The mean difference between the groups in change from baseline to two year follow-up was −12.2 (95% confidence interval −21.3 to −3.1) for low back pain and 5.8 (2.5 to 9.1) for SF-36 physical summary. On the seven point global rating scale at two years, 63% (46) of patients in the surgery group and 39% (26) in the rehabilitation group (P=0.005 for difference between treatment groups) considered themselves completely recovered or much improved. Self efficacy for pain favoured the surgical group. SF-36 mental summary, EQ-5D, FABQ work and physical, HSCL-25, return to work, and drug consumption did not differ at two year follow-up. At the start of the study, 28% (46) of patients were at work full or part time; at two year follow-up, this had increased to 56% (n=74). There was a "net back to work" rate of 31% (n=21) in the surgical group and 23% (n=15) in the rehabilitation group (P=0.31) (table 5). Scores on the back performance scale did not differ significantly between the groups (−0.8, −1.8 to 0.2; P=0.10). The Prolo sum score favoured the surgical group, with a mean difference of 0.9 (0.1 to 1.6; P=0.019).

Unplanned analyses

In the mixed model analysis, low back pain (table 6), SF-36 physical summary (table 8), and EQ-5D, HSCL-25, and self efficacy for pain (table 9) improved significantly more in the surgical group than the rehabilitation group at all time points. The mean difference between the groups in change from baseline to two year follow-up for low back pain was −12.7 (95% confidence interval −21.1 to −4.2, table 6) and SF-36 physical summary 4.3 (0.8 to 7.9, table 8). Further analyses are shown in tables 7, 8, and 9.

DISCUSSION

This randomised trial comparing disc prosthesis with multidisciplinary rehabilitation showed a significant difference in the primary outcome variable (Oswestry disability index after two years) in favour of surgery. The difference between groups of 8.4 points on the index (with intention to treat analysis) at two year follow-up, however, was smaller than the difference of 10

Table 6 | Unplanned analysis in secondary outcome in patients with low back pain and degenerative disc randomised to disc prosthesis surgery or rehabilitation. Mean (SD) outcome values for back pain* at follow-up and treatment effect (difference (95% confidence interval))

	Intention to treat analysis		
	Surgery	Rehabilitation	Treatment effect†
Baseline	64.9 (15.3)	73.6 (13.9)	—
6 weeks	34.7 (27.5)	51.1 (24.6)	−16.5 (−24.8 to −8.2)
3 months	29.3 (25.0)	55.4 (23.4)	−26.2 (−34.5 to −17.8)
6 months	36.1 (28.5)	50.0 (24.5)	−13.8 (−22.3 to −5.3)
1 year	33.0 (29.4)	48.7 (28.9)	−15.7 (−24.3 to −7.0)
2 years	32.7 (28.8)	45.3 (28.6)	−12.7 (−21.1 to −4.2)

*See table 1 for score details.
†Negative values indicate larger improvement in outcome with surgery. All P<0.001 for trend in treatment effect over time. Two sided *t* test.

points that the study was designed to detect. As evident in the confidence intervals, the data are consistent with a wide range of differences between the groups, including values well below 10 points. There is, as far as we know, no agreement on the size of the clinically important difference between two treatment groups. As an alternative we can assess the proportion of patients achieving a clinically meaningful improvement.[31] By using a clinically meaningful improvement for an individual patient of 15 points on the Oswestry disability index,[8] 70% (n=51) of patients in the surgical group and 47% (n=31) of those in the rehabilitation group achieved at least this improvement (intention to treat). We will publish data on the estimated minimal clinically important change elsewhere, but the changes are in agreement with recommendations from FDA studies. As there is no consensus based agreement of how large a difference between groups must be to be of clinical importance it is impossible to conclude whether the effect found in our study is of clinical importance. As such a decision must be made before a new treatment can be recommended in clinical practice; our study underlines the need for such a consensus agreement.

The change in the Oswestry disability index score in our study is comparable with those seen in previous studies. In our study, the mean score was reduced by 29% (12.4 points) in the rehabilitation group (intention to treat analysis). Brox et al[4] found a similar reduction of 29% (12.0 points) at one year follow-up, while Fairbank et al[6] and Fritzell et al[3] observed a smaller reduction at two year follow-up (8.7 and 5.5 points, respectively). In our study, there was a mean reduction in score of 50% (20.8 points) in the surgical arm (intention to treat analysis). Similar reductions have been reported in other studies,[8 9 11] though Zigler et al used the "chiropractor version" of the Oswestry index.[32] This questionnaire has not been sufficiently validated and consequently it is difficult to compare the outcome.[18]

It could be argued that patients who withdrew after randomisation or dropped out during or after treatment had a superior or inferior outcome. We therefore sent a questionnaire to such patients. The nine patients who withdrew after surgery experienced a reduction in Oswestry score of 30.2 (SD 4.5) points. The six who withdrew after rehabilitation had a reduction of 11.8 (SD 3.0), and the 11 patients who withdrew without treatment had no change (1.0 (SD 4.5) points) (see table B in appendix 3 on bmj.com). This might support the assumption of no improvement in outcome after drop-out, justifying use of the last value carried forward analysis.

Most changes in secondary variables measuring disability and pain favoured surgical treatment, though there were no significant differences between groups in FABQ work, FABQ physical, SF-36 mental health, EQ-5D, HSCL-25, drug consumption, return to work, and the back performance scale in the main analysis. In the surgical group we found a similar "net back to work" rate as reported by Fritzell et al.[3] Nevertheless, it has been argued that sick leave, to a large extent, is influenced by factors outside the domain of medical and therapeutic interventions.[33] The somewhat smaller difference between groups in the back performance scale than in the Oswestry disability index might be explained by differences in psychometric properties between the outcome measurements or by patients overstating the effect in a subjective questionnaire.

Table 7 | Unplanned analysis in secondary outcomes in patients with low back pain and degenerative disc randomised to disc prosthesis surgery or rehabilitation. Mean (SD) outcome values for SF-36*

	Baseline		3 months		6 months		1 year		2 years		
Variable	Surgery	Rehabilitation	Surgery	Rehabilitation	Surgery	Rehabilitation	Surgery	Rehabilitation	Surgery	Rehabilitation	P value†
Physical function	52.7 (17.6)	50.6 (17.8)	76.2 (18.3)	64.0 (22.4)	76.0 (21.4)	63.7 (21.0)	79.5 (20.6)	66.7 (22.9)	78.9 (20.2)	69.6 (22.2)	<0.001
Role physical	25.3 (24.2)	23.9 (18.7)	50.0 (31.5)	45.6 (31.9)	57.2 (35.1)	47.8 (31.2)	58.9 (37.3)	55.9 (33.9)	66.4 (33.5)	55.1 (35.0)	0.135
Bodily pain	24.9 (16.5)	24.4 (12.1)	48.2 (22.4)	34.9 (16.1)	50.8 (29.1)	39.1 (20.8)	52.5 (30.8)	43.5 (24.6)	55.5 (29.1)	44.4 (23.0)	<0.001
General health	57.9 (19.7)	55.9 (19.9)	67.6 (22.8)	60.7 (24.7)	65.5 (24.3)	60.1 (24.4)	68.1 (26.8)	61.7 (22.1)	65.7 (26.0)	61.1 (24.8)	0.125
Vitality	37.8 (20.2)	33.1 (20.0)	50.3 (21.6)	44.4 (22.1)	55.6 (23.7)	45.7 (22.9)	57.5 (27.5)	48.2 (24.9)	55.0 (27.1)	46.8 (23.5)	0.003
Social function	53.0 (30.6)	57.6 (26.7)	72.8 (25.0)	68.8 (25.6)	75.0 (28.6)	71.1 (26.7)	76.7 (25.7)	74.3 (26.8)	78.3 (26.8)	77.9 (27.4)	0.725
Role emotion	72.5 (33.3)	67.6 (32.7)	85.1 (23.3)	69.0 (34.7)	83.3 (26.3)	74.5 (29.8)	80.4 (31.0)	79.2 (26.3)	83.9 (25.6)	79.2 (29.0)	0.010
Mental health‡	71.7 (18.0)	65.8 (18.7)	78.6 (15.6)	72.4 (17.9)	79.5 (16.8)	74.1 (16.4)	80.4 (17.5)	73.8 (20.9)	78.3 (18.2)	75.8 (17.5)	0.007

*See table 1 for score details.
†For trend in treatment effect over time.
‡Values are not adjusted for significantly different baseline scores.

Table 8 | Unplanned analysis in secondary outcome in patients with low back pain and degenerative disc randomised to disc prosthesis surgery or rehabilitation. Mean (SD) outcome values for physical and mental component summary scores on SF-36* at follow-up and treatment effect (difference (95% confidence interval))

	Surgery	Rehabilitation	Treatment effect†
SF-36 physical component summary			
Baseline	30.5 (7.1)	30.8 (6.5)	—
3 months	40.3 (10.9)	37.3 (8.9)	3.0 (−0.6 to 6.6)
6 months	41.4 (12.3)	37.2 (9.2)	4.2 (0.6 to 7.8)
1 year	43.5 (12.7)	39.4 (11.5)	4.2 (0.6 to 7.7)
2 years	43.9 (11.9)	39.6 (10.4)	4.3 (0.8 to 7.9)
SF-36 mental component summary			
Baseline	47.7 (13.0)	45.2 (13.2)	—
3 months	50.9 (10.4)	47.0 (12.9)	3.9 (−0.2 to 8.0)
6 months	52.0 (9.7)	49.5 (10.5)	2.5 (−1.6 to 6.6)
1 year	51.7 (11.6)	49.7 (12.0)	2.0 (−2.0 to 6.1)
2 years	51.0 (11.0)	50.5 (11.0)	0.5 (−3.4 to 4.5)

*See table 1 for score details.
†Positive treatment effect indicates larger improvement in outcome for surgery. P=0.002 for physical and 0.166 for mental for trend in treatment effect over time.

Table 9 | Secondary outcomes in patients with low back pain and degenerative disc randomised to disc prosthesis surgery or rehabilitation. Mean (SD) outcome values on EQ-5D, HSCL-25, FABQ, and self efficacy at follow-up and treatment effect (difference (95% confidence interval))

Variable*	Surgery	Rehabilitation	Treatment effect†
EQ-5D‡			
Baseline	0.30 (0.30)	0.27 (0.31)	—
6 weeks	0.59 (0.30)	0.55 (0.29)	0.04 (−0.06 to 0.15)
3 months	0.70 (0.23)	0.48 (0.31)	0.22 (0.12 to 0.33)
6 months	0.68 (0.28)	0.51 (0.33)	0.17 (0.06 to 0.27)
1 year	0.67 (0.35)	0.54 (0.32)	0.13 (0.02 to 0.23)
2 years	0.68 (0.34)	0.60 (0.30)	0.08 (−0.03 to 0.18)
HSCL-25§			
Baseline	1.81 (0.50)	1.88 (0.51)	—
3 months	1.38 (0.34)	1.66 (0.51)	−0.27 (−0.44 to −0.11)
6 months	1.44 (0.45)	1.66 (0.49)	−0.22 (−0.38 to −0.05)
1 year	1.45 (0.50)	1.59 (0.49)	−0.14 (−0.30 to 0.03)
2 years	1.47 (0.49)	1.55 (0.50)	−0.09 (−0.25 to 0.08)
FABQ work§			
Baseline	25.8 (11.2)	27.4 (9.9)	—
3 months	20.0 (12.9)	24.3 (11.9)	−4.3 (−8.6 to 0.1)
6 months	18.7 (12.9)	23.0 (12.7)	−4.3 (−8.7 to 0.1)
1 year	18.2 (13.9)	21.3 (13.2)	−3.1 (−7.4 to 1.2)
2 years	16.7 (13.5)	18.5 (12.5)	−1.8 (−6.1 to 2.5)
FABQ physical§			
Baseline	14.0 (5.8)	12.5 (5.6)	—
3 months	8.8 (5.3)	9.1 (6.3)	−0.3 (−2.4 to 1.7)
6 months	8.6 (6.3)	9.3 (6.7)	−0.7 (−2.8 to 1.3)
1 year	8.0 (6.3)	8.9 (5.8)	−0.8 (−2.9 to 1.2)
2 years	8.0 (6.0)	8.3 (5.7)	−0.3 (−2.3 to 1.7)
Self efficacy‡			
Baseline	3.4 (1.5)	3.6 (1.6)	—
3 months	6.1 (2.3)	5.0 (2.2)	1.0 (0.2 to 1.9)
6 months	6.0 (2.6)	5.6 (2.4)	0.5 (−0.3 to 1.3)
1 year	6.4 (3.3)	5.5 (2.5)	0.9 (0.0 to 1.7)
2 years	6.2 (2.7)	5.6 (2.7)	0.5 (-1.4 to 2.8)

*See tables 1 and 5 for score details.
†EQ-5D P<0.001, HSCL P<0.001, FABQ work P=0.057, FABQ physical P=0.548, self efficacy P=0.019 for trend in treatment effect over time.
‡Positive scores indicate larger improvement in outcome with surgery.
§Negative scores indicate larger improvement in outcome with surgery.

Strengths and limitations

Our study has several strengths. It was randomised and had few patients who crossed over to the other treatment regimen. In addition, an independent research assistant collected the data, the observers at the two year evaluation were blinded, the interventions were standardised, and the financing of the study was public. Choosing magnetic resonance imaging criteria for inclusion could be a strength or limitation. To our knowledge, there are no specific criteria to determine which degenerative changes should be operated on. When designing the study we wanted the inclusion of patients across centres to be as unanimous as possible, treating the same population, although this possibly would lead to less external validity of the study. It could also possibly lead to inclusion of more severe degenerated discs in our study compared with other studies.[8 9]

One limitation of our study is the lack of a placebo or sham group. The regression to the mean and the natural resolution of chronic low back pain must also be considered in both groups. When balancing a non-operative regimen with an operative treatment, there is probably a difference in placebo effect that is difficult to untangle from the treatment effect.[34-37] The placebo effect might be higher in the surgical group, although the possible placebo effect of rehabilitation over several weeks with personal contact with a therapist should not be underestimated. Furthermore, it could be argued that the patients included in the study wanted surgery, but the number of patients not wanting the rehabilitation programme was similar to the number of patients not wanting surgery (see figure and appendix 1 on bmj.com). Brox et al found no difference in treatment effect between patients who did and did not "believe" in surgery,[4 5] and a recent study found no significant relation between baseline expectations and follow-up scores.[38] On the other hand, "expectation being fulfilled" might be a predictor of global outcome.[38] During the inclusion process, we emphasised the advantages and disadvantages of the two treatment options and that none of the treatments are documented as superior to another. It is still possible, however, that patients in the rehabilitation group found themselves faced with "more of the same." The lack of routine rehabilitation in the surgical arm could be another limitation in the study. We wanted to avoid the postoperative treatment containing elements from the rehabilitation programme. Hence, patients received only general advice when they were discharged from the hospital and received no rehabilitation in the first weeks after surgery. At six weeks, however, patients could be referred if required to a physiotherapist at their home for functional mobilisation and general muscle training.

WHAT IS ALREADY KNOWN ON THIS TOPIC

In patients with chronic low back pain, compared with fusion, the clinical outcome with disc prosthesis has been at least equivalent

Compared with multidisciplinary rehabilitation, improvement in disability and pain are similar

WHAT THIS STUDY ADDS

Surgery with disc prosthesis resulted in a significantly greater improvement in scores on the Oswestry disability index and variables measuring disability and pain, although the difference in Oswestry score between groups was lower than the study was designed to detect

There were no differences in return to work and several outcomes measuring mental health

Furthermore, some surgical patients underwent a second operation but repeat rehabilitation was not considered. Patients did not request a second chance for rehabilitation, though they were advised during follow-up consultations. Another weakness in our study is the difference in compliance between groups and the high drop-out rate. This difference in adherence to the protocol probably leads to an underestimate of the true effect of surgery, especially in the intention to treat analysis. In similar studies comparing surgery with rehabilitation, the drop-out rates were similar to ours.[6 39-41] The patients we included in our study were highly selected, with one or two level degenerative changes and good general health. Thus, our results are valid only in similar patients. Furthermore, we examined several secondary outcome variables that could lead to the detection of differences by chance. Although we conducted several unplanned analyses (not recorded in the original protocol), in common with similar studies, we consider it as an important asset to our data. Lately, similar studies have applied repeated measurements by using mixed models.[40] Using unplanned analysis could be considered a weakness, but our findings in these analyses support our main analyses and strengthen our conclusion. Nevertheless, caution should be used in interpreting the results of non-prespecified analyses.

Potential harms of disc prosthesis surgery

Surgery carries a risk of serious complications, as seen in one of our patients. In a review by Inamasu et al, the perioperative vascular injury rate for anterior lumbar interbody fusion was 0-18% (mean 3%).[42] This is an important drawback of surgery. No major differences in complication rates between insertion of a disc prosthesis and fusion have been found in a randomised setting.[8 9 11] The short term reoperation rate in our study was 6.5% (n=5) and the vascular injury rate was 6.5% (n=5) (table 2). Although vascular complications are reported, serious consequences like amputation and mortality are rare.[43] Recently Kurtz et al looked at the rates of short term revision and mortality total disc replacement.[43] They found similar reoperation rates as with anterior fusion surgery and hip arthroplasty. Four retrospective studies have reported long term reoperation rates of up to 13%.[44-47] Data on the anterior revision rate of the prosthesis is difficult to extract from these studies but seems considerably lower. The potential long term revision rate with a higher complication rate on revisions needs to be considered.[48]

Earlier addressed but unresolved questions are the incidence of adjacent level degeneration after total disc replacement and distinct characteristics of patients associated with good outcome. Some studies have examined these issues but more information is needed.[49-51] In a univariate analysis we found indications that patients with Modic I or II changes have a superior result in the surgical arm and that patients with high Oswestry scores seem to be more suitable for rehabilitation. A full multivariate analysis of good outcomes will be published soon to answer these questions. Another important issue is the incidence of degeneration in the facet joints of the operated level. An analysis of adjacent level degeneration and degeneration of the operated level in addition to a full health economic analysis will be published later.

The total blood loss and operation time were higher in our study than in similar studies. The learning curve might be quite flat, and perhaps the participating surgeons should have carried out disc prosthesis surgery in more patients before the start of the study. Using a surgeon to expose the disc (access surgeon), might also have reduced the blood loss and operation time. Blumenthal et al and Zigler et al performed one level surgery, while a third of our patients underwent two level surgery.[8 9] This could explain some of the increased blood loss and operation time in our study. Because of the complexity of the surgery and the risk of serious complications, we think this kind of surgery should be confined to a few specialist centres with experienced spine surgeons and available vascular surgeons. A high quality rehabilitation programme should be available.

Our study was not designed to evaluate specific mechanisms of reduction of pain and disability. Possible explanations for the pain reduction are removal of the disc in the surgical group and better coping in the rehabilitation group, but the patients were heterogeneous and probably had a mixed aetiology difficult to separate. Even though we did not have a control group, the mixed causes of chronic low back pain, the association of surgery with potentially serious complications, and the considerable improvement in the rehabilitation group suggest that it is reasonable to consider a rehabilitation programme before surgery.

We thank the patients participating in the study; Coast Hospital for Physical Medicine and Rehabilitation, Stavern, for videos and material for lectures for the rehabilitation intervention; Hege Andresen at St Olavs Hospital, Trondheim, for data coordination; Per Farup at St Olavs Hospital, Trondheim, for organising the web randomisation system; Astrid Woodhouse and Kirsti Vanvik from St Olavs Hospital for performing the two year control; and Lucy Hyatt for paid editorial assistance.
The Norwegian Spine Study Group
University Hospital North Norway, Tromso (eight patients): Odd-Inge Solem (department of orthopaedic surgery), Jens Munch-Ellingsen (department of neurosurgery), and Franz Hintringer, Anita Dimmen Johansen, Guro Kjos (department of physical medicine and rehabilitation).

Trondheim University Hospital, Trondheim (21 patients): Hege Andresen, Helge Rønningen, Kjell Arne Kvistad (national centre for spinal disorders, department of neurosurgery), Bjørn Skogstad, Janne Birgitte Børke, Erik Nordtvedt, Gunnar Leivseth (multidiscipline spinal unit, department of physical medicine and rehabilitation).
Haukeland University Hospital, Bergen (64 patients): Sjur Braaten, Turid Rognsvåg, Gunn Odil Hirth Moberg (Kysthospitalet in Hagevik, department of orthopaedic surgery), Jan Sture Skouen, Lars Geir Larsen, Vibeche Iversen, Ellen H Haldorsen, Elin Karin Johnsen, Kristin Hannestad (Outpatient Spine Clinic, department of physical medicine and rehabilitation).
Stavanger University Hospital, Stavanger (27 patients): Endre Refsdal (department of orthopaedic surgery).
Oslo University Hospital, Oslo (53 patients): Vegard Slettemoen, Kenneth Nilsen, Kjersti Sunde, Helenè E Skaara (department of orthopaedics), Anne Keller, Berit Johannessen, Anna Maria Eriksdotter (department of physical medicine and rehabilitation).

Contributors: All authors had full access to the data, were responsible for study concept and design, and critically revised the manuscript for important intellectual content. Acquisition of data: CH, LGJ, KS, OPN, MR, OG acquired the data, which were analysed and interpreted by HC , LGJ, KS, OPN, JIB, LS, IR, and OG. CH drafted the manuscript. CH and LS did the statistical analysis. CH, LJ, KS, OPN, JIB, MR, and OG provided administrative, technical, or material support. CH, KS, OPN, JIB, IR, LS, and OG supervised the study. CH is guarantor.

Funding: The study was funded by the South Eastern Norway Regional Health Authority and EXTRA funds from the Norwegian Foundation for Health and Rehabilitation, through the Norwegian Back Pain Association.

Competing interests: All authors have completed the Unified Competing Interest form at www.icmje.org/coi_disclosure.pdf (available on request from the corresponding author) and declare: no support from any organisation for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

Ethical approval: The study was evaluated and approved by the regional committees for medical research ethics in east Norway and all participants gave written informed consent. We did not obtain participants' informed consent for data but the presented data are anonymised and risk of identification is low.

Data sharing: Dataset available from the corresponding author at christian.hellum@medisin.uio.no.

1 Waddell G. *The back pain revolution*. 2nd ed. Churchill Livingstone, 2004.
2 Thomas E, Silman AJ, Croft PR, Papageorgiou AC, Jayson MI, Macfarlane GJ. Predicting who develops chronic low back pain in primary care: a prospective study. *BMJ* 1999;318:1662-7.
3 Fritzell P, Hagg O, Wessberg P, Nordwall A, for the Swedish Lumbar Spine Study Group. 2001 Volvo Award Winner in Clinical Studies: lumbar fusion versus nonsurgical treatment for chronic low back pain: a multicenter randomized controlled trial from the Swedish Lumbar Spine Study Group. *Spine* 2001;26:2521-32.
4 Brox JI, Sorensen R, Friis A, Nygaard O, Indahl A, Keller A, et al. Randomized clinical trial of lumbar instrumented fusion and cognitive intervention and exercises in patients with chronic low back pain and disc degeneration. *Spine* 2003;28:1913-21.
5 Brox JI, Reikeras O, Nygaard O, Sorensen R, Indahl A, Holm I, et al. Lumbar instrumented fusion compared with cognitive intervention and exercises in patients with chronic back pain after previous surgery for disc herniation: a prospective randomized controlled study. *Pain* 2006;122:145-55.
6 Fairbank J, Frost H, Wilson-MacDonald J, Yu LM, Barker K, Collins R. Randomised controlled trial to compare surgical stabilisation of the lumbar spine with an intensive rehabilitation programme for patients with chronic low back pain: the MRC spine stabilisation trial. *BMJ* 2005;330:1233.
7 Brox JI, Nygaard O, Holm I, Keller A, Ingebrigtsen T, Reikeras O. Four-year follow-up of surgical versus non-surgical therapy for chronic low back pain. *Ann Rheum Dis* 2009;69:1643-8.
8 Blumenthal S, McAfee PC, Guyer RD, Hochschuler SH, Geisler FH, Holt RT, et al. A prospective, randomized, multicenter Food and Drug Administration investigational device exemptions study of lumbar total disc replacement with the CHARITE artificial disc versus lumbar fusion: part I: evaluation of clinical outcomes *Spine* 2005;30:1565-75.
9 Zigler J, Delamarter R, Spivak JM, Linovitz RJ, Danielson GO III, Haider TT, et al. Results of the prospective, randomized, multicenter Food and Drug Administration investigational device exemption study of the ProDisc-L total disc replacement versus circumferential fusion for the treatment of 1-level degenerative disc disease. *Spine* 2007;32:1155-62.
10 Guyer RD, McAfee PC, Banco RJ, Bitan FD, Cappuccino A, Geisler FH, et al. Prospective, randomized, multicenter Food and Drug Administration investigational device exemption study of lumbar total disc replacement with the CHARITE artificial disc versus lumbar fusion: five-year follow-up. *Spine J* 2009;9:374-86.
11 Berg S, Tullberg T, Branth B, Olerud C, Tropp H. Total disc replacement compared to lumbar fusion: a randomised controlled trial with 2-year follow-up. *Eur Spine J* 2009;18:1512-9.
12 Gibson JN, Waddell G. Surgery for degenerative lumbar spondylosis. *Cochrane Database Syst Rev* 2005;4:CD001352.
13 Van den Eerenbeemt KD, Ostelo RW, van Royen BJ, Peul WC, van Tulder MW. Total disc replacement surgery for symptomatic degenerative lumbar disc disease: a systematic review of the literature. *Eur Spine J* 2010;19:1262-80.
14 Masharawi Y, Kjaer P, Bendix T, Manniche C, Wedderkopp N, Sorensen JS, et al. The reproducibility of quantitative measurements in lumbar magnetic resonance imaging of children from the general population. *Spine* 2008;33:2094-100.
15 Modic MT, Steinberg PM, Ross JS, Masaryk TJ, Carter JR. Degenerative disk disease: assessment of changes in vertebral body marrow with MR imaging. *Radiology* 1988;166:193-9.
16 Aprill C, Bogduk N. High-intensity zone: a diagnostic sign of painful lumbar disc on magnetic resonance imaging. *Br J Radiol* 1992;65:361-9.
17 Luoma K, Riihimaki H, Luukkonen R, Raininko R, Viikari-Juntura E, Lamminen A. Low back pain in relation to lumbar disc degeneration. *Spine* 2000;25:487-92.
18 Fairbank JCTM, Pynsent PBP. The Oswestry disability index. *Spine* 2000;25:2940-53.
19 Grotle M, Brox JI, Vollestad NK. Cross-cultural adaptation of the Norwegian versions of the Roland-Morris disability questionnaire and the Oswestry disability index. *J Rehabil Med* 2003;35:241-7.
20 Ware JE Jr, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Med Care* 1992;30:473-83.
21 Ware JE Jr. SF-36 health survey update. *Spine* 2000;25:3130-9.
22 Euroquol Group. EuroQol—a new facility for the measurement of health-related quality of life. *Health Policy* 1990;16:199-208.
23 Derogatis LR, Lipman RS, Rickels K, Uhlenhuth EH, Covi L. The Hopkins symptom checklist (HSCL): a self-report symptom inventory. *Behav Sci* 1974;19:1-15.
24 Waddell G, Newton M, Henderson I, Somerville D, Main CJ. A fear-avoidance beliefs questionnaire (FABQ) and the role of fear-avoidance beliefs in chronic low back pain and disability. *Pain* 1993;52:157-68.
25 Lorig K, Chastain RL, Ung E, Shoor S, Holman HR. Development and evaluation of a scale to measure perceived self-efficacy in people with arthritis. *Arthritis Rheum* 1989;32:37-44.
26 Ostelo RW. Clinically important outcomes in low back pain. *Best Pract Res Clin Rheumatol* 2005;19:593-607.
27 Strand LI, Moe-Nilssen R, Ljunggren AE. Back performance scale for the assessment of mobility-related activities in people with back pain. *Phys Ther* 2002;82:1213-23.
28 Prolo DJ, Oklund SA, Butcher M. Toward uniformity in evaluating results of lumbar spine operations. A paradigm applied to posterior lumbar interbody fusions. *Spine* 1986;11:601-6.
29 Hagg O, Fritzell P, Nordwall A. The clinical importance of changes in outcome scores after treatment for chronic low back pain. *Eur Spine J* 2003;12:12-20.
30 Altman DG. Confidence intervals for the number needed to treat. *BMJ* 1998;317:1309-12.
31 Guyatt GH, Juniper EF, Walter SD, Griffith LE, Goldstein RS. Interpreting treatment effects in randomised trials. *BMJ* 1998;316:690-3.
32 Fairbank JC. Use and abuse of Oswestry disability index. *Spine* 2007;32:2787-9.
33 Nachemson A. Chronic pain—the end of the welfare state? *Qual Life Res* 1994;3(suppl 1):S11-7.
34 Pauza KJ, Howell S, Dreyfuss P, Peloza JH, Dawson K, Bogduk N. A randomized, placebo-controlled trial of intradiscal electrothermal therapy for the treatment of discogenic low back pain. *Spine J* 2004;4:27-35.
35 Freeman BJ, Fraser RD, Cain CM, Hall DJ, Chapple DC. A randomized, double-blind, controlled trial: intradiscal electrothermal therapy versus placebo for the treatment of chronic discogenic low back pain. *Spine* 30:2369-77.
36 Buchbinder R, Osborne RH, Ebeling PR, Wark JD, Mitchell P, Wriedt C, et al. A randomized trial of vertebroplasty for painful osteoporotic vertebral fractures. *N Engl J Med* 2009;361:557-68.
37 Rousing R, Hansen KL, Andersen MO, Jespersen SM, Thomsen K, Lauritsen JM. Twelve-months follow-up in forty-nine patients with acute/semiacute osteoporotic vertebral fractures treated

conservatively or with percutaneous vertebroplasty: a clinical randomized study. *Spine* 2010;35:478-82.
38 Mannion AF, Junge A, Elfering A, Dvorak J, Porchet F, Grob D. Great expectations: really the novel predictor of outcome after spinal surgery? *Spine* 2009;34:1590-9.
39 Weinstein JN, Lurie JD, Tosteson TD, Skinner JS, Hanscom B, Tosteson AN, et al. Surgical vs nonoperative treatment for lumbar disk herniation: the Spine Patient Outcomes Research Trial (SPORT) observational cohort. *JAMA* 2006;296:2451-9.
40 Weinstein JN, Lurie JD, Tosteson TD, Hanscom B, Tosteson AN, Blood EA, et al. Surgical versus nonsurgical treatment for lumbar degenerative spondylolisthesis. *N Engl J Med* 2007;356:2257-70.
41 Peul WC, van Houwelingen HC, van den Hout WB, Brand R, Eekhof JA, Tans JT, et al. Surgery versus prolonged conservative treatment for sciatica. *N Engl J Med* 2007;356:2245-56.
42 Inamasu J, Guiot BH. Vascular injury and complication in neurosurgical spine surgery. *Acta Neurochir (Wien)* 2006;148:375-87.
43 Kurtz SM, Lau E, Ianuzzi A, Schmier J, Todd L, Isaza J, et al. National revision burden for lumbar total disc replacement in the United States: epidemiologic and economic perspectives. *Spine (Phila Pa 1976)* 2010; published online 26 Feb.
44 Putzier M, Funk JF, Schneider SV, Gross C, Tohtz SW, Khodadadyan-Klostermann C, et al. Charite total disc replacement—clinical and radiographical results after an average follow-up of 17 years. *Eur Spine J* 2006;15:183-95.
45 Lemaire JP, Carrier H, Sariali E, Skalli W, Lavaste F. Clinical and radiological outcomes with the Charite artificial disc: a 10-year minimum follow-up. *J Spinal Disord Tech* 2005;18:353-9.
46 Tropiano P, Huang RC, Girardi FP, Cammisa FP Jr, Marnay T. Lumbar total disc replacement. Seven to eleven-year follow-up. *J Bone Joint Surg Am* 2005;87:490-6.
47 David T. Long-term results of one-level lumbar arthroplasty: minimum 10-year follow-up of the CHARITE artificial disc in 106 patients. *Spine* 2007;32:661-6.
48 McAfee PC, Geisler FH, Saiedy SS, Moore SV, Regan JJ, Guyer RD, et al. Revisability of the CHARITE artificial disc replacement: analysis of 688 patients enrolled in the US IDE study of the CHARITE artificial disc. *Spine* 2006;31:1217-26.
49 Siepe CJ, Zelenkov P, Sauri-Barraza JC, Szeimies U, Grubinger T, Tepass A, et al. The fate of facet joint and adjacent level disc degeneration following total lumbar disc replacement: a prospective clinical, x-ray, and magnetic resonance imaging investigation. *Spine (Phila Pa 1976)* 2010;35:1991-2003.
50 Siepe CJ, Mayer HM, Wiechert K, Korge A. Clinical results of total lumbar disc replacement with ProDisc II: three-year results for different indications. *Spine* 2006;31:1923-32.
51 Guyer RD, Siddiqui S, Zigler JE, Ohnmeiss DD, Blumenthal SL, Sachs BL, et al. Lumbar spinal arthroplasty: analysis of one center's twenty best and twenty worst clinical outcomes. *Spine* 2008;33:2566-9.

Accepted: 25 March 2011

Paper II

Is not included due to copyright

Paper III

Is not included due to copyright

Paper IV

Johnsen et al. BMC Musculoskeletal Disorders 2013, 14:148
http://www.biomedcentral.com/1471-2474/14/148

BMC Musculoskeletal Disorders

RESEARCH ARTICLE Open Access

Comparison of the SF6D, the EQ5D, and the oswestry disability index in patients with chronic low back pain and degenerative disc disease

Lars G Johnsen[1,2,3*†], Christian Hellum[4], Øystein P Nygaard[1,3,7], Kjersti Storheim[4,6], Jens I Brox[4], Ivar Rossvoll[1,2,3], Gunnar Leivseth[5] and Margreth Grotle[8,9†]

Abstract

Background: The need for cost effectiveness analyses in randomized controlled trials that compare treatment options is increasing. The selection of the optimal utility measure is important, and a central question is whether the two most commonly used indexes - the EuroQuol 5D (EQ5D) and the Short Form 6D (SF6D) – can be used interchangeably. The aim of the present study was to compare change scores of the EQ5D and SF6D utility indexes in terms of some important measurement properties. The psychometric properties of the two utility indexes were compared to a disease-specific instrument, the Oswestry Disability Index (ODI), in the setting of a randomized controlled trial for degenerative disc disease.

Methods: In a randomized controlled multicentre trial, 172 patients who had experienced low back pain for an average of 6 years were randomized to either treatment with an intensive back rehabilitation program or surgery to insert disc prostheses. Patients filled out the ODI, EQ5D, and SF-36 at baseline and two-year follow up. The utility indexes was compared with respect to measurement error, structural validity, criterion validity, responsiveness, and interpretability according to the COSMIN taxonomy.

Results: At follow up, 113 patients had change score values for all three instruments. The SF6D had better similarity with the disease-specific instrument (ODI) regarding sensitivity, specificity, and responsiveness. Measurement error was lower for the SF6D (0.056) compared to the EQ5D (0.155). The minimal important change score value was 0.031 for SF6D and 0.173 for EQ5D. The minimal detectable change score value at a 95% confidence level were 0.157 for SF6D and 0.429 for EQ5D, and the difference in mean change score values (SD) between them was 0.23 (0.29) and so exceeded the clinical significant change score value for both instruments. Analysis of psychometric properties indicated that the indexes are unidimensional when considered separately, but that they do not exactly measure the same underlying construct.

Conclusions: This study indicates that the difference in important measurement properties between EQ5D and SF6D is too large to consider them interchangeable. Since the similarity with the "gold standard" (the disease-specific instrument) was quite different, this could indicate that the choice of index should be determined by the diagnosis.

Keywords: Utility measures, Outcome assessment, Measurement properties, Health economics, Low back pain, Lumbar disc prosthesis, EQ5D, SF6D

* Correspondence: lars.gunnar.johnsen@ntnu.no
†Equal contributors
[1]Neuroclinic; National Center of spinal disorder, St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway
[2]Clinic of Orthopedics and Rheumatology, Department of Orthopaedic Surgery, St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway
Full list of author information is available at the end of the article

Background

An important way of assessing the effects of treatment in health economic evaluations is the use of utility indexes. The outcome scores of general health-related quality of life (HRQoL) questionnaires are stratified into different health states [1,2] that can then be validated in a community population [3,4]. Treatment benefit is thus expressed in a way that allows health states that are considered less preferable (0) to full health (1) to be given quantitative values. Because these quantitative values represent a valuation or preference of health states for the patients, they are called utility indexes (more utility for the patient with increasing value). When combined with a follow up period, health utility indexes are used to calculate quality-adjusted life years (QALYs). There are several utility indexes that could be used, and discrepancies exist regarding which index is most suitable [1,5]. These discrepancies could have implications for calculating cost-effectiveness when comparing alternative treatment options for the same disease [6-10]. Two of the most widely used indexes are the EuroQuol 5D (EQ5D) and the Short Form 6D (SF6D) [4,7].

Two papers assessed the impact that the measure has on cost-utility estimates [8,9]. Sach et al. found that the SF6D and EQ5D favored different treatment options for alleviating knee pain when applying the same cost per QALY threshold. Søgaard et al. [11] reported on the interchangeability of the two indexes. When plotting difference between change scores of SF6D and EQ5D against their average in a Bland-Altman plot , they found that the expected between-measure variation was 0.546 [12]. They conclude that although both indexes appear to be psychometrically valid for generic assessment of long-lasting back pain, the variation between them was too great to be considered interchangeable.

From other studies, we could hypothesize that there would be a discrepancy between the EQ5D and SF6D because of differences in valuing similar health states, evidence of a floor effect in the SF6D and a ceiling effect in the EQ5D, and because the SF6D can describe severe health states better than EQ5D [7,13,14].

Further work is required in this field to understand these discrepancies. Therefore, the aim of this study was to evaluate change scores values of the EQ5D and SF6D utility indexes in terms measurement error, structural validity, criterion validity, responsiveness, and interpretability according to the COSMIN taxonomy. The psychometric properties of the two utility indexes were compared to a disease-specific instrument, the Oswestry Disability Index (ODI), in the setting of a randomized controlled trial for degenerative disc disease.

Methods

Details about the RCT on which this work is based is reported in detail in Hellum et al. [15]. Between April 2004 and September 2007, 172 patients with diagnosed chronic low back pain and degenerative disc disease were randomized to either surgery with total disc replacement or multidisciplinary rehabilitation. The results from this study have been published previously [15].

Briefly, data were collected in a multicentre randomized controlled trial involving the five university hospitals in Norway. Inclusion criteria included age between 25 and 55 years, LBP for more than a year, degenerative changes in the intervertebral disc in one of the two lowest levels of the lumbar spine and an Oswestry Disability Index score of 30% points or more. Exclusion criteria included generalized chronic pain syndrome and degeneration established in more than two levels. Part of this study was an economic evaluation of chronic low back pain treatment. Patients were randomized to either surgery with insertion of an artificial disc or to non-surgical treatment (a multidisciplinary back rehabilitation program).

The outcomes of patients who completed the SF6D, EQ5D, and ODI at baseline and at 2-year follow up were included in this study.

Instruments

ODI

The ODI is a back-specific questionnaire [16,17]. Patients rate physical disability in activities of daily living due to low back pain in 10 questions, each of which has verbal response alternatives. Ratings are summed to yield a score ranging from 0 (not disabled at all) to 100 (completely disabled). We used the Norwegian translation of the validated questionnaire (version 2.0) [18].

SF6D

The SF6D utility index is comprised of 11 items from the SF-36 [19] that were revised into a six-dimensional health state classification system. The six dimensions are physical functioning, role limitations, social functioning, pain, mental health, and vitality. It reflects a continuous outcome scored on a 0.29–1.00 scale, with 1.00 indicating full health [3]. SF6D health states were evaluated against a normal population using the Standard Gamble (SG) method. We used the United Kingdom (UK) tariff [3]. The SF6D was calculated based on the Norwegian SF-36 (version 2) with the use of syntax files in SPSS 17 (SPSS, New York, US). The syntax files were kindly provided by Dr J. Brazier, University of Sheffield, UK.

EQ5D

For the EQ5D utility index, responses on a questionnaire with five dimensions, each comprised of three levels, are revised into an index with a range from −0.59–1, with 1.00 indicating full health. The 243 possible health states on the EQ5D are evaluated against a normal population using the time trade off method (TTO) [20,21]. We used

the Norwegian version of the EQ5D and syntax files obtained from the EQ5D society using the UK tariff to calculate the index.

Seven-point scale for patient assessment

Many authors suggest a seven-point scale to assess patient outcome in terms of a global score [22]. On the question: "How much benefit do you think you have had from the treatment you have received?" patients answered on a 7-category response scale that ranged from "I am completely disabled" to "I am completely recovered".

Data analysis

We followed the definitions and recommendations from The COSMIN (COnsensus-based Standards for the selection of health Measurement INstruments) checklist when analyzing the psychometric properties of the two utility indexes and ODI in this study [23].

If not otherwise mentioned, SPSS version 17 was used in the statistical analysis.

Measurement error

Measurement error concerns the systematic and random error of a patient`s score that is not attributed to true changes in the construct to be measured [24]. We used the standard error of measurement (SEM) to express instrument imprecision [22,25-27]. The advantage of using SEM is that it is considered to be an attribute of the measure and not a characteristic of the sample itself [28]. The SEM value could be calculated from a test-retest study or in a group of stable patients. The SEM in this study was calculated as:

$$S_w = SEM = \sqrt{\frac{1}{2n}\Sigma d_t^2}$$

where s_w is the within-subject standard deviation, d is the difference between two observations in patients *i* who reported "unchanged" on a four-point scale between 3 and 6 months follow up and n is the number of subjects [29]. The s_w statistics is also called the $SEM_{consistency}$ [30].

The lowest change that exceeds measurement error and noise at a 95% confidence level is defined as:

$$MDC_{95} = 1.96 * \sqrt{2} * SEM = 2.77 * SEM$$

Here, the $* \sqrt{2}$ is introduced because there are two measurements for each patient. The minimum detectable change (MDC) at a 95% confidence level, is denoted MDC_{95} [31]. With a scale value $\geq MDC_{95}$, we can be 95% certain that a change in the measured underlying construct has really occurred [32].

To assess the agreement between EQ5D and SF6D, a Bland Altman plot was constructed. [12]. The average EQ5D and SF6D change score values were plotted against the mean difference in change score values of both instruments. Limits of Agreement (LoA) based on a +/− $1.96*SD_{difference}$ interval for the differences were also constructed.

Structural validity

Structural validity concerns the degree to which the scores of an instrument are an adequate reflection of the dimensionality of the construct to be measured [33]. Both EQ5D and SF6D are constructed to measure the dimension of general health related quality of life (HRQoL) alongside a continuous scale (from low to high). Using Item Response Theory (IRT), the unidimensionality of the two utility indexes was tested. The category ordering of the questionnaire items (the probability of moving from an easier to a harder accomplished category of item answers in parallel with being increasingly disabled) was also tested.

We employed the unrestricted (Partial-Credit) polytomous model of the Rasch model (for general information about fit to the Rasch model, see Additional file 1) and the test proposed by Smith to reveal unidimensionality [34]. The SF6D and EQ5D were tested for unidimensionality in a principal component analysis (PCA) [35]. We performed a test equating procedure with baseline values from the SF6D and the EQ5D. The response of each patient to a question was tested against what was predicted by the Rasch model. Deviation from the model is expressed in residuals. Independent t-tests were used to test if the magnitude of the residuals represents a significant deviation. The CI calculated for this was 95%. We carried out a binominal test for the proportion of t-tests outside the range of −1.96–1.96. The software used in the Rasch analysis was RUMM 2020 (RUMM Laboratory Pty Ltd.).

Criterion validity

Criterion validity concerns the degree to which the scores of an instrument are an adequate reflection of a "gold standard" when this is present [33]. In this analysis we compared the scores of the EQ5D and SF6D to the disease specific instrument ODI. The rationale was that the ODI has been found to be a responsive and valid measure for patients with LBP [16,18,36] and that an improvement assessed by the ODI should be correlated with an improvement assessed by the two utility indexes.

Spearman rank correlation coefficient (r) with 1000 bootstrap replications of the *baseline* scores was calculated to assess the correlation between the scores of the EQ5D and ODI and SF6D and ODI.

Responsiveness

Responsiveness is defined as the ability of an instrument to detect change over time in the construct to be measured [33]. Responsiveness was assessed by using the

ODI and the seven-point global scores at 2-year follow-up as "gold standard". First, we calculated the Spearman rank correlation coefficient (r) with 1000 bootstrap replications for the correlation between *change* scores from baseline to 2 year FU for the EQ5D, SF6D and ODI. Second, we analyzed the area under the Receiver Operator Curve (ROC) for the change scores of the EQ5D, SF6D and ODI by using a dichotomization of the patient global scores as follows: Categories 1 to 3 was considered "improved" and categories 4 to 7 were "non-improved". Sensitivity was defined as the proportion of patients who were correctly classified as "improved" and specificity was defined as the proportion of patients who were correctly classified as "non-improved". A receiver operating characteristic (ROC) curve was then calculated by plotting every possible change score from baseline to 2 year FU for EQ5D, SF6D and ODI using the global score as an anchor [37,38]. The area under the ROC curve (AUC) was then calculated. This value corresponds to the possibility of correctly diagnosing a patient as having improved when this is really the case [38] and reflects how responsive the instruments are to detect a change in the underlying construct.

The calculation of ROC curves was performed with MedCalc Statistica software (version 11.1.1. for Windows, Brussels, Belgia).

Interpretability

Interpretability concerns the qualitative meaning of quantitative scores or change in scores. A core question is: "What is the smallest change in score in the construct to be measured which patients consider important? This is expressed as the Minimal Important Change (MIC) value [33], and is calculated based on the sensitivity and specificity results from the ROC analysis described above. The cut-off value for differentiating between patients with or without improvement at optimum sensitivity and specificity was determined using ROC analysis [38]. This corresponds to the upper left point on the ROC curve and it can be interpreted as the point or value that yields the lowest overall misclassification [25,39].

Study approval

The study was evaluated and approved by the regional Committee for Medical Research Ethics in east Norway. Storage of data was allowed by the Norwegian Data Inspectorate. The study was conducted in accordance with the Helsinki Declaration and the ICH-GCP guidelines and registered at clinicaltrial.gov under the identifier NCT00394732.

Results

At inclusion, there were 52,6% females. Mean age was 41 years and mean (SD) duration of low back pain was 6 years (5,74). Response rates at baseline and 2-year follow up and pre- and post-treatment scores are presented in Table 1. At baseline, 133 out of 173 patients had completely filled out the ODI, the EQ5D, and the SF-36, so values for each of the instruments could be calculated. At 2-year follow up, 113 patients had values for all three instruments, so change scores could be calculated.

Table 1 Response rate at baseline and two year follow up together with pre- and post-treatment scale scores

	Response rate		Mean scale score (SD)	
	Baseline	2 years	Baseline	2 year
ODI	99%	100%	42,29 (0,81)	23 (16)
SF6D	82%	90%	0.555 (0,007)	0.692 (0.143)
EQ5D	93%	99%	0.292 (0.026)	0.642 (0.318)

N = 133.

Measurement error

The SEM values calculated for patients who were stable for a period of 3 months are presented in Table 2.

The smallest change score that could be said to represent a real change beyond measurement error with 95% probability in one individual (MDC_{95}) are presented in Table 2.

The proportion of patients with a change score value ≥MCD_{95} was 69% for ODI, 57% for SF6D, and 45% for EQ5D.

Figure 1 shows a Bland-Altman plot of the SF6D and EQ5D baseline values. It illustrates a systematic variation (proportional error) in the EQ5D and SF6D scores, with less healthy individuals tending to have a higher score on the SF6D and healthier individuals tending to have a higher score on the EQ5D. The 95% Limits of Agreement (LOA) varied from −0.3 to 0.83 with a mean difference in scale scores (SD) of 0.23 (0.29).

Structural validity

When the SF6D items were used as one subset and the EQ5D items as another, the binominal test showed overlap of the 5% expected value with the 95% CI for each of the indexes. When the EQ5D and SF6D items were combined on a common scale, no overlap was identified. This finding could indicate that the indexes are unidimensional

Table 2 SEM and MDC_{95} values

	SEM	MDC_{95}
ODI	4.24	11.75
SF6D	0.056	0.157
EQ5D	0.155	0.429

The SEM represents the standard error of measurement. The MDC_{95} is the minimal detectable change value that falls outside the measurement error of the instrument with 95% probability.

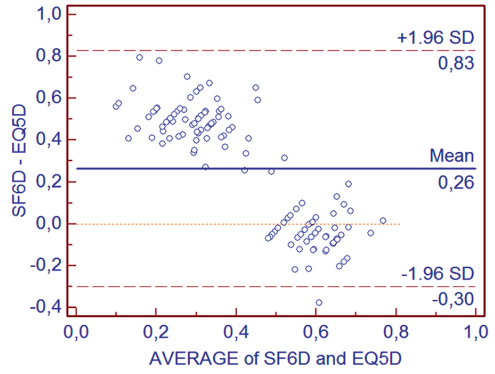

Figure 1 Bland-Altman plot.

when considered separately, but that they do not exactly measure the same underlying construct [34,40].

Figures 2 and 3 are graphic representations of the targeting of the SF6D and EQ5D items. Patients "ability" (level of health-related quality) and the item location (moving from an easy to a more difficult category of item answers in parallel with being increasingly disabled) are plotted on the same logarithmic scale. The bars in the top panels represent patient responses, and the bars in the bottom panels represent item thresholds on the scales. A threshold is the 0.5 probability point between adjacent item categories [41]. HRQoL levels (i.e., scoring values) decrease from left to right. Scoring responses outside the range of items represent a floor effect (to the right) or a ceiling effect (to the left). Responses outside the range of the scale give no additional information, and the test cannot discriminate between patients who fall in this area.

From Figures 2 and 3 it can be seen that the EQ5D was relatively well targeted for this group, with no sign of floor or ceiling effects, i.e., all responders were captured within the scale. With a mean person-location value of −0.132, the patients were at a slightly higher level of HRQoL than the scale could express. No floor or ceiling effect could be seen in the SF6D, but here the mean person-location was 1.423. This indicates that there is a tendency for patients to score at the lower end of the scale of this index.

Three of the items in the SF6D showed disordered threshold: question 1: Physical functioning, question 2: Role limitation and question 4: Pain. A better fit to the model was achieved if some of the response categories of these items were omitted. None of the questions in the EQ5D showed disordered thresholds.

Criterion validity

The correlation between *baseline* scores of ODI and EQ5D was r = 0,58 (n = 114, p=0.000) and for ODI and SF6D: r = 0.38 (n = 114, p = 0.000).

Responsiveness

a) The correlation between *change* scores of ODI and EQ5D was r = 0,64 (n = 108, p=0.000) and between ODI and SF6D change scores: r = 0.77 (n = 108, p = 0.000).
b) Spearman's rho for the correlation between change scores of the instruments and global score categories was 0.84, 0.55 and 0.76 for ODI, EQ5D and SF6D respectively. The area under the ROC curve, the possibility of correctly discriminating between "improved" or "non-improved" patients with a 95% CI was: 94% (87.5–97.6) for ODI, 90% (82.1–94.6) for SF6D, and 83% (75–90) for EQ5D. The ROC curves are presented in Figure 4.

Interpretability

The MIC values defined as the most optimal cut-off point of change scores plotted on the ROC curve was for ODI: 12.88,(sensitivity 88%, specificity 85%), EQ5D:

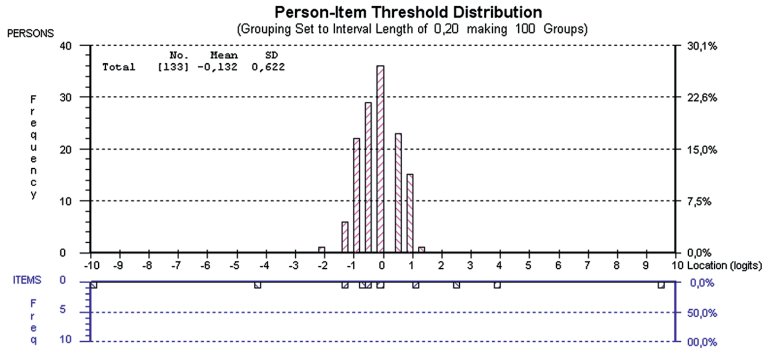

Figure 2 Person-item threshold distribution for EQ5D.

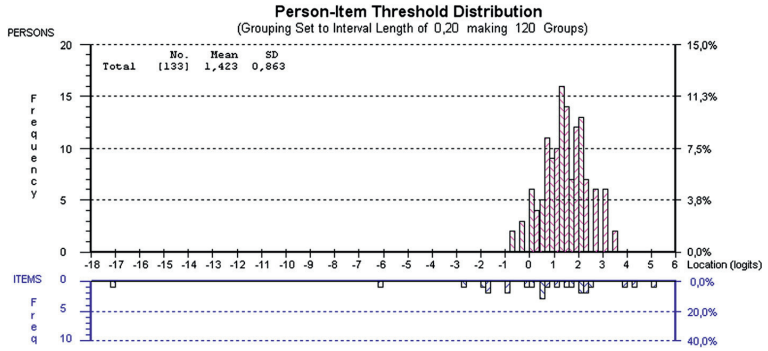

Figure 3 Person-item threshold distribution for SF6D.

0.173 (sensitivity: 73%, specificity 79%) and SF6D: 0,031 (sensitivity 93%, specificity 78%) (Figure 4).

Discussion

The present study failed to show similarity between EQ5D and SF6D in several important measurement properties. EQ5D had a higher value of inherent measurement error than SF6D. The mean difference between baseline score values had a wide 95% Limits of Agreement in the Bland-Altman plot signifying a low degree of agreement between the instruments [12,42]. Rasch analysis showed that although EQ5D and SF6D separately seem to have unidimensional scale properties they probably do not measure the same underlying construct. SF6D show less similarity with the baseline scores of the disease specific instrument but were more responsive to detect a change in the underlying construct in addition to better ability to correctly diagnosing a patient as having improved when this was really the case even though it did not reach the level of the ODI. The MIC values were quite different and SF6D had a better ability to identify truly change in scale score beyond measurement error.

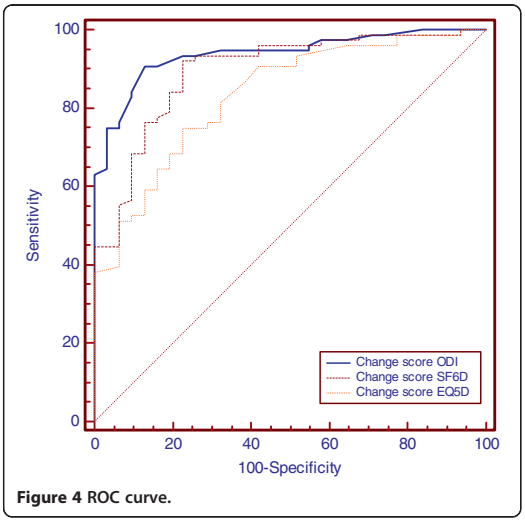

Figure 4 ROC curve.

Van Stel et al. showed that the EQ5D and the SF6D yield dissimilar scores in patients with coronary heart disease, and consequently, they cannot be used interchangeably [43]. This is in line with the Bland-Altman plot pattern we found in our study and in agreement with other previously published reports [6,13,43]. Furthermore, we observed that the magnitude of difference between the two instruments in the Bland-Altman plot was beyond the MIC for both instruments and therefore interpreted as clinically significant.

In this study, sensitivity was defined as the proportion of patients that truly improved (true-positive rate), and the sensitivity was the proportion of patients that did not actually improve (true-negative rate). The EQ5D diagnosed fewer patients as clinically improved (change score values beyond MIC). This was also reflected in the MIC/MDC_{95} ratio (the proportion of patients who truly changed with a possibility of 95% predicted by the instruments): For the MIC value to reach the MDC_{95}, the specificity for the SF6D would have to increase from 78.1 to 87.5, but the sensitivity would then fall from 92.5 to 73.7. For the EQ5D, this would necessitate an increase in specificity from 78.9 to 86.8 and a decrease in sensitivity from 72.8 to 57.6. In other words, to reach a value beyond the 95% CI for measurement error, the probability of correctly classifying a patient as improved would fall dramatically for the EQ5D, nearly reaching 50% or classifying by chance. The effect was not as dramatic for the SF6D, which would still correctly classify over 70% of patients as “improved”.

We found that the difference in the range of the scales between the SF6D and the EQ5D could be reflected in

their targeting properties. Based on the Rasch analysis (Figures 3 and 4), we could hypothesize that patients were at a lower level of HRQoL than the SF6D could express (floor effect). The range of patient abilities was better captured within the EQ5D scale. Barton et al. [6] compared the performance of the EQ5D and the SF6D in 1865 individuals over ≥45 years old. They found that healthier individuals had higher scores on the EQ5D, and less healthy individuals such as patients with back pain had higher scores on the SF6D. In a study that compared the SF6D and the EQ5D in liver transplant patients, Longworth et al. observed that the SF6D does not describe health states at the lower end of the utility scale but is more sensitive than the EQ5D in detecting small changes at the top of the scale [14]. This result is somewhat confusing because the same group later published a paper in which they conclude that the SF6D can describe some "poor health states including states that (according to the EQ5D scoring algorithm) are viewed as worse than the state of being dead" [13].

The Rasch analysis also revealed that some of the SF6D items did not function as intended. A better fit to the model was achieved if some of the response categories of these items were collapsed (i.e., the category was removed from the item). An interpretation of this is that for these items, patients could not differentiate between two adjacent response categories and the information in the removed categories was therefore redundant. None of the items in the EQ5D showed similar signs of dysfunction. When treated as separate scales, both instruments showed signs of unidimensionality, but significant invariance across items was noted when analyzed as one scale (all items from the SF6D and the EQ5D put together). The interpretation of this was that the two scales seem to measure different aspects of HRQoL. Walters and Brazier mentioned that a fundamental assumption in their comparison of the EQ5D and the SF6D was that the instruments should measure the same underlying HRQOL variable [44].

Strengths and limitations of the study

Compared to Brazier et al. [7], SF6D in our study had a higher percentage of missing data at both assessment time points (baseline and 2-year follow up). As Brazier mentioned in another paper, this has important consequences for data quality [45].

The use of global assessment score has been questioned in several studies [46,47]. Criticism of the reliability of anchor based methods includes no standardization of anchors, time dependence of patients perception of health, dependence on only one question and failure of the anchor question to differentiate between quantitative and qualitative perception of change [48]. The COSMIN study did not reach any consensus about which method to use to determine the MIC value but conclude that there is an ongoing discussion about this in the literature [23]. Some authors now suggest ROC analysis for determining MIC values mainly because it uses all available data and maximizes the number of individuals correctly classified [49]. The question and answer categories in our 7-point global scale was not a standardized scale but Spearman`s rho for the correlation between change scores of the instruments and global score categories used in the ROC analysis was considered acceptable (0.84, 0.55 and 0.76 for ODI, EQ5D and SF6D respectively) [46,50,51].

Conclusions

EQ5D and SF6D measure different aspects of HRQoL. The difference in psychometric properties between them and the lack of agreement is probably clinically significant. Because the ability to detect a change in the underlying construct and similarity to a disease-specific instrument is quite different, the choice of instrument should probably be guided by diagnosis and/or treatment choice. In our study of patients with chronic low back pain, the SF6D had the best ability to detect change and correctly identify patients as improved or non-improved beyond a 95% confidence level of measurement error.

Finally, our study supports the findings of Soegaard et al. [11]. They concluded that the SF6D and EQ5D cannot be used interchangeably for measurement of preference value and that sensitivity analysis examining the impact of between-measure discrepancy remains a necessary condition for cost-utility evaluation results.

Additional file

Additional file 1: Rasch analysis.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

LGJ takes responsibility for the integrity of the data and the accuracy of the data analysis. LGJ performed the statistical analysis. LGJ, MG, IR and CH participated in the design of the study. LGJ and CH: Acquisition of data. LGJ, CH, ØPN, KS, JIB, IR, GL and MG conceived of the study and helped to draft the manuscript. All authors had full access to the data. All authors read and approved the final manuscript.

Authors' information

LGJ: M.D. orthopaedic surgeon, CH: M.D. orthopaedic surgeon., KS: Ph.D. physiotherapist, ØPN: M.D, Ph.D. neurosurgeon, professor, JIB: M.D., Ph.D. specialist in physical medicine and rehabilitation, Ivar Rossvoll: M.D., Ph.D. orthopaedic surgeon, GL: M.D., Ph.D. specialist in physical medicine and rehabilitation, professor.

Acknowledgements

We want to thank the patients participating in the study, the South Eastern Norway Regional Health Authority and EXTRA funds from the Norwegian Foundation for Health and Rehabilitation, through the Norwegian Back Pain Association, for financial support and Hege Andresen at St.Olavs Hospital, Trondheim, for data coordination. Editorial assistance was delivered by Charlesworth Publishing Services at a price of 360$.

Financial disclosures
All authors involved declare that they have no conflict of interests and no financial disclosures to report.

Author details
[1]Neuroclinic; National Center of spinal disorder, St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway. [2]Clinic of Orthopedics and Rheumatology, Department of Orthopaedic Surgery, St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway. [3]Department Of Neuromedicine, Faculty of Medicine, Norwegian University of Science and Technology, Trondheim, Norway. [4]Clinic for Surgery and Neurology, Department of Orthopedics, Oslo University Hospital and University of Oslo, Oslo, Norway. [5]Department of Clinical Medicine, Neuromuscular Diseases and Research Group, University of Tromsø, Tromsø, Norway. [6]Clinic for Surgery and Neurology, Oslo University, Oslo, Norway. [7]Department of neurosurgery, St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway. [8]FORMI, Clinic for surgery and neurology, Ullevaal, Oslo N-0407, Norway. [9]Faculty of health Sciences, Department of Physiotherapy, Oslo and Akershus University College of Applied Sciences, Oslo, Norway.

Received: 14 May 2012 Accepted: 19 April 2013
Published: 26 April 2013

References

1. Brazier J: *Measuring and valuing health benefits for economic evaluation.* New York: Oxford University Press; 2007.
2. Nord E: **Health state values from multiattribute utility instruments need correction.** *Ann Med* 2001, **33**(5):371–374.
3. Brazier J, Roberts J, Deverill M: **The estimation of a preference-based measure of health from the SF-36.** *J Health Econ* 2002, **21**(2):271–292.
4. Dolan P: **Modeling valuations for EuroQol health states.** *Med Care* 1997, **35**(11):1095–1108.
5. Drummond MF: ***Methods for the economic evaluation of health care programmes**, 3rd edn. Oxford.* New York: Oxford University Press; 2005.
6. Barton GR, Sach TH, Avery AJ, Jenkinson C, Doherty M, Whynes DK, Muir KR: **A comparison of the performance of the EQ-5D and SF-6D for individuals aged >or= 45 years.** *Health Econ* 2008, **17**(7):815–832.
7. Brazier J, Roberts J, Tsuchiya A, Busschbach J: **A comparison of the EQ-5D and SF-6D across seven patient groups.** *Health Econ* 2004, **13**(9):873–884.
8. Grieve R, Grishchenko M, Cairns J: **SF-6D versus EQ-5D: reasons for differences in utility scores and impact on reported cost-utility.** *Eur J Health Econ* 2009, **10**(1):15–23.
9. Sach TH, Barton GR, Jenkinson C, Doherty M, Avery AJ, Muir KR: **Comparing cost-utility estimates: does the choice of EQ-5D or SF-6D matter?** *Med Care* 2009, **47**(8):889–894.
10. Soegaard R: **Interchangeability of the EQ-5D and the SF-6D in Long-Lasting Low Back Pain Source: Value in Health 12, no. 4 (2009): 606–612 Additional Info: Blackwell Publishing; 20090601 Standard No: ISSN: 1098–3015.** *Value Health* 2009, **12**(4):606–612. doi:10.1111/j.1524-4733.2008.00466.x.
11. Sogaard R, Christensen FB, Videbaek TS, Bunger C, Christiansen T: **Interchangeability of the EQ-5D and the SF-6D in long-lasting low back pain.** *Value Health* 2009, **12**(4):606–612.
12. Bland JM, Altman DG: **Statistical methods for assessing agreement between two methods of clinical measurement.** *Lancet* 1986, **1**(8476):307–310.
13. Bryan S, Longworth L: **Measuring health-related utility: why the disparity between EQ-5D and SF-6D?** *Eur J Health Econ* 2005, **6**(3):253–260.
14. Longworth L, Bryan S: **An empirical comparison of EQ-5D and SF-6D in liver transplant patients.** *Health Econ* 2003, **12**(12):1061–1067.
15. Hellum C, Johnsen LG, Storheim K, Nygaard OP, Brox JI, Rossvoll I, Ro M, Sandvik L, Grundnes O: **Surgery with disc prosthesis versus rehabilitation in patients with low back pain and degenerative disc: two year follow-up of randomised study.** *BMJ* 2011, **342**:d2786.
16. Fairbank JC, Couper J, Davies JB, O'Brien JP: **The Oswestry low back pain disability questionnaire.** *Physiotherapy* 1980, **66**(8):271–273.
17. Fairbank JC, Pynsent PB: **The Oswestry Disability Index.** *Spine* 2000, **25**(22):2940–2952. discussion 2952.
18. Grotle M, Brox JI, Vollestad NK: **Cross-cultural adaptation of the Norwegian versions of the Roland-Morris Disability Questionnaire and the Oswestry Disability Index.** *J Rehabil Med* 2003, **35**(5):241–247.
19. Ware JE Jr, Sherbourne CD: **The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection.** *Med Care* 1992, **30**(6):473–483.
20. Dolan P, Gudex C, Kind P, Williams A: **The time trade-off method: results from a general population study.** *Health Econ* 1996, **5**(2):141–154.
21. The EuroQol Group: **EuroQol–a new facility for the measurement of health-related quality of life.** *Health Policy* 1990, **16**(3):199–208.
22. Ostelo RW, de Vet HC: **Clinically important outcomes in low back pain.** *Best Pract Res Clin Rheumatol* 2005, **19**(4):593–607.
23. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM, De Vet HC: *COSMIN checklist manual.* 2012.
24. Mokkink LB, Terwee CB, Knol DL, Stratford PW, Alonso J, Patrick DL, Bouter LM, de Vet HC: **The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content.** *BMC Med Res Methodol* 2010, **10:**22.
25. 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):578–582.
26. Wyrwich KW, Nienaber NA, Tierney WM, Wolinsky FD: **Linking clinical relevance and statistical significance in evaluating intra-individual changes in health-related quality of life.** *Med Care* 1999, **37**(5):469–478.
27. 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.** *J Clin Epidemiol* 1999, **52**(9):861–873.
28. Crosby RD, Kolotkin RL, Williams GR: **Defining clinically meaningful change in health-related quality of life.** *J Clin Epidemiol* 2003, **56**(5):395–407.
29. Bland JM, Altman DG: **Measurement error.** *BMJ* 1996, **313**(7059):744.
30. de Vet HC, Terwee CB, Knol DL, Bouter LM: **When to use agreement versus reliability measures.** *J Clin Epidemiol* 2006, **59**(10):1033–1039.
31. Beaton DE: **Understanding the relevance of measured change through studies of responsiveness.** *Spine* 2000, **25**(24):3192–3199.
32. Hagg O, Fritzell P, Nordwall A: **The clinical importance of changes in outcome scores after treatment for chronic low back pain.** *Eur Spine J* 2003, **12**(1):12–20.
33. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM, de Vet HC: **The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes.** *J Clin Epidemiol* 2010, **63**(7):737–745.
34. Smith EV Jr: **Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals.** *J Appl Meas* 2002, **3**(2):205–231.
35. Chou Y-T, Wang W-C: **Checking Dimensionality in Item Response Models With Principal Component Analysis on Standardized Residuals.** *Educ Psychol Meas* 2010, **70**(5):717–731.
36. Fairbank JC, Pynsent PB, 22: **The Oswestry Disability Index.** *Spine (Phila Pa 1976)* 2000, **25:**2940–2952. discussion 2952.
37. Zweig MH, Campbell G: **Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine.** *Clin Chem* 1993, **39**(4):561–577.
38. Deyo RA, Centor RM: **Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance.** *J Chronic Dis* 1986, **39**(11):897–906.
39. 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.** *Spine J* 2008, **8**(6):968–974.
40. Tennant A, Conaghan PG: **The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper?** *Arthritis Rheum* 2007, **57**(8):1358–1362.
41. Pallant JF, Tennant A: **An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS).** *Br J Clin Psychol* 2007, **46**(Pt 1):1–18.
42. Bland JM, Altman DG: **Measuring agreement in method comparison studies.** *Stat Methods Med Res* 1999, **8**(2):135–160.
43. van Stel HF, Buskens E: **Comparison of the SF-6D and the EQ-5D in patients with coronary heart disease.** *Health Qual Life Outcomes* 2006, **4:**20.

44. Walters SJ, Brazier JE: **Comparison of the minimally important difference for two health state utility measures: EQ-5D and SF-6D.** *Qual Life Res* 2005, **14**(6):1523–1532.
45. Brazier J, Deverill M: **A checklist for judging preference-based measures of health related quality of life: learning from psychometrics.** *Health Econ* 1999, **8**(1):41–51.
46. de Vet HC, Ostelo RW, Terwee CB, van der Roer N, Knol DL, Beckerman H, Boers M, Bouter LM: **Minimally important change determined by a visual method integrating an anchor-based and a distribution-based approach.** *Qual Life Res* 2007, **16**(1):131–142.
47. Guyatt GH, Norman GR, Juniper EF, Griffith LE: **A critical look at transition ratings.** *J Clin Epidemiol* 2002, **55**(9):900–908.
48. Terwee CB, Roorda LD, Dekker J, Bierma-Zeinstra SM, Peat G, Jordan KP, Croft P, de Vet HC: **Mind the MIC: large variation among populations and methods.** *J Clin Epidemiol* 2010, **63**(5):524–534.
49. Turner D, Schunemann HJ, Griffith LE, Beaton DE, Griffiths AM, Critch JN, Guyatt GH: **The minimal detectable change cannot reliably replace the minimal important difference.** *J Clin Epidemiol* 2010, **63**(1):28–36.
50. Cella D, Hahn EA, Dineen K: **Meaningful change in cancer-specific quality of life scores: differences between improvement and worsening.** *Qual Life Res* 2002, **11**(3):207–221.
51. Guyatt GH, Jaeschke RJ: **Reassessing quality-of-life instruments in the evaluation of new drugs.** *Pharmaco Economics* 1997, **12**(6):621–626.

doi:10.1186/1471-2474-14-148
Cite this article as: Johnsen *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 Musculoskeletal Disorders* 2013 **14**:148.

Appendix

APPENDIX

Rasch analysis

Rasch analysis is a probabilistic model that tests the extent to which the observed pattern of responses fits the pattern expected by the model [377 378 379]. The model shows what should be expected in responses to items if measurement (at the metric level) is to be achieved [240]. Two qualities are central: the ability of a person and the difficulty of an item. The ability can be any clinical sign, such as low back pain. The difficulty of an item could be seen as a measure of the extent to which a person has the ability (e.g., more or less low back pain). The model states that the probability that a person will affirm an item is a logistic function of the difference between a person's level of, for example, reduced physical function due to low back pain (θ) and the level of reduced functional level expressed by the item (b) and only a function of that difference [240]

$$\ln\left(\frac{P_{nij}}{1-P_{nij-1}}\right)=\theta_n-b_{ij}$$

P_{nij} is the probability that a person *n* will answer in the "affirm" category *j* of item *i* (or be able to do the level of a task specified by that category within the item). Rasch analysis also offers the possibility of converting ordinal raw data into a linear scale if the data fit the model. We used a polytomous variant of the Rasch model, which is known as the partial credit model [380].

In PCA, we explore the relationship of the items to the components that contribute most to the variation in data after the Rasch component is removed[293]. This is done by comparing fit residuals for each person for each item using independent t-tests [292]. The first component in the PCA is the component that accounts for the most variance in the data and can be seen as a "second dimension." To examine this, we used the subsets of items that loaded the most strongly on the first component because these were the most likely to breach the assumption of unidimensionality. In other words, if these two subsets showed a significant difference from the overall scale, then the assumptions of unidimensionality could be broken. If the data fit the model, then analysis of any subsets of items should produce equivalent person measures within measurement error.

Threshold order

A threshold is defined as the 0.5 probability point between adjacent categories of an item[240]. The probability of affirming one category response is illustrated by probability density curves for

each of the categories (fig 6). The 0.5 probability is then at the top of the curve. For categories in increasingly or decreasingly order the top of curve 1 should come before curve 2, the top of curve 2 should come before curve 3 etc. When this is not the case, the thresholds are disordered (fig 7). For SF-6D, disordered thresholds were found between category 3 and 4 in “Physical”, between category 2 and 3 in “Role” and between category 1 and 2 in “Pain”.

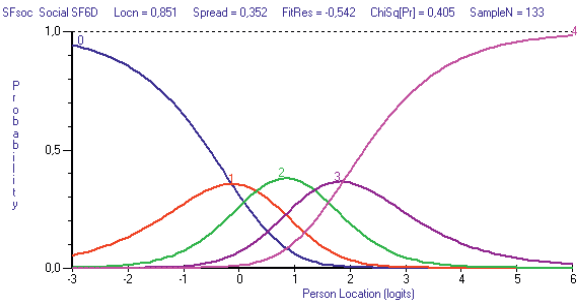

Figure6. Example of item with ordered thresholds

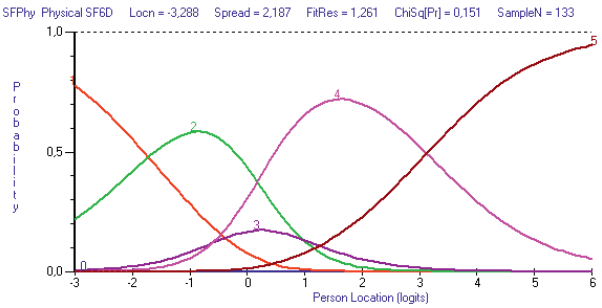

Figure7. Example of item with disordered thresholds