MRI findings in candidates for lumbar disc prosthesis: reliability and relationship to disability and pain

Linda Berg

Dissertation for the degree philosophiae doctor (PhD)
at the University of Bergen

2014

Dissertation date: November 21, 2014

Year: 2014

Title: MRI findings in candidates for lumbar disc prosthesis: reliability and relationship to disability and pain

Author: Linda Berg

Print: AIT OSLO AS / University of Bergen

Scientific environment

UNIVERSITAS BERGENSIS

Department of Clinical Medicine, Faculty of Medicine and Dentistry,
University of Bergen

HELSE BERGEN
Haukeland universitetssjukehus

Department of Radiology, Haukeland University Hospital

Acknowledgements

The present study was carried out at Section for Radiology, Department of Clinical Medicine (K1), Faculty of Medicine and Dentistry, University of Bergen and Department of Radiology, Haukeland University Hospital.

I want to thank the patients participating in the study, the co-authors, colleagues and supervisors for good cooperation and support, and **Section for Radiology, Department of Clinical Medicine (K1), University of Bergen** and **Department of radiology, Haukeland University Hospital** for support and goodwill.

In particular I want to thank my main supervisor **Ansgar Espeland** for sharing his comprehensive knowledge and his great interest in research, and for his attention to details and commitment to good scientific quality. His guiding, support and commitment has been immensely important.

A special thanks to my second supervisor **Geir Egil Eide** for sharing his great knowledge and methodological insight, especially for his important contributions to the statistical part of the project.

Thanks to **Øivind Gjertsen** and **Gesche Neckelmann** for their important and extensive participation in the rating of MRI images.

Thanks to **Christian Hellum** for sharing his knowledge and interest in back pain research as well as contributions and feedback on the papers before publishing.

Thanks also to **Lars Gunnar Johnsen**, **Kjersti Storheim**, and **Jens Ivar Brox** for both important and interesting input, and good cooperation.

The following additional investigators participated in the study:

From the University Hospital North Norway, Tromsø (included *n*=8 patients): Department of Orthopaedic Surgery: Odd-Inge Solem, MD; Department of Neurosurgery: Jens Munch-Ellingsen, MD, PhD; and Department of Physical Medicine and Rehabilitation: Franz Hintringer, MD, Anita Dimmen Johansen, Ergonomist, Guro Kjos, PT. **From Trondheim University Hospital, Trondheim** (included *n*=21 patients): National Centre for Spinal Disorders, Department of Neurosurgery: Hege Andresen, RN, Øystein P Nygaard, MD, Professor, Ivar Rossvoll, MD, PhD Helge Rønningen, MD, Professor, Kjell Arne Kvistad, MD, PhD; and Multidiscipline Spinal Unit, Dep. of Physical Medicine and Rehabilitation: Bjørn Skogstad, MD, Janne Birgitte Børke, PT, MSc, Erik Nordtvedt, PT, Gunnar Leivseth, MD, Professor, Magne Rø, MD. **From Haukeland University Hospital, Bergen** (included *n*=64 patients): Kysthospitalet in Hagevik, Department of Orthopedic Surgery: Sjur Braaten, MD, Turid Rognsvåg, PT, MSc, Gunn Odil Hirth Moberg, Secretary; and The Outpatient Spine Clinic, Department of Physical Medicine and Rehabilitation: Jan Sture Skouen, MD, Professor, Lars Geir Larsen, PT, Vibeche Iversen, RN, Ellen H Haldorsen, PhD, cand Psychol, Elin Karin Johnsen, RN, Kristin Hannestad, PT. **From Stavanger University Hospital, Stavanger** (included *n*=27 patients): Department of Orthopaedic Surgery: Endre Refsdal, MD. **From Oslo University Hospital, Oslo (included *n*=53 patients)**: Department of Orthopaedics: Vegard Slettemoen, RN, Kenneth Nilsen, RN, Kjersti Sunde, RN, Helenè E Skaara, PT, MSc; and Department of Physical Medicine and Rehabilitation: Berit Johannessen, PT, Anna Maria Eriksdotter, PT, MSc, Anne Keller, MD, PhD.

Thanks to the **Western Norway Regional Health Authority** and to **Haakon and Sigrun Ødegaard's Fund** at the Norwegian Society of Radiology for financial support. This work was part of a larger project also supported by the **South Eastern Norway Regional Health Authority** and the **Norwegian ExtraFoundation for Health and Rehabilitation** through the Norwegian Back Pain Association.

Abbreviations

AP- AnteroPosterior

BMI- Body Mass Index

β - regression coefficient

CC- CranioCaudal

CI- Confidence Interval

cm- centimetre

CNR- Contrast to Noise Ratio

CNS- Central Nervous System

CT- Computer Tomography

DD- Disc Degeneration

DICOM- Digital Imaging and Communications in Medicine

DRIVE- DRIVen Equilibrium

FA- Facet Arthropathy

FLAIR- Fluid Attenuated Inversion Recovery

FOV- Field Of View

FS- Fat Suppressed

FSE- Fast Spin Echo

GEE- Generalized Estimating Equations

HIZ- High-Intensity Zone in the intervertebral disc

k- kappa magnitude

LBP- Low Back Pain

LOWESS- LOcally WEighted Scatterplot Smoothing

mm- millimetre

MRI- Magnetic Resonance Imaging

NA- Not Applicable

NEX- Number of EXcitations

NOK- Norwegian Kroner

ODI- Oswestry Disability Index

OR- Odds Ratio

PABAK- Prevalence-And Bias-Adjusted Kappa

PACS- Picture Archiving and Communication System

PD- Proton Density

r- correlation coefficient

SD- Standard Deviation

SE- conventional Spin Echo

SNR- Signal to Noise Ratio

STIR- Short Tau Inversion Recovery (also called Short T1 Inversion Recovery)

T- Tesla (indicating magnetic field strength of the MRI machine)

TE- Echo Time

TIRM- Turbo Inversion Recovery Magnitude

TR- Repetition Time

VAS- Visual Analogue Scale

Definitions

Bias: "Bias is the extent to which the raters disagree on the proportion of positive (or negative) cases and is reflected in a difference between cells b and c" (in a 2x2 cross-table) [104].

Bonferroni correction for multiple comparisons: When several independent significance tests are carried out simultaneously on the same body of data (multiple significance testing) probability is high of finding a significant result just by chance. Bonferroni correction is a method to keep the overall probability of type I error (the risk of rejecting the null hypothesis when there is no real difference) below a certain level (α, usually below 0.05). If we perform *k* comparisons the Bonferroni correction gives us the new significance level = α/*k.* The new significance level tends to be over-corrected [103, 109].

Bootstrapping: Empirical method of obtaining confidence interval (CI) for estimates (for example mean) when assumptions for using existing statistical methods are not satisfied, for example the common assumption of independent observations. The method implies taking a large number of repeated samples ("constructed samples") from a single data set using a computer. For example, CI for the mean will be calculated by finding the mean for each "constructed sample" and then the distribution of the "constructed sample" means [103].

Central pain sensitization: Distorted or amplified pain so that that the degree, duration, and spatial extent of pain "no longer directly reflects a peripheral noxious stimuli, but rather the particular functional states of circuits in the CNS" [26]. CNS means central nervous system.

Chronic low back pain: Low back pain (LBP) with more than 12 weeks duration [2].

Clustered data: Observations in one cluster tend to be more similar to each other than to the rest of the sample. Clustering on subject occur when multiple measures are made in each subject, and consequently the assumption that observations are independent is violated [173].

Diagnostic triage: The diagnostic process of sorting patients to determine priority of medical treatment based on symptoms and signs [3].

Disc degeneration: No universally accepted standard definition exists, but it is commonly used for the following imaging findings: nucleus pulposus signal loss, intervertebral disc height decrease, bulging or herniation of the disc, endplate irregularities, and vertebral osteophytes [24].

Interobserver agreement: Agreement between ratings made by two or more observers [104].

Intraobserver agreement: Agreement between ratings made by the same observer on two or more occasions [104].

Low back pain (LBP): Pain and discomfort, localised below the costal margin and above the inferior gluteal folds, either with or without referred leg pain [2].

Magnetic resonance imaging (MRI) indication for lumbar disc prosthesis in the present study: (a) ≥ 40 % disc height decrease and/or (b) at least two of the following: Modic changes type I and/or II, posterior high-intensity zone (HIZ) in the disc, and dark/black nucleus pulposus on T2-weighted images; evaluated separately at L4/L5 and at L5/S1 [42].

Nociception: Activation of sensory nerve signal sending information about potential tissue damage [26, 165].

Reliability: Reliability is the extent to which the observers agree in their ratings [104]. Reliability is dependent on both repeatability (when measurement is repeated under the same conditions) and reproducibility (when measurement is repeated under different conditions) [103].

Sensitization: Increased response to stimulation [37].

Spinal motion unit or segment: A spinal motion unit or segment, also called functional spinal unit, is defined as the smallest physiological motion unit of the lumbar spine; consisting of two adjacent vertebrae with the intervertebral disc, facet joints, ligaments, and muscles between them [3].

Contents

Appendix

Papers I-III

Summary

Background and objective: In candidates for lumbar disc prosthesis, limited reliability data exist for magnetic resonance imaging (MRI) findings relevant to planning of treatment and to evaluation of outcome after treatment. In this subgroup of patients with low back pain (LBP), we assessed the reliability of degenerative MRI findings and change in such findings over time. How the sum of MRI findings relates to the degree of disability and LBP was also evaluated since this is not clear.

Materials and Methods: 170 of 173 patients aged 25-55 years, with LBP ≥ 1 year, Oswestry Disability Index (ODI) score ≥ 30 %, and localized degenerative MRI changes at L4/L5 and/or L5/S1 were included. On pre-treatment images three experienced radiologists independently rated Modic changes, disc findings, and facet arthropathy (FA) at L3-S1. Two of the radiologists rated progress and regress of the same findings on 2-year follow up images (n = 126). Agreement was analysed using the kappa statistic. How pre-treatment MRI total score related to the ODI (n = 170) and LBP intensity scores (n = 165) was analysed using multiple linear regression adjusting for age, gender, body mass index (BMI), smoking, and anxiety/depression.

Results: Overall interobserver agreement was generally moderate or good (kappa 0.40-0.77) at L4-S1 for Modic changes, nucleus pulposus signal, disc height, posterior HIZ, and disc contour, and fair (kappa 0.24) at L4/L5 for FA. Intraobserver agreement was mostly good or very good (kappa 0.60-1.00). Image comparison indicated good interobserver agreement on progress and regress (prevalence and bias adjusted kappa (PABAK) 0.63–1.00) for Modic changes, posterior HIZ, disc height, and disc contour at L3-S1 and for nucleus pulposus signal and FA at L3/L4; and moderate interobserver agreement (PABAK 0.46–0.59) on decreasing nucleus signal and increasing FA at L4-S1. The MRI total score was not related to ODI (regression coefficient 0.12, *P* = 0.79) or LBP intensity (regression coefficient 0.64, *P* = 0.37).

Conclusions and consequences: In candidates for lumbar disc prosthesis or fusion, Modic and disc findings, and change in these findings over time, have acceptable reliability for use in treatment planning and research, but the sum of these findings are unlikely to explain variation in current pre-treatment disability and pain.

List of included papers

I. Berg L., G. Neckelmann, O. Gjertsen, C. Hellum, L.G. Johnsen, G.E. Eide, and A. Espeland, *Reliability of MRI findings in candidates for lumbar disc prosthesis.* Neuroradiology, 2012. **54**(7): p. 699-707.

II. Berg, L., O. Gjertsen, C. Hellum, G. Neckelmann, L.G. Johnsen, G.E. Eide, and A. Espeland, *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**(12): p. 1547-57.

III. Berg, L., C. Hellum, O. Gjertsen, G. Neckelmann, L.G. Johnsen, K. Storheim, J.I. Brox, G.E. Eide, and A. Espeland, *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. **42**(11): p. 1593-602.

Reprints were made with permission from Springer Science+Business Media.

1.0 Introduction

Low back pain (LBP) is the leading single cause of disability worldwide [1]. Better understanding of the aetiology, pathophysiology and prognosis of LBP and improved prevention, diagnosis and treatment are important to reduce disability due to LBP. The first part of this introduction contains an overview of LBP and chronic non-specific LBP in particular. The next part addresses imaging for LBP and degenerative findings on magnetic resonance imaging (MRI). The final part details the specific background and motivation for the present studies on patients with longstanding LBP and localized disc degeneration who were candidates for surgery with lumbar disc prosthesis. These studies concern agreement and disagreement on MRI findings and change in such findings over time (paper I and II), and association between MRI findings and the degree of disability and pain (paper III).

The literature search made for this thesis was completed May 28, 2014.

1.1 Low back pain

- Definition

Low back pain (LBP) is defined as pain and discomfort, localised below the costal margin and above the inferior gluteal folds, either with or without referred leg pain [2].

- Classification

Classification means to put things into groups. Classification of LBP is important because it enables us to organise information about LBP for example according to aetiology, treatment effects and prognosis. Classification of LBP is challenging, because the risk, aetiology, and prognosis of LBP involve multiple biological, psychological and social factors [3]. LBP has been classified based on diagnosis (assuming that a specific pathoanatomical condition induces specific symptoms and signs), treatment effect (assuming that specific treatments are best for specific groups of patients), and prognostic factors (assuming that specific patient groups will have a common prognosis); however, there is no consensus on which imaging findings, specific patient characteristics, and symptoms/signs to use for grouping of LBP patients [4]. An episode of LBP can be classified as acute (duration 0-6 weeks), sub-acute (duration 6-12 weeks), or chronic (duration more than 12 weeks) [3].

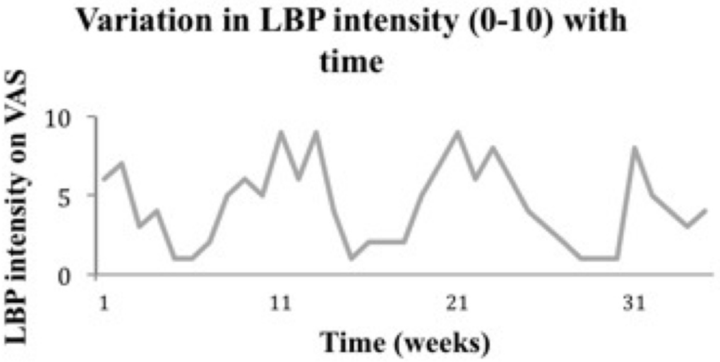


Figure 1 The figure shows the relapsing and variable course of low back pain (LBP) intensity over time in a hypothetical patient. VAS= visual analogue scale ranging from 0 (no pain) to 10 (worst pain imaginable).

Since LBP often fluctuates [2] (**Figure 1**) a distinction between recurrent LBP ("a new episode after a symptom-free period of 6 months") and an exacerbation of chronic LBP has been proposed [5]. LBP may still not fit these acute and chronic categories, and the number of days with LBP the past year has been proposed as a better measure [6]. Yet, the differentiation between acute and chronic LBP is useful both in clinical practice and in research. Risk factors, prognostic factors and advised treatment differ between acute and chronic LBP [6].

A further system for classifying LBP is based on symptoms and signs, and the importance of recognising conditions that may affect prognosis and/or treatment choice. In this system, LBP patients are divided into three groups (diagnostic triage). One group (1-5 % of patients) presents with "red flags" i.e. symptoms or signs indicating serious underlying spinal pathology such as tumour, infection, inflammatory disorder, fracture, or cauda equina syndrome. Examples of red flags are violent trauma, previous cancer, unexplained weight loss, and widespread neurological symptoms [5, 6]. Another group (5-10 % of patients) have symptoms or signs of radiculopathy (nerve root affection), such as radiating pain in foot/toes with dermatomal pattern and motor, sensory or reflex changes limited to one nerve root. The largest group (80-90 % of patients) does not have symptoms of serious spinal pathology or radiculopathy, and is classified as non-specific LBP [3, 6].

- Prevalence and course of LBP

LBP is common and disables 11-12 % of the population [2]. The adult incidence of LBP is probably about 5 % per year, the point prevalence 12-33 %, the one-year prevalence 22-65 %, and the lifetime prevalence 11-84 % [5, 7]. The prevalence in Norway was found to be similar (point prevalence 13.4 %, one-year prevalence 40.5 %, and lifetime prevalence 60.7 %) [8]. Probably less than 10 % of LBP episodes in the general population lead to consultation with a general practitioner [9, 10]. As much as 90 % of patients with acute LBP improve within 6 weeks [5] and 20 % recover completely after 1 year [11, 12], but 42-75 % of LBP patients still experience LBP after 1 year [5, 11, 12]. In a review on long-term course of LBP in the general population, Hestbaek et al reported that about 60 % (range 44-78 %) of persons with LBP experienced recurrence of LBP during 1 to 5 years follow up (inhomogeneous definition of LBP and methodological variations made meta-analysis of the reviewed studies impossible) [12]. Leboeuf-Yde et al recently published a one-year study of the general population where LBP was present in 65 % and could be classified as either "episodic" (in 30 %) or "mainly persistent" (in 35 %) [13]. In a population study of 40-year old men and women [14], 22 % reported 1-7 days total duration of LBP the past year, 23 % reported 8-30 days, and 25 % reported > 30 days. In the same study, 66 % reported more than one previous LBP episode, and 30 % reported non-trivial LBP episodes the past year. Non-trivial LBP episode was defined as LBP for more than 30 days during the past year with at least 1) care seeking or reduced activity *or* 2) previous LBP episode(s) with mean duration of more than 6 weeks *or* 3) self-reported disc herniation [14].

- Economic aspects and costs

In Norway LBP was the subgroup of musculoskeletal disorders that caused most sick leave and the largest National Health Insurance Office expenditures in 2010, and in 2009 13% of sick leaves lasting more than 8 weeks were due to back pain related conditions [6]. The total cost of back pain related conditions were estimated to 13-15 billion Norwegian kroner (NOK) in 2003 [6] and approximated to 15-24 billion NOK in 2011 [15]. From 2000 to 2012 sick leave and disability pensions caused by back pain related conditions decreased both in Norway and some other European countries [6]. This may be due to a change in diagnostic practice and not to a real reduction in back pain related conditions, because in the same period sick leave due to depressions and mild psychiatric disease increased accordingly, and the point prevalence of self-reported back pain increased slightly from 1995-1997 to 2006-2008 in a Norwegian population study (from 32.8 % in HUNT 1 to 35.7 % in HUNT 2) [6]. Van Tulder et al concluded that 93 % of LBP related costs in the Netherlands were due to indirect costs including sick leave and disability costs [16]. Martin et al reported a 65 % increase in back pain related expenditures from 1997 to 2005 in the United States, without evidence of better outcome for patients [17].

- Risk factors for LBP

Many potential risk factors for LBP have been studied, and although results have been conflicting, possible risk factors include heavy physical work load, frequent bending, twisting, lifting, static and repetitive work, vibrations, distress, stress, anxiety/depression, job dissatisfaction, and cognitive dysfunction and pain behaviour [5]. The point prevalence of LBP is about twice as high for those with *versus* those without a history of LBP (14-93 % vs. 7-39 %) [12]. LBP has been associated with degeneration of the lumbar disc in many studies, but in a systematic review Endean et al concluded that disc height reduction, reduced nucleus signal, disc protrusion, annular tear/ HIZ), endplate changes (irregular or defects) and FA could not be established as causes of LBP [18].

1.2 Chronic non-specific LBP

- Prevalence, course and costs of chronic non-specific LBP

According to the “European guidelines for the management of chronic nonspecific low back pain” (2006) little scientific evidence exists on the prevalence of chronic non-specific LBP, but the best estimates suggest a point prevalence of 23 % [2]. Frequency, duration and intensity of LBP vary over time for most patients with chronic non-specific LBP and both prevalence estimates and risk factors vary depending on the definition of chronic LBP used [2, 19]. Accordingly, chronic non-specific LBP includes patients with a broad spectrum of symptoms from a patient experiencing LBP of 12 weeks duration for the first time to a patient experiencing frequent recurring LBP for many years, and from severe disability to normal function despite of pain. These variations make it difficult to compare results from different studies. More than one third of patients with recent onset, non-radicular chronic LBP are reported to recover within 12 months (in one study 35 % recovered completely by nine months and 41 % by 12 months) [20].

Chronic LBP may have a serious social and economic impact both on the individual and on the community. About 9 % of disability pensions (2006) and 11 % of sick leaves (2008) in Norway were due to LBP [21].

- Comorbidity

Chronic LBP is often associated with pain in other musculoskeletal locations, most commonly with pain in the neck or pain due to osteoarthritis of the knees and hips [3]. In a community survey by Carnes et al the point prevalence of chronic LBP was 25 %, but only 3-4 % when patients with other musculoskeletal pain conditions were excluded [22]. Sleep disorders, stress-related symptoms, headache, and mental disorders are also quite common among patients with chronic LBP. In a study by Hagen et al LBP patients (sick listed for 8-12 weeks for LBP) had more neck pain, upper back pain, pain in the feet during exercise, headache, sleep problems, flushes/heat sensations, anxiety, and sadness/depression than a general Norwegian reference population (odds ratios = ORs ranged from 1.6 to 3.4) [23].

- Pathogenesis and risk factors for chronic non-specific LBP

The aetiology and pathophysiology of chronic non-specific LBP is - by definition - principally unknown [19]. Probably determined variables (such as genetics and age) and environmental variables (such as trauma, physical and psychological burden of life) interact in chronic LBP [19, 24, 25] (**Figure 2).**

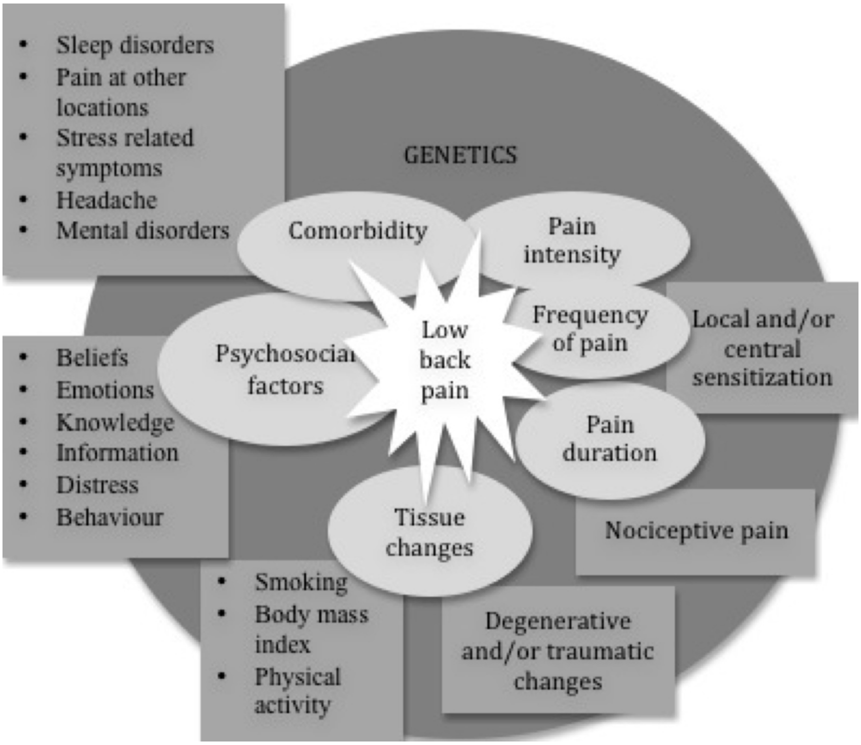


Figure 2 The figure illustrates that the symptom low back pain has a complex background involving biological, psychological and social factors constantly interacting with each other, and with genetics as a backdrop influencing all factors.

Genetics: Genetic variation may explain much of the variation in LBP between people. The heritability - the part of the variation in LBP between people explained by genetic variation - for severe disabling LBP was 57 % in a twin study by McGregor et al [25]. The prevalence of severe disabling LBP (ever had LBP of duration > 1 month with associated disability) was 18 % among the 1064 monozygotic and dizygotic female twin pairs, and more severe definitions of pain were associated with a higher heritability [25].

Factors affecting chronicity: Predictors of chronicity are psychosocial distress, depressive mood, severity and functional impact of pain, extreme symptom report, negative expectations and beliefs, prior LBP episodes, radicular pain, and heavier occupations with no modified duty/ lack of work place support [2]. Many LBP patients have comorbidities that may contribute to disability and pain experience [23]. Chronic pain is associated with structural changes in the brain and pathological or dysfunctional pain not related to damaging injury or inflammation, and accordingly chronic LBP may be influenced by such amplified sensory signals in the central nervous system (CNS) [26].

Tissue changes: Degeneration of the lumbar spine (disc degeneration and FA) has been a proposed cause of chronic LBP. However, degenerative MRI findings are prevalent in people without LBP, and associations between chronic LBP (≥ 3 months) and degenerative MRI findings are weak (OR usually 2.0-3.0 or less) [27]. This topic is detailed further later in the introduction.

Pain inducing mechanisms: In a review on lumbar degenerative disease, Modic and Ross discuss that both mechanical factors and inflammatory parameters may be involved in pain inducing mechanisms [28].

Mechanical factors: Restricted, excessive or irregular movements of one or more spinal motion units (see definitions) may be a result of degeneration of the facet joints (degenerative spondylolisthesis) and/or disc degeneration [28], but the association between abnormal movements and LBP is still unclear [29]. In a review Mulholland reports that movement is usually restricted in severe disc degeneration, whereas increased angular and translational movement is seen in some normal discs and in mild disc degeneration [29]. In the same review Mulholland discusses the importance of abnormal loading as a cause of LBP, and that abnormal movement or instability of a degenerated vertebral segment may be associated with but not a cause of LBP. This view of abnormal loading as a cause of LBP is supported by results from studies on different types of treatment (for example fusion, cage and disc prosthesis surgery), results for realignment surgery in other joints, studies on the mechanical load bearing capacity of the normal and degenerated disc, and LBP patients' history of what aggravates and alleviates pain (for example lying down, sitting, movement) [29].

Inflammatory parameters: Nucleus pulposus has inflammatory properties, nerve ingrowths in annulus fibrosus have been shown in degenerated discs, and vertebral endplate pathologies may be even more innervated than intervertebral disc pathologies [28, 30]. The extent of change in innervation varies between individuals and may partly be explained by differences in genetics [31].

<u>In summary</u>, obesity or overweight [32], lack of physical conditioning or too much physical load [33, 34], smoking [35], and genetics [24, 36] may be associated with chronic LBP. Central sensitization (amplified sensory signals in the central nervous system, see definitions) may be important in the development and maintenance of chronic LBP [37]. A weak association between chronic LBP and degenerative MRI findings has been shown. However, the high prevalence of such findings in people without LBP and lack of evidence for an association between new MRI findings and development of LBP [38], makes it difficult to state a direct link and causal relationship between chronic LBP and degenerative lumbar MRI findings.

- Treatment for chronic non-specific LBP

In the "European guidelines for the management of chronic non-specific LBP" (2006) cognitive behavioural therapy, supervised exercise, brief educational interventions and multidisciplinary (bio-psycho-social) treatment are advised. Surgery for chronic non-specific LBP is not recommended unless 2 years of all recommended conservative treatments - including multidisciplinary approaches with combined programs of cognitive intervention and exercises - have failed [2]. Thus, the two main treatments for chronic non-specific LBP are:

1) Non-surgical/ conservative treatment
Monodisciplinary treatments such as general practioner care/analgesics and exercise therapy are recommended to be tried first, and then – if necessary and available - multidisciplinary rehabilitation with focus on normal activity, intensive exercise programme and cognitive behavioural intervention [2]. In a prospective study, common degenerative MRI findings were not associated with bad outcome after conservative treatment (active physiotherapy, muscle reconditioning or low impact aerobic/stretching exercises) [39].

2) Surgical/ invasive treatment
In selected patients with failed conservative treatment, localized disc degeneration seen on MRI is a target for lumbar surgery with fusion or disc prosthesis [40-42]. Surgical treatment with fusion is based on the assumption that movement in a degenerated spinal segment can cause pain and that stabilisation of the segment reduces or eliminates pain [29, 43]. Fusion of one or more segments may over time lead to degeneration of adjacent spinal segments because of increased stress [44]. To avoid this disadvantage of fusion, and based on the good results for artificial prostheses in major peripheral joints, mobile intervertebral disc prostheses were developed [45]. The clinical outcome, especially the long-term outcome of surgical treatment with disc prosthesis compared to conservative treatment and natural course is still unclear [43, 45]. In a recently published long-term (average 11 years) follow up of chronic LBP patients randomized to multidisciplinary rehabilitation or spinal fusion, patients' self-reported outcomes did not differ between treatment groups [46]. Short-term (2 years) results show no clinical important difference in pain relief or disability between groups treated with fusion vs. disc prosthesis surgery [45]. At 2-year follow-up of patients treated with disc prosthesis vs. rehabilitation, the Oswestry Disability Index (ODI) score had improved more in the disc prosthesis group;

however, the difference did not exceed the pre-specified minimally important clinical difference of 10 points between groups [42].

The indications for surgery are not well defined, but studies on surgical treatment of non-specific LBP have usually included patients with longstanding LBP (more than 1 or 2 years) and maximum two lumbar levels with degenerative findings on MRI assumed to be painful [2, 43, 45]. However, as stated above, associations between MRI findings and LBP are weak [18, 47]. No test for identification of the painful segment exist [29], and no prognostic tests have been established to aid in clinical decision-making regarding the surgery [48].

In many studies and in clinical practice subgrouping of patients with non-specific LBP based on pathoanatomical imaging findings has been tried based on the assumption that a degenerated segment gives segmental pain. The result is many different terms not recognized as diagnostic entities such as degenerative disk disease, discogenic pain, spinal instability and degenerative lumbar spondylosis [2]. Such terms are also used in Cochrane reviews on fusion and disc prosthesis surgery [43, 45]. Abnormal movement or abnormal disc load (often denoted as spinal or segmental instability) has been proposed as possible causes of LBP, but relationship between instability on imaging and symptoms is controversial [29, 43, 49]. In a review, Rahme and Moussa concluded that Modic type I seems to be associated with LBP and segmental instability, and appear to predict better outcome after lumbar fusion [50]. Pre-treatment presence of Modic type I and/or II is reported to predict better outcome after disc prosthesis surgery (total disc replacement) [51].

In summary, treatment of chronic non-specific LBP is still a challenge despite many years with research, improved imaging (from x-ray to MRI) and more advanced treatment options (from inactivity and bed rest to multidisciplinary rehabilitation and surgical treatment with fusion or disc prosthesis) [52].

1.3 Imaging in patients with LBP

- Indications and imaging technique

There is consensus [2, 3, 6, 53] that lumbar MRI is recommended in LBP patients with 1) red flags indicating serious underlying disease (for example cancer, infection, fracture) and / or 2) LBP duration 4-6 weeks without improvement. Henschke et al evaluated 25 red flag questions for detecting serious spinal pathologies in primary care. They found that 80 % of 1172 patients had at least one red flag, but only 0.9 % had serious disease (11 cases, 8 with fractures) [54]. There is consensus about the importance of diagnostic triage [2], but the red flag approach will inadvertently lead to imaging of many patients without serious disease [54]. According to European LBP guidelines [2], "Individual red flags do not necessarily link to a specific pathology, but indicate a higher probability of an underlying condition that may require further investigation. Multiple red flags need further investigation. Screening procedures for diagnoses that benefit from urgent treatment should be sensitive." High sensitivity but low specificity of red flags was confirmed in a recent review, where it was stated that "The available evidence indicates that in patients with LBP, an indication of spinal malignancy should not be based on the results of one single "red flag" question" [54].

Today MRI is the first choice image evaluation method in patients with LBP because of its excellent anatomical depiction and non-ionizing imaging technique [2, 55]. A general-purpose lumbar MRI protocol is recommended to include [55, 56]:

- Sagittal T1, usually spin echo (SE), fast spin echo (FSE) or fluid-attenuated inversion recovery (FLAIR)
- Sagittal T2, usually FSE
- Axial T2 (usually FSE) and/or T1 (SE or FSE)

The sagittal images should cover from TH12 to S1 and the axial images L3 to S1 (**Figure 3**) [55, 56]. T1-weighted sequences depict anatomy very well whereas T2 weighted sequences depict fluid as bright signal and are well suited to detect soft tissue and bone marrow oedema, and assess the intervertebral disc [56].

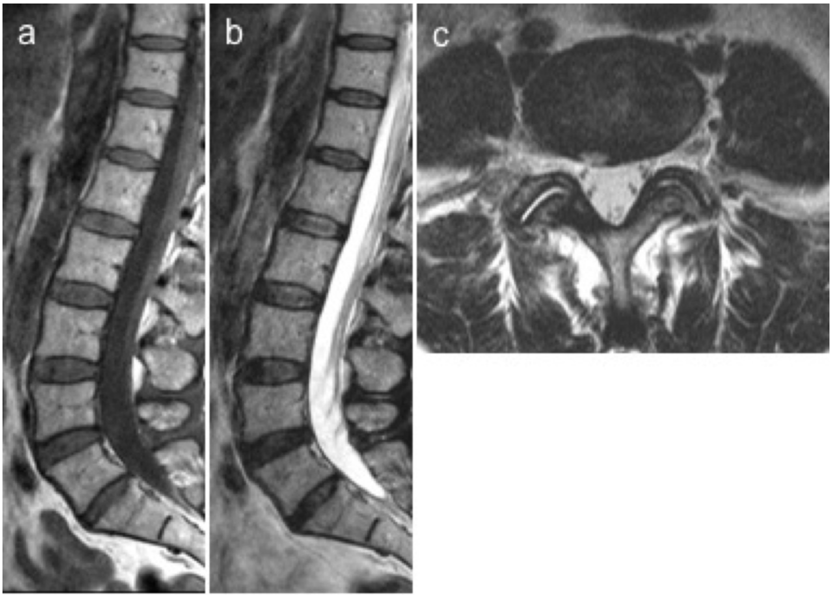


Figure 3 The figure illustrates the recommended sequences a) sagittal T1 Fast Spin Echo (FSE), b) sagittal T2 FSE, and c) axial T2 FSE.

- Degenerative MRI findings

The lumbar MRI findings commonly regarded as degenerative are disc height decrease, loss of nucleus pulposus signal on T2-weighted sequences, annulus fibrosus fissures (including HIZ), Modic changes, changes in disc contour (disc bulging, protrusion, extrusion or sequestration), endplate irregularities and Schmorl´s nodes, osteophyte formation, and FA. Malalignment, ligamentous signal changes, spinal stenosis, and fluid/ vacuum/ calcification in the disc are also regarded as degenerative MRI findings [19, 24, 28]. Degenerative findings are most prevalent at the lowest lumbar levels, which are the most highly loaded areas of the lumbar spine. This is in accordance with the localisation of osteoarthritis in the knee and hip [57].

In the spine as in other anatomical locations, pathological degeneration is difficult to differentiate from normal aging. A conventional understanding has been that aging and degeneration are two similar processes, and that degeneration is accelerated aging [28]. In contrast some researchers propose two different processes for disc degeneration: 1) normal aging affecting mainly annulus fibrosus and adjacent apophyses, and 2) degeneration affecting mainly nucleus pulposus and vertebral endplates with extensive fissuring in the annulus fibrosus [28, 58, 59].

No universally accepted standard definition for disc degeneration exists and this has resulted in a wide range of different definitions, and consequently difficulties in comparing results from different studies [24]. Different degenerative MRI findings are defined and described in the following, with emphasize on findings studied in this thesis.

Modic changes: Modic changes are signal intensity changes in the vertebral body marrow adjacent to the endplate. Three types of Modic changes (type I, II and III) have been described, but mixed types also exist (**Figures 4 and 5**) [50, 60]. Type I has hypointense T1 signal and hyperintense T2 signal, and histologic studies show endplate disruption and fissuring, thickened trabeculae, and vascularized fibrous tissue in the adjacent marrow [60], maybe secondary to inflammation and/or trabecular micro-damage [59]. Type II has hyperintense T1 signal and iso- or slightly hyperintense T2 signal, and histologic studies show disruption of the endplate with markedly thickened trabeculae and granulation tissue suggesting chronic repetitive trauma, and fat replacement of adjacent hematopoietic marrow [60]. Type III has hypointense T1 and T2 signal, may show sclerosis on radiographs, and probably reflects relative absence of bone marrow in areas with densely woven bone (sclerosis) [60].

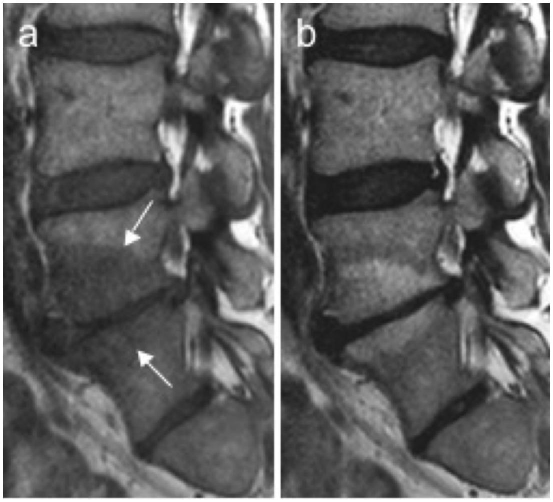


Figure 4 Sagittal T1-weighted (a) and T2-weighted (b) MRI images from one patient illustrating Modic type I (white arrows) superior and inferior to the disc at L4/L5.

In a review, a median of 16 % of individuals (with and without LBP) had Modic type I (at one or more levels), 26 % had type II, less than 1 % had type III, and 13 % had mixed types, and any type of Modic changes in about 36 % of individuals [47]. Based on meta-analyses of reviewed studies Jensen et al concluded that the prevalence of any Modic changes increases with age (11 % per 10 years) [47]. Modic type I changes usually convert to type II over time (often at least 1 year), but may sometimes convert to normal bone marrow signal [61]. Modic type II is more stable than type I, but may convert to type I or type III over time [60, 62, 63].

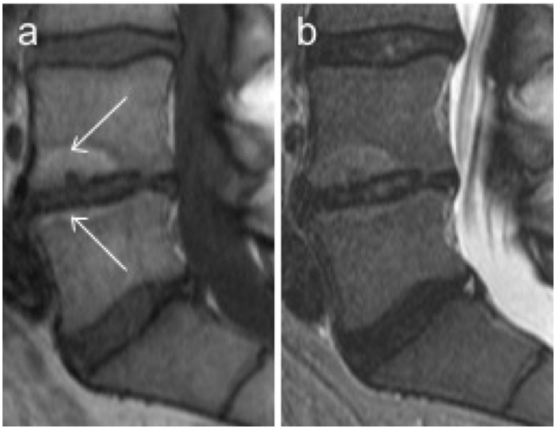


Figure 5 Sagittal T1-weighted (a) and T2-weighted (b) MRI images from one patient illustrating Modic type II superior and inferior (white arrows) to the disc at L4/L5.

High-intensity zone (HIZ): A posterior HIZ is an area of high-signal intensity in the posterior annulus fibrosus that is brighter than nucleus pulposus on T2-weighted MRI images, and is surrounded superiorly, inferiorly and anteriorly by the low-intensity (black) signal of the annulus fibrosus (**Figure 6**) [64, 65].

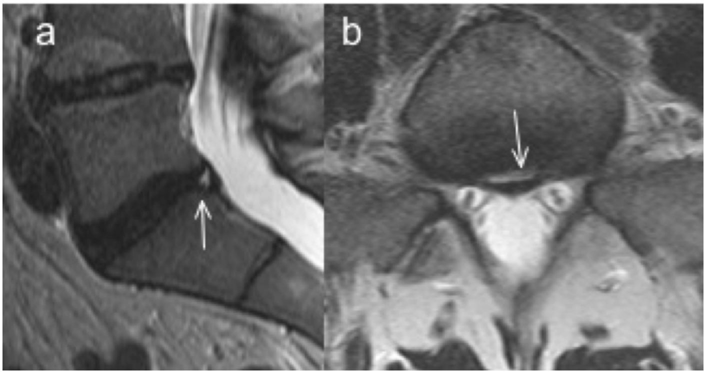


Figure 6 Sagittal T2-weighted (a) and axial T2-weighted (b) MRI images from one patient illustrating high-intensity zone (HIZ) in the posterior disc at level L5/S1 (white arrows).

Annulus fibrosus consists primarily of collagen type I fibres, and annulus ruptures (concentric or radial) are regarded as ruptures or avulsions of these fibres [28, 66, 67]. HIZs correspond to grade 3 or grade 4 annular ruptures on computed tomography (CT) discograms [64, 68]. Grade 3 ruptures extend into the outer third of the annulus fibrosus, and grade 4 is a grade 3 with a circumferential component of more than 30° (**Figure 7**) [64].

Only about half of grade 3 and 4 radial tears on CT discography were seen as a HIZ on MRI [64]. In a study by Smith et al the sensitivity of HIZ for detecting a grade 4 annular rupture was only 31 %, but the specificity was 90 %. According to Smith et al, the varying association between HIZ and grade 4 annular tears may be due to different stages of healing and patient-specific healing responses, and the presence of a HIZ cannot be used to accurately predict a painful disc [68].

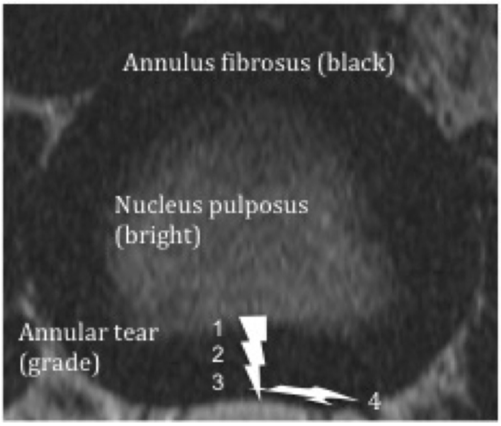


Figure 7 The figure illustrates the grading of annular tears (grade 1 affect inner third, grade 2 middle third, grade 3 outer third, and grade 4 equals grade 3 but has a circumferential component more than 30° in addition [64, 68].

Nucleus pulposus signal: Nucleus pulposus consists of proteoglycan and water held together by collagen type II and elastin fibres. Signal loss in nucleus pulposus is thought to result from change in glycosaminoglycan concentration and water state [58, 69]. On sagittal T2-weighted MRI images nucleus pulposus signal can be visually graded using CSF as intensity reference as bright, grey, dark or black [70] (**Figure 8**).

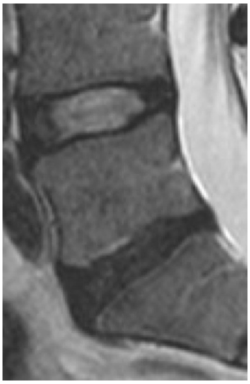


Figure 8 Sagittal T2-weighted MRI image illustrating normal disc signal (bright with clear distinction between nucleus pulposus and annulus fibrosus) and disc height at the upper disc level, and at the lower disc level moderate disc height decrease (disc narrower than the disc above) and dark nucleus pulposus signal.

Disc height decrease: In the normal spine disc height gradually increases from L1 to L5, but is more variable and usually lower at L5/S1 than at L4/L5. Disc height decrease can be measured, or subjectively rated in relation to the height of nearest normal disc above (higher, as high as, narrower, or endplate almost in contact) [71-74] (**Figure 8**).

Disc contour: According to the recommendations for nomenclature and classification of lumbar disc pathology proposed by Fardon and Milette [67], a disc bulge (symmetrical or asymmetrical) extends beyond the edges of the disc in the axial plane and includes more than 50 % of the disc circumference ("usually less than 3 mm beyond the edges of the vertebral body apophyses"). A disc herniation is "a localized displacement of disc material beyond the normal margins of the intervertebral disc space" and can be subdivided into focal or broad-based protrusion, extrusion and sequestration [67].

Facet arthropathy (FA): The facet joints are synovial joints. Many different terms such as facet arthritis, facet joint syndrome, facet disease, facet hypertrophy, degenerative facet joints, and FA are used for facet joint pathology. In general, these terms are synonymous and imply findings consistent with degeneration of the facet joint [75]. Different classification systems exist for FA, but they are commonly based on registration of joint space narrowing, osteophytes, subchondral sclerosis, articular process hypertrophy and erosions [76, 77] (**Figure 9**). Based on a systematic review of existing grading systems, Kettler and Wilke recommended to grade facet joint degeneration in 3-5 grades with grade 0 denoting a normal facet joint [76]. Fujiwara et al have developed such a grading system for FA (0: normal, 1: mild (joint space narrowing or mild osteophyte), 2: moderate (sclerosis or moderate osteophyte), 3: severe (marked osteophyte) [77].

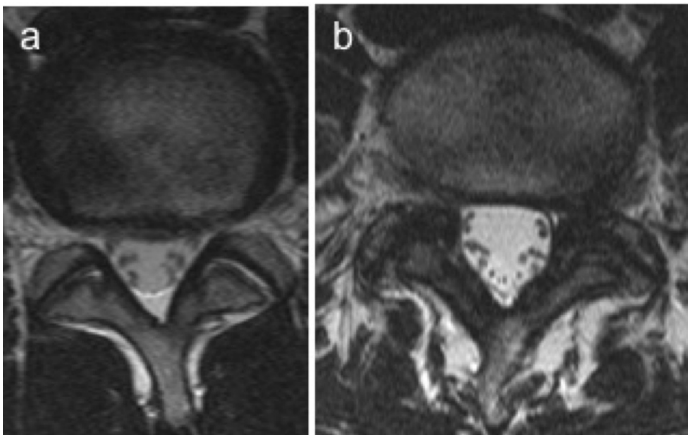


Figure 9 Figure illustrating normal facet joints (a), and moderate to severe FA with joint space narrowing, sclerosis and moderate to severe osteophytes (b).

Some surgeons consider moderate or severe FA as a contraindication to disc prosthesis surgery [41, 78, 79]. Other surgeons do not consider FA as a contraindication [42].

Other degenerative findings: *Osteophytes* at the vertebral margins can be found anteriorly, posteriorly and laterally, and is usually graded according to localisation and size [72]. *Endplate defects* include intravertebral disc herniations through the endplate (Schmorl´s nodes/localised endplate defects) and/or *irregular endplates* (intact but irregular endplates) [67, 80]. *Malalignment*, *ligamentous signal changes*, and *spinal stenosis* are often regarded as complications of disc degeneration or FA [28]. *Fluid*, *vacuum*, and *calcification* in the disc may also be a part of disc degeneration. Vacuum represents gas in disc fissures with negative pressure, formed in a stiff, degenerated disc more vulnerable to compressive and rotational forces because of reduced water and proteoglycan content and increased fibrosis [28]. We seldom see gas and often see fluid in the disc on MRI [81]. Gas filled spaces in the disc (vacuum) are gradually filled with fluid during 20 minutes in a MRI machine (oral impart by Dr. Noubauer-Huhmann on the "Advanced MRI imaging of the musculoskeletal system", School of MRI 2013 in Bergen November 7-9, 2013).

Coexistence of degenerative MRI findings: Degenerative MRI findings often coexist. In a study by Kleinstuck et al, all disc bulges, HIZs, and Modic type I and/or II changes occurred at discs with at least grade 3 degeneration (on a 5-point scale) and nearly all (89-100 %) occurred at discs with severe degeneration (grade 4 or 5) [39]. Disc bulge was associated with HIZ ($P = 0.0001$) and with Modic type I and/or II changes ($P = 0.0001$) [39]. Similarly, Albert et al reported that Modic changes were associated with disc degeneration [82]. At 3 year follow up, Kuisma et al only found new Modic changes at discs with at least some degeneration at baseline [83]. Modic changes have also been reported to occur at a late stage of disc degeneration [84].

Other studies have shown relationships between reduced disc signal and reduced disc height [85] and between disc degeneration, disc herniation and HIZ [86]. Emch [66] and Modic [28] reported disc height decrease only in discs with reduced T2 signal on MRI. They also discussed that FA can be secondary to disc height decrease (because of changed biomechanics and load to the facet joints) but can also occur independent of disc degeneration [28, 66].

Change in degenerative MRI findings over time: Most degenerative lumbar MRI findings change slowly. In a population study of Finnish male twins, an average age increase of 17.4 years was associated with the presence of one more endplate with Modic changes [87]; After 3 years, Kuisma et al [83] found new Modic changes at 6 % (13/230) of disc levels without Modic changes at baseline; Modic type was unaltered at 86 % (60/70) of discs levels with Modic changes at baseline, but the extent of existing Modic changes type II and mixed type I/II had increased significantly. Battie et al reported that of the common measures of disc degeneration (disc signal, height, and bulge), disc signal changed most over a 5-year period [24].

Prevalence of degenerative MRI findings in people without LBP: In a systematic review by Chou et al the prevalence of degenerative lumbar MRI findings (disc degeneration, disc protrusion, reduced disc height, annular tear, HIZ, endplate changes, Modic changes, FA) in persons without LBP was up to 60-80 % [27]. In people without LBP, the reported prevalence of degenerative MRI findings at one or more levels from L1-S1 was for

- Modic changes 6 % (median; range about 0-25 %; systematic review) [47].
- HIZ 28 % (range 6-56 %; systematic review) [18].
- Any disc height decrease 56 % [88]; 38 % for disc lower than a normal disc above or endplates almost in contact [89] but 28 % for disc as low as or lower than a normal disc above in another study [72].
- Disc degeneration 54 % (range 7-85 %; based on reduced disc height and/or reduced T2 signal from the disc; systematic review) [18].
- Reduced disc signal 53 % (mean; range 30-83 %; systematic review; grading of signal differed between studies) [18].
- Reduced T2 signal from nucleus pulposus 30 % [89].
- Disc herniation or bulge 25-50 % [90-92], and disc protrusion or extrusion 22 % [89].
- FA 3-76 % (range; systematic review; definition of FA varied substantially between studies) [18], and slight or severe FA 36 % [89].

Osteophytes at the anterior and lateral margins of the vertebral are found in 100 % of individuals older than 40 years and are regarded as a part of normal aging whereas posterior osteophytes are rare even in individuals older than 80 years and are regarded as pathological [28]. Endplate defects have been reported in 20-30 % of people without LBP [18].

Association between MRI findings and LBP: In a systematic review on Modic changes, Jensen et al included studies of acute and chronic LBP patients and of individuals with disc herniations with and without sciatica [47]. They found that:

- Modic changes were associated with LBP (median OR 3.4, range 2.0-19.9) based on data from the 10 included studies.
- The median prevalence of Modic changes was 43 % (range about 15-90 %) in patients with non-specific LBP.

In another systematic review, Endean et al [18] reported OR for LBP for different MRI findings:

- Disc degeneration: OR 2.5 (95 % CI, 2.0-7.4). Separate results for disc height and disc signal were not available. Definition of disc degeneration was heterogeneous among included studies but based on presence or severity of reduced disc height and/or reduced T2 signal from the disc.
- Disc herniation: OR 3.6 (95 % CI, 1.8-7.0).
- FA: Not related to LBP, OR (95 % CI) reported in included studies was 1.1 (0.7-1.6) and 4.4 (0.9-21). The definition of FA varied substantially between studies and no meta-estimate was calculated for OR.
- Endplate defects: Not related to LBP, OR 0.9 (95 % CI 0.6-1.4).

The reviewed studies included acute and chronic LBP patients with and without radicular pain. The way absence of LBP was defined varied between studies. In the review they chose to give preference to those who had been pain free for at least 12 months [18].

In a study of male twin pairs, Videman et al [72] found that:

- Signal intensity of nucleus pulposus was not significantly associated with any LBP parameter neither in univariate nor in multivariable analyses.
- Osteophytes were not significantly associated with any LBP parameter, when controlling for age, clustering by twin pairs, and other MRI findings.

In a population study of 40 years old [89] the reported prevalence was for:

- HIZ (defined as bright white signal located in the substance of the posterior annulus fibrosus and clearly dissociated from nucleus pulposus) 41 %, and those with HIZ had an OR for LBP the past year of 2.5 (95 % CI, 1.6-3.9).
- Hypointense nucleus signal 45 % and the finding was significantly related to LBP during the past year (OR 2.6, 95 % CI, 1.7-4.0).
- Moderate or severe disc height reduction (disc narrower than the disc above (if normal) or endplates almost in contact) 53 %, and in this group OR for LBP the past year was 2.5 (95 % CI, 1.6-3.9).

- Disc herniation (less than 50 % of the circumference of the disc) 25 %, but this finding was not related to LBP the past year (OR 1.3; 95% CI, 0.8-2.1).
- Slight or severe FA at one or more levels 37 %, but this finding was not associated with LBP the past year (OR 1.1; 95% CI, 0.7-1.6).

Few studies have investigated the association between LBP related disability and MRI findings. In one study, Kaapa et al [93] found that:

- Patients with Modic type I reported worse disability ($P = 0.0156$) and LBP intensity scores ($P = 0.0451$) than patients with mixed type I and II.
- Size of Modic type I (in % of the sagittal area of the corresponding vertebrae) was not related to disability or LBP pain intensity scores.

Risk factors for lumbar spine degeneration: Genetics, mechanical strain, and metabolic factors are considered as important factors in the degenerative process of the lumbar spine [19].

Heredity may be a major determinant of disc degeneration. Familial aggregation (early environmental influence and heredity combined) explained 61 % of variance in disc degeneration in the TH12-L4 region and 32 % in the L4-S1 region in a male twin study - age explained 16 % and occupational physical loading 11 % [24]. Similarly, in another twin study, heredity explained 74 % of lumbar disc degeneration - disc height decrease, disc bulge, and osteophytes, but disc signal did not appear highly heritable. They concluded that disc signal probably is influenced predominantly by environmental factors such as age dependent change in hydration and proteoglycan content [94]. Yet, in a recent systematic review on genetics and association with lumbar disc degeneration “The phenotype definition of lumbar disc degeneration was highly variable between the studies and replications were inconsistent. Most of the associations presented with a weak level of evidence” [95].

Lumbar degenerative changes in women tend to occur approximately ten years later than in men [96]. Reduced disc signal is the lumbar MRI degenerative finding most strongly associated with age, and is reported in 90 % of people without LBP older than 60 years [24, 66, 97].

Routine physical loading may have a small positive effect on the disc with respect to disc degeneration whereas smoking may have a small negative effect [24]. Load at work is probably not associated with disc degeneration [98]. Overweight and obesity are reported to be associated with presence, extent and severity of disc degeneration [99].

Risk factors for FA include age, genetics, anatomical properties (for example malalignment, orientation of the facet joint space, spinal level), overweight, occupational factors (for example heavy physical loading and sedentary work), and disc degeneration [75].

In summary, LBP seems only weakly related to degenerative MRI findings (Modic changes, HIZ, degenerated discs) and probably no direct link between individual degenerative MRI findings and LBP exist [18, 27, 47, 66]. A systematic review on radiographic findings showed similar results for degenerative radiographic findings (reduced disc height, osteophytes, sclerosis) and indicated no firm evidence neither for the absence nor the presence of a causal relationship between such findings and non-specific LBP [100].

Although degeneration of the lumbar spine (disc degeneration and FA) has been a proposed cause of chronic LBP the results from studies on this topic have been conflicting [27]. Conflicting results may be due to low methodological quality, different definitions of LBP (location, duration, frequency and intensity), different populations (gender, age, physical work load and leisure time activity), and differences in imaging (type and quality of equipment, grading of findings) [101]. According to Jarvik and Deyo, for an imaging test or finding to be useful it must be associated with clinical findings, help to differentiate disease states and/or have prognostic value – but first it must demonstrate adequate reliability [102]. Although good reliability of an imaging finding does not imply that the finding is clinically helpful, it does provide a basis for further studies, e.g. of relationships between the imaging finding and clinical variables.

1.4 Reliability of MRI findings

- Definition

Reliability in the context of clinical tests is defined as the extent to which the observers agree in their ratings, and reliability is dependent on both repeatability (agreement when measurement is repeated under the same conditions, intraobserver agreement) and reproducibility (agreement when measurement is repeated under different conditions, interobserver agreement) [103, 104].

- General considerations on reliability of radiological findings

Reliability of diagnostic procedures has been studied for more than 60 years and many factors influencing agreement have been elucidated (for example available clinical history, double interpretation, observer experience, interpretation time used, and availability of previous images). Despite of this, only moderate agreement on radiological findings can be expected [105, 106]. Only moderate or poor agreement on findings can be expected also for clinical tests and diagnoses [105]. Lack of agreement may be due to error (for example poor technique, failures of perception, lack of knowledge, misjudgements) or variation in interpretation (cases where experts fail to achieve consensus) [106]. Disagreement rate in radiology has been reported to be about 30 % if only abnormal images are rated and about 4 % if a mixture of normal and abnormal images are rated [105].

- Methods to measure agreement on type of MRI finding

Percentage of agreement has been used as a measure of agreement, but does not take into account agreement expected purely by chance, and according to Sim and Wright only agreement beyond that expected by chance is considered as “true” agreement [104].

The kappa statistic is a recommended measure of agreement for categorical variables [103, 104, 107]. Kappa measures the proportion of agreement beyond the agreement expected by chance alone; kappa is the ratio of observed non-chance agreement to possible non-chance agreement [108]. Kappa is defined as the difference between observed and expected agreement (by chance) expressed as a fraction of the maximum difference. Kappa = (observed agreement - expected agreement) / (1 - expected agreement) [104]. The range of possible kappa values is from -1 to 1 (the minimum value tends towards 0 for more than 2 observers) [103, 107]. Kappa (k) takes the value of 1 when agreement is perfect, zero when agreement is no better (and no worse) than chance alone and negative values when agreement is less than expected by chance alone. There are no clear cut-off values indicating acceptable agreement, but the following interpretation proposed by Altman has been widely used: $k \leq 0.20$: poor, 0.21- 0.40: fair, 0.41- 0.60: moderate, 0.61- 0.80: good and 0.81- 1.00: very good agreement beyond chance [104, 109].

The magnitude of kappa is influenced by the prevalence of the MRI finding, systematic disagreements between observers (bias), and non-independent ratings [104]. When the prevalence of a positive rating is either very high (> 90 %) or very low (< 10 %) the chance agreement is very high and consequently kappa is reduced. This effect of prevalence on kappa is more pronounced for large values of kappa than for small values [104]. Large bias results in an increase of kappa, which is more pronounced for small kappa values than for large kappa values [104]. The effect of prevalence and bias induces two paradoxes: kappa can be lowered despite high actual agreement and increased despite low actual agreement [104]. Independent ratings mean that the observers must be blinded to each other's ratings and their own prior ratings [104].

When the number of rating categories is increased, disagreement is potentially higher and a lower unweighted kappa value can be expected [104]. *Unweighted* kappa treats all disagreements equally (e.g., disagreement between category 1 and 4 on a scale is considered equally serious as disagreement between category 1 and 2) and is therefore considered inappropriate for ordinal scales [104]. If there are more than two categories we can choose to use *weighted* kappa, which reflects the seriousness of the disagreement (e.g., gives more weight to disagreements between category 1 and 4 than between category 1 and 2 on a scale). Different weighting systems can be applied (such as linear or quadratic) for weighted kappa, and will have different impact on the magnitude of kappa [104].

Table 1 Cross-table for presence of a finding (Yes/No) rated by observers A and B

		Observer A		
		Yes	**No**	**Totals**
Observer B	**Yes**	**a**	**b**	**a+b**
	No	**c**	**d**	**c+d**
	Totals	**a+c**	**b+d**	**N**

Prevalence- and bias-adjusted kappa (PABAK) has been developed to compensate for the influence of prevalence and bias on the magnitude of kappa. The adjustments imply to substitute the actual values a and d in a contingency table (**Table 1**) with the mean of a and d, and substitute the actual values b and c with the mean of b and c. PABAK reflects a hypothetical situation with no bias and no difference in prevalence between observers [104]. There is some opposition to the use of PABAK [104], since important information regarding prevalence and systematic differences in rating is not taken into account. In this hypothetical situation, it is equally easy/difficult to agree on “Yes” as it is to agree on “No” (**Table 1**). This may not be true in the actual situation, where it may for example be easier to agree on lack of a finding (No) than to agree on presence of the finding (Yes). Although PABAK has limitations, it can be helpful for assessing agreement when the prevalence is low and the ordinary kappa value may be difficult to interpret.

- Methods to measure disagreement on prevalence of MRI findings

Kappa cannot be used to assess disagreement between observers (b and c, **Table 1**) and whether disagreement is random (due to chance) or due to a consistent pattern (systematic differences) [104]. An imbalance between the magnitude of b and c (**Table 1**) suggests a systematic difference in the rating between the two observers (bias). McNemars test can be used to analyse whether bias is significant by comparing the proportion of positive findings either between two observers or between one observer's first and second rating [103]. There are also methods for comparing the proportion of findings between more than two observers [110].

1.5 Specific background for studies included in the thesis

- Content of and motivation for the studies

We studied patients with chronic non-specific LBP and localized degeneration (at L4/L5 and/or L5/S1) who were candidates for surgery with lumbar disc prosthesis. We examined the reliability for type and prevalence of degenerative MRI findings and for change in such findings over time (paper I and II) and the relationship of the sum of MRI findings to the degree of pain/disability (paper III). We analysed Modic changes, HIZ, disc contour, disc height, nucleus pulposus signal, and FA.

In candidates for lumbar disc prosthesis, such degenerative MRI findings and change in these findings over time are relevant to study for several reasons. Localized disc degeneration seen on MRI is a presumed source of pain and a target for lumbar surgery with fusion or disc prosthesis [40-43, 45]. It is therefore relevant to evaluate the reliability of these MRI findings and their association with complaints; a clear association would support their use in decisions on surgery. Reliable evaluation of the findings' change over time is needed to assess outcome and adverse effects (such as adjacent level degeneration) after treatment [111-113]. Disagreement on type or prevalence of MRI findings can lead to underestimation of the findings' potential relationship to clinical features, incorrect treatment decisions, and faulty assessment of beneficial and adverse effects of treatment [102, 112, 114, 115].

- Previous research on reliability of the studied MRI findings

Inter- and intraobserver agreement: Moderate to good agreement has been reported for most lumbar degenerative MRI findings with generally slightly better intra- than interobserver agreement (**Table 1 in appendix**). The reported range of kappa values was for:

- Modic changes; presence/type 0.31-0.85 (interobserver = inter) and 0.64-1.00 (intraobserver = intra), extent: 0.43-0.80 (inter) and 0.60-0.83 (intra).
- HIZ 0.44-0.86 (inter) and 0.67-0.97 (intra)
- Nucleus pulposus signal 0.38-0.93 (inter) and 0.75-0.91 (intra)
- Disc height reduction; subjective 0.45-0.74 (inter) and 0.51-0.81 (intra), measured 0.58 (inter, one study) and 0.77-0.99 (intra)
- Disc contour 0.55-0.75 (inter) and 0.69-0.79 (intra)
- FA 0.07-0.54 (inter) and 0.26-0.76 (intra)

In the previous studies there was considerable variation in MRI equipment, classification of MRI findings, number of findings and spinal levels evaluated, size and characteristics of the patient sample, prevalence of MRI findings, number of observers, years of experience and speciality of observers, statistics used (weighted or unweighted kappa), and in time from first to second rating of images for intraobserver agreement analysis (**Table 1 in appendix**). Few of the previous reliability studies reported and took into account the effect of prevalence [65, 71, 73, 80, 116, 117] and bias on kappa [65, 71, 118], and clustering of data [65, 116, 117] (**Table 1 in appendix**). Differences between observers in reported prevalence of MRI findings received little attention [65, 119]. Only one prior study concerned disc prosthesis patients and it was restricted to FA [120]. It seems that no previous study has addressed the reliability of combined MRI findings used as indication for disc prosthesis.

<u>Agreement on change in MRI findings over time – comparison of images:</u> To our knowledge, only one study (of Modic changes) has examined the reliability of change in lumbar spine MRI findings over time [80]. Comparison of old and new images, as in daily clinical practice, provided only moderate reliability; therefore, independent evaluation of initial and follow-up images (non-comparison) was recommended when studying the course of Modic changes [80]. However, it was unknown whether

comparison of images is more or less reliable than non-comparison when evaluating change in MRI findings over time. Both approaches are used in research [83, 113, 121-123].

Influence of disc prosthesis artefacts on agreement: Disc prosthesis causes artefacts on MRI. It had not been assessed how such artefacts might influence the reliability when evaluating MRI findings adjacent to the prosthesis.

- Research on sum of MRI findings in relation to pain/disability

As previously summarized the presence of LBP has been reported to be only weakly related to degenerative findings on MRI [18, 27, 47, 66]. However, it is not clear how the *sum* of such MRI findings may be related to the *degree* of disability and pain within specified groups of LBP patients [18, 89]. Degenerative MRI findings often coexistence, e.g. it is rare for Modic changes [39, 82] or FA to appear without disc degeneration [75, 124]. Kleinstuck et al discussed whether the presence of a MRI finding in itself is enough to induce LBP or if a certain number of levels must be involved and/or a certain severity of findings is necessary [39]. Studies on combined MRI findings [18] and more accurate and reproducible stratification of patient cohorts [28] are advised to improve our understanding of the relationship between MRI findings and clinical features in LBP patients.

However, few have studied the association between multiple lumbar MRI findings combined and LBP or disability. The combination of various MRI findings provided little explanation for disabilities in two mixed samples of LBP patients with and without radicular pain or sciatica [125, 126]. Mariconda et al based a MRI summary score on disc signal/bulge, disc height, disc herniation, FA, spinal stenosis, and degenerative spondylolisthesis. This score was weakly related to pain duration (regression coefficient = 9.38; 95 % CI, 2.21 to 16.55; $P = 0.011$) and degree of disability (regression coefficient = 0.92; 95 % CI, -0.002 to 1.84; $P = 0.050$) [125]. Arana et al found that combined MRI findings (disc and facet findings, spinal stenosis, and other pathologies) were not related to disability or LBP [126].

In a twin study, McGregor et al used a MRI severity sum score based on a 4 point grading of nucleus pulposus signal, disc height, disc bulge, and anterior osteophytes. This score was the covariate with the strongest association to severe LBP (with associated disability) with OR 3.6 (95 % CI, 1.8 to 7.3) for LBP in the quartile group with highest MRI score vs. the quartile group with lowest MRI score. They concluded that the association between MRI score and LBP probably was mediated genetically [25]. Kjaer et al recorded the prevalence of LBP variables (such as LBP past year, seeking care for LBP, previous LBP episodes) and found higher prevalence in persons with disc degeneration and Modic changes than in patients with disc degeneration only (e.g., LBP past year 91.8 % vs. 70.9 %) [14].

In patients with chronic non-specific LBP accepted for disc prosthesis surgery, the localized degenerative MRI findings are assumed to reflect relevant pain sources. Therefore, in this patient group, a dose-response relationship may exist between the extent of the MRI findings and the degree of disability and pain. To our knowledge no previous studies have compared the sum of MRI findings to the degree of LBP and disability in such a specific and well-defined patient cohort.

2.0 Aims of the thesis

The overall purpose of this work was to assess the reliability of lumbar spine MRI findings in candidates for lumbar disc prosthesis, the reliability of change in such MRI findings over time, and their relationship to disability and intensity of LBP.

I. The aim in **paper I** was to assess the reliability of pre-treatment lumbar spine MRI findings in chronic LBP patients who were accepted candidates for lumbar disc prosthesis. At each disc level for each MRI finding - and for combined MRI findings used as MRI indication for disc prosthesis - among experienced radiologists we aimed to evaluate:

a) differences in reported prevalence, and
b) interobserver and intraobserver agreement.

II. The aims in **paper II** were:

a) to assess the reliability of change in lumbar MRI findings over time after disc prosthesis surgery or non-surgical treatment,
b) to assess the impact of the prosthesis on the reliability of adjacent MRI changes, since the prosthesis caused image artefacts, and
c) to compare the reliability between the image evaluation methods of comparison versus non-comparison of images.

III. The aims in **paper III** were as follows:

a) The primary *a priori* aim was to examine whether the sum of MRI findings was related to the degree of disability (as indicated by the ODI score) and the intensity of LBP.
b) The second *a priori* aim was to assess whether each individual of these MRI findings was related to ODI and intensity of LBP.
c) *Post hoc*, we also analysed whether adding FA to the sum of MRI findings and as an individual MRI finding, respectively, affected the results.

3.0 Material and methods

3.1 Patients and eligibility criteria

This PhD project is a radiological subproject of a Norwegian, multi-centre trial on 173 patients with chronic LBP and one- or two-level disc degeneration who were randomized to disc prosthesis surgery or non-surgical treatment (multidisciplinary rehabilitation, including training). The project thus included accepted candidates for disc prosthesis, not chronic LBP patients in general. The trial was initiated at Oslo University Hospital and National Centre for Diseases of the Spine, St. Olavs Hospital and involves all University Hospitals in Norway. Patients were included in 2004-2007; 2-year follow-up was completed in 2009. The pre-treatment and 2-year follow-up MRI examinations form the basis for the present PhD project.

- Inclusion and exclusion criteria

The inclusion criteria used in the trial were:

- age between 25 and 55 years,
- LBP as main symptom for at least one year,
- insufficient effect of structured physiotherapy or chiropractic treatment for at least 6 months,
- ODI score ≥ 30 %, and
- the following MRI findings (defined in section 3.2) reported by the enrolling physicians at L4/L5 and/or at L5/S1 (levels suitable for disc prosthesis):
  (a) ≥ 40 % disc height decrease and/or
  (b) at least two of the following: Modic changes type I and/or II, posterior HIZ in the disc, and dark/black nucleus pulposus on T2-weighted images.

Exclusion criteria were:

- any of the four MRI findings in (a) or (b) at any higher lumbar level (L1-L4)

- or any of the following: spondylolysis, isthmic spondylolisthesis, arthritis (e.g. ankylosing spondylitis, psoriatic/rheumatoid arthritis), osteoporosis (based on dual-energy x-ray absorptiometry or known osteoporotic fractures), prior fracture at L1-S1, prior spinal fusion, congenital or acquired deformity, symptomatic disc herniation / spinal stenosis (based on radiating pain to the lower extremity and corresponding MRI findings), generalized chronic musculoskeletal pain (e.g. fibromyalgia, widespread myofascial pain), drug abuse, on-going psychiatric or somatic disease that excluded either or both treatment alternatives, or lacking understanding of Norwegian language.

Thus, lumbar disc degeneration above the L4/L5 and L5/S1 levels, and symptoms of nerve root affection, were exclusion criteria. FA was not an exclusion criterion.

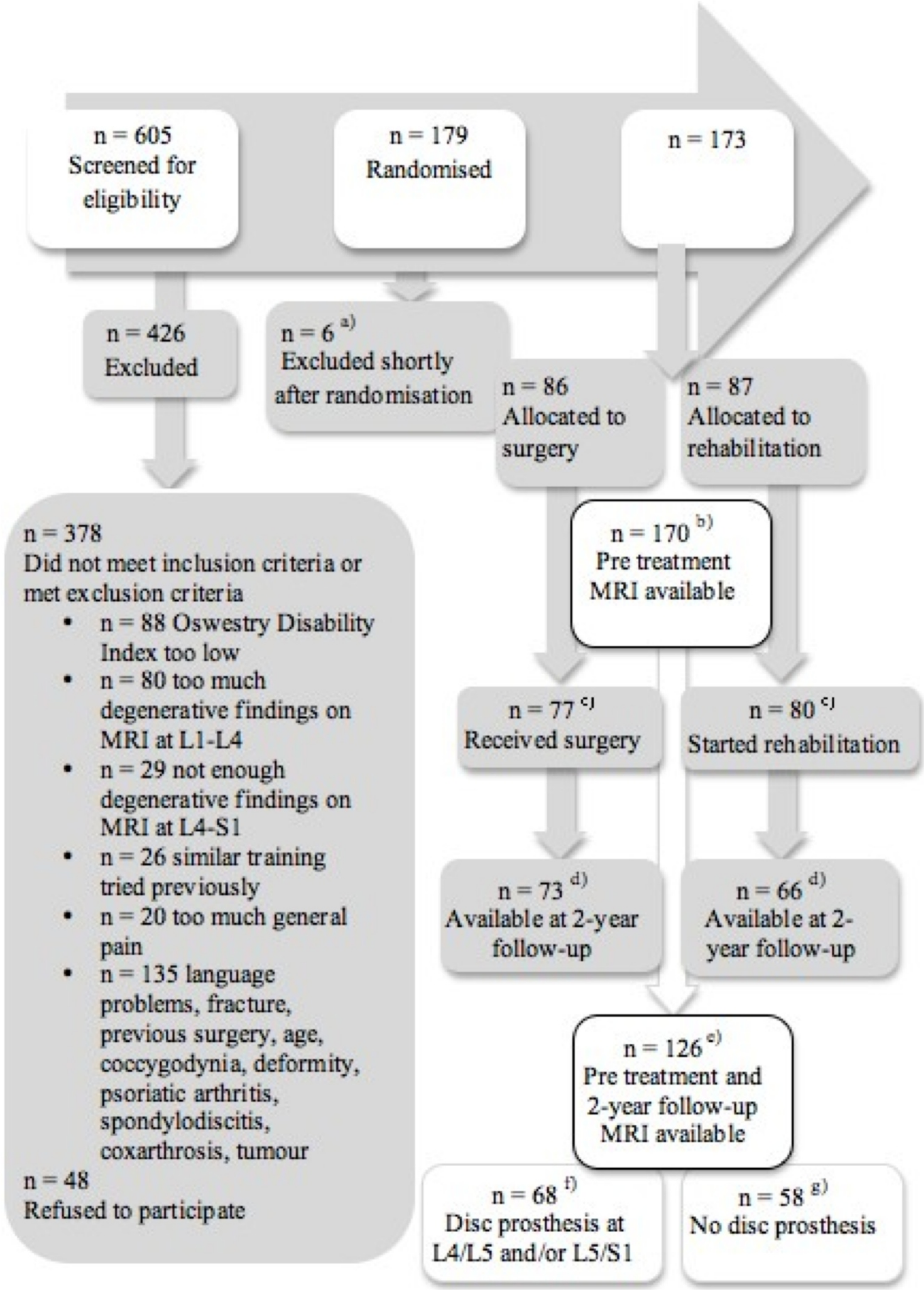
- Ethics

The Regional Committees for Medical Research Ethics in east Norway approved the study. The trial was registered at clinicaltrials.gov (identifier NCT 00394732) and was carried out in accordance with the Helsinki Declaration. All patients provided informed written consent prior to study participation.

- Enrolment and study flow

Patients from all health regions in Norway were recruited in usual clinical practice without any supplemental recruitment attempt [42]. They were referred from local hospitals or primary care to the nearest University Hospital where they were screened by an orthopaedic surgeon or a specialist in physical medicine and rehabilitation. Patients were included after a new appointment where both physicians were present and agreed on inclusion [127]. Two physicians independently evaluated disc and Modic changes on MRI; if they disagreed on inclusion, a third physician independently evaluated the MRI and simple majority decided inclusion. (These initial MRI reports were not used later in this PhD project.) **Figure 10** shows the flow of patients through the trial.

Figure 10 Enrolment, randomisation and follow-up of patients in the trial


n = 605
Screened for eligibility
n = 179
Randomised
n = 173
n = 426
Excluded
n = 6 a)
Excluded shortly after randomisation
n = 86
Allocated to surgery
n = 87
Allocated to rehabilitation
n = 378
Did not meet inclusion criteria or met exclusion criteria
• n = 88 Oswestry Disability Index too low
• n = 80 too much degenerative findings on MRI at L1-L4
• n = 29 not enough degenerative findings on MRI at L4-S1
• n = 26 similar training tried previously
• n = 20 too much general pain
• n = 135 language problems, fracture, previous surgery, age, coccygodynia, deformity, psoriatic arthritis, spondylodiscitis, coxarthrosis, tumour
n = 48
Refused to participate
n = 170 b)
Pre treatment MRI available
n = 77 c)
Received surgery
n = 80 c)
Started rehabilitation
n = 73 d)
Available at 2-year follow-up
n = 66 d)
Available at 2-year follow-up
n = 126 e)
Pre treatment and 2-year follow-up MRI available
n = 68 f)
Disc prosthesis at L4/L5 and/or L5/S1
n = 58 g)
No disc prosthesis

a) 3 in surgery group and 3 in rehabilitation group due to heart attack (n = 1) or obvious exclusion criterion discovered after randomisation (n = 5) i.e. large previous abdominal operation (n = 1), not enough degenerative MRI findings at L4-S1 (n = 2) and too much degenerative findings at L1-L4 (n = 2).
b) For 3 of 173 included patients the pre-treatment MRI was no longer available (never sent to the research database from the examination centre, and not found on request).
c) 9 patients allocated to surgery and 7 patients allocated to rehabilitation changed their mind after randomisation and declined surgery or did not attend for rehabilitation.
d) In the surgery group 4 patients dropped out after treatment (1 serious complication with leg amputation, 2 did not want to attend follow-up, 1 could not be contacted). In the rehabilitation group 14 patients dropped out after treatment start; 6 during treatment (1 did not find the rehabilitation program good enough, 1 could not manage it, 1 operated for lumbar disc herniation, 1 developed diabetes, 1 psychosocial reasons, 1 hypertension and recommended not to train) and 8 after completed treatment (1 participated in another study, 1 did not complete questionnaire, 1 moved, 1 died of cancer, 3 did not want to attend follow-up, 1 unknown).
e) Of the 139 patients available at the 2-year follow-up, 128 patients underwent 2-year follow-up MRI, but the pre-treatment MRI was not available for 2 of these 128 patients.
f) Included are 5 patients from the rehabilitation group operated with disc prosthesis (1 after 6 months, 4 after 1 year) and 1 patient operated with both fusion and disc prosthesis.
g) Includes 1 patient allocated to rehabilitation and operated with fusion only.

Figure 10 shows that 170 of 173 patients allocated to treatment had pre-treatment MRI available for this PhD study (paper I and III); 126 of the 170 patients had 2-year follow-up MRI available as well and were included in paper II. At the 2-year follow-up, 68 (54%) of the 126 patients had disc prosthesis at L4/L5 and/or L5/S1.

3.2 Magnetic resonance imaging

- MRI parameters

MRI was performed as part of clinical practice, using various protocols and magnets, as shown in **Table 2** (n = 170 sample) and **Table 3** (n = 126 sample).

Pre-treatment images in the n = 170 sample (paper I and III): The magnets used were 1.5 T (155 of 170 cases), 1.0 T (12 cases), and 3.0 T, 0.5 T, and 0.2 T magnets (1 case each). The number of examinations performed with 1.5 T magnets is corrected from 150 given in the Materials and methods section in paper I to 155 because more information became available in the meantime. MRI sequences advised for evaluating lumbar spine degeneration were applied [55, 56]. All examinations included sagittal

T2-weighted FSE images, all but two included sagittal T1-weighted images, and all but two included axial images of the L4/L5 and L5/S1 levels (**Table 2**). Slice thickness was 3-5 mm. Matrix varied from 160×256 to 640×640 but was typically 512×512 (**Table 2**). The images were obtained directly in Digital Imaging and Communications in Medicine (DICOM) format or, in seven cases, as digitized printed film hard copies stored in DICOM format. The examinations were de-identified before being evaluated.

Pre-treatment and 2-year follow-up images in the n = 126 sample (paper II): The magnetic field strength was 1.5 T in 235 of the 252 examinations of the 126 patients. The number of examinations performed with 1.5 T magnets is corrected from 230 given in the Materials and methods section in paper II to 235 because more information became available in the meantime. The magnets used in the 126 pre-treatment examinations were 1.5 T (116 of 126 cases), 1.0 T (7 cases), and 3.0 T, 0.5 T, and 0.2 T magnets (1 case each), and in the 126 2-year follow up examinations the magnets used were 1.5 T (119 of 126 cases), 3.0 T (6 cases) and 0.5 T (1 case). MRI sequences advised for evaluating lumbar spine degeneration were applied [55, 56], and relevant MRI parameters are given in **Table 3**. All but one examination included sagittal T1-weighted images and all but one examination (one at 2 year - but this examination included sagittal STIR) included sagittal T2-weighted images. Matrix was typically 512×512 (varied from 256×256 to 1024×1024) (**Table 3**). The images were obtained directly in DICOM format or, for six pre-treatment examinations, as digitized printed film hard copies stored in DICOM format, and were de-identified before being evaluated.

Table 2 Pre-treatment MRI parameters, n = 170 sample (paper I and III)

	Pre-treatment MRI (170 examinations)	**N**
Magnetic field strength	1.5 T	155
Range	0.2-3.0 T	170
Sagittal T1		168
Spin echo, TR/TE (ms)	350-911/ 7-22	159
FLAIR only, TR/TE (ms)	1984-2130/ 20-22	9
Sagittal T2		170
Fast spin echo, TR/TE (ms)	2511-4760/ 91-140	170
DRIVE only	DRIVE not used	0
Sagittal FS, TR/TE (ms) [a)]	4300-5070/ 56-101	5
Axial images [b)]		168
T1		33
T2		135
PD		21
Slice thickness/gap, range (mm) [c)]	3-5/0-2.2	
Matrix		
Sagittal images [d)]	512×512	115
Axial images	512×512	89
Field of view		
Sagittal images**,** range (cm)	19-38	
Axial images**,** range (cm)	15-32	

Abbreviations:

MRI = magnetic resonance imaging
n = number of examinations with the stated MRI parameter(s), range or sequence
T = tesla, indicating magnetic field strength of the MRI machine
T1 = T1-weighted MRI images
T2 = T2-weighted MRI images
mm = millimetre
cm =centimetre
TR/TE = repetition time in milliseconds (ms) / echo time in ms during image acquisition
FLAIR = Fluid Attenuated Inversion Recovery
DRIVE (Driven Equilibrium) = fast spin echo with 90° Flip-Back Pulse
FS = fat suppressed images such as STIR (Short Tau Inversion Recovery) or TIRM (turbo inversion recovery magnitude)

a) Included T2 TIRM with TR/TE: 4300-5070/70-101 (n=4) and T1 STIR with TR/TE: 4600/56 (n=1). All examinations with FS sequences had FSE T2 in addition.

b) Some examinations include more than one series of axial images, for example both T1- and T2-weighted axial images.

c) Slice thickness and gap concern both sagittal and axial images.

d) The number of examinations with the typical matrix used during acquisition of both T1- and T2-weighted images. If the stated matrix was used only during T1- or T2-weighted image acquisition (not both) the examination was not included in the number n.

Table 3 Pre-treatment and 2-year follow-up MRI parameters, n = 126 sample (paper II)

	Pre-treatment MRI (126 examinations)	**n**	**2-year follow-up MRI (126 examinations)**	**n**
Magnetic field strength	1.5T	116	1.5T	119
Range	0.2-3.0T	126	0.5-3.0T	126
Sagittal T1		125		126
Spin echo, TR/TE (ms)	360-911/ 7-20	118	375-724/ 7.4-20	126
FLAIR only, TR/TE (ms)	1984-2130/ 20-21.8	7	NA	0
Sagittal T2		125		122
Fast spin echo, TR/TE (ms)	2511-4760/ 91-140	125	2000-5070/ 70-140	110
DRIVE only, TR/TE (ms)	NA	0	700/ 135-140	12
Sagittal FS, TR/TE (ms) [a)]	4300-5070/56-101	5	2200/ 20	4
Axial images [b)]		124		122
T1		26		5
T2		93		120
PD		20		1
Slice thickness/gap, range (mm) [c)]	3-5/0-2.0		3-5/0-1.4	
Matrix				
Sagittal images [d)]	512×x512	86	512×X512	90
Axial images	512×x512	74	512×X512	86
Field of view				
Sagittal images, range (cm)	19-36		28-41	
Axial images, range (cm)	15-32		14-30	

Abbreviations:
MRI = magnetic resonance imaging
n = number of examinations with the stated MRI parameter(s), range or sequence
T = tesla, indicating magnetic field strength of the MRI machine
T1 = T1-weighted MRI images, T2 = T2-weighted MRI images
NA = not applicable
mm = millimetre, cm = centimetre
TR/TE = repetition time in milliseconds (ms) / echo time in ms during image acquisition
FLAIR = Fluid Attenuated Inversion Recovery
DRIVE = fast spin echo with 90° Flip-Back Pulse
FS = fat suppressed images such as STIR (Short Tau Inversion Recovery) or TIRM (turbo inversion recovery magnitude)

a) Pre-treatment FS images included T2 TIRM (n = 4, TR/TE 4300-5070/70-101) and T1 STIR (n = 1, TR/TE 4600/56), and all examinations that included FS sequences also included FSE T2 in addition. All 2-year follow-up FS images included T1 STIR and did not include T2 FSE or DRIVE sequences in addition.

b) Some examinations include more than one series of axial images, for example both T1- and T2-weighted axial images.

c) Slice thickness and gap concern both sagittal and axial images.

d) The number of examinations with the typical matrix used during acquisition of both T1- and T2-weighted images. If the stated matrix was used only during T1- or T2-weighted image acquisition (not both) the examination was not included in the number n.

- Observers and viewing equipment

One radiologist experienced in musculoskeletal MRI (observer A) and two neuroradiologists (observers B and C) from three different institutions participated in the rating of MRI findings and change in MRI findings over time. Each observer had more than 10 years' experience in reporting lumbar spine MRI findings. Observers A and C viewed the images on a clinical Picture Archiving and Communication System (PACS) unit. Observer B viewed the 170 pre-treatment examinations on a personal computer; he viewed the 2-year follow-up images and re-viewed the pre-treatment images in the n = 126 sample on a PACS unit. Observers A and B used the eFilm Lite image reading software version 2.1.2 (Merge Healthcare, Hartland, Wisconsin), while observer C used the Agfa Impax 4.5 (Agfa HealthCare, Mortsel, Belgia).

- Pilot study

To achieve a common understanding of the rating criteria, the three observers independently assessed six pilot examinations from another study. Observers A and B then discussed ratings and criteria at a joint two-hour meeting. Observer C did not attend the meeting but compared ratings with observers A and B and discussed with the main supervisor of this thesis, who had attended the meeting.

- MRI ratings

We used existing MRI rating criteria for Modic changes, posterior HIZ in the disc, nucleus pulposus signal, disc height (subjective and measured), disc contour, and FA:

Modic changes in the vertebral body marrow adjacent to the endplate are of type I (hypointense T1 signal and hyperintense T2 signal), type II (hyperintense T1 signal and iso- or hyperintense T2 signal), and type III (hypointense T1- and T2 signal) [28, 67, 84]. Their type, maximal anteroposterior (AP) extent (< 25 %, 25 – 50 %, or > 50 % of AP endplate diameter), and maximal craniocaudal (CC) extent (minimal/small dots, < 25 %, 25 – 50 %, or > 50 % of vertebral body height) were rated [80].

Posterior HIZ (graded as present or not present) was defined as an area of high-signal intensity in the posterior annulus fibrosus that is brighter than the nucleus pulposus on T2-weighted images and is surrounded superiorly, inferiorly, and anteriorly by the low-intensity signal of the annulus fibrosus [64, 65].

Nucleus pulposus signal (graded as bright, grey, dark, or black) was assessed on sagittal T2-weighted images [70].

Disc height reduction was graded subjectively as none (disc higher than normal disc above), slight (disc as high as normal disc above), moderate (disc narrower than normal disc above), or severe (endplates almost in contact) [71-73]. The grading was based on experience for the L5/S1 disc and when the disc above appeared abnormal [73].

Measured disc height decrease was calculated as < 40 % or ≥ 40 % of the nearest normal above disc height, based on the measured distance between the mid-inferior and the mid-superior disc borders on the mid-sagittal T2-weighted image [74].

Disc contour was graded as normal, bulge (base > 1/2 of disc circumference), or herniation (includes protrusion, extrusion, and sequestration) [67].

FA was graded for the worst side (right/left) as normal, mild (joint space narrowing or mild osteophyte), moderate (sclerosis or moderate osteophyte), or severe (marked osteophyte) based on the grading system of Fujiwara et al [77] and illustrations from the Spine Pain Outcomes Research Trial [65].

The observers received published illustrations of Modic changes, posterior HIZ, and FA [65]. They rated each variable at each of the disc levels L3/L4, L4/L5 and L5/S1, and rated Modic changes both inferiorly and superiorly to the disc. They were asked to also rate the variables on images of sub-optimal quality, since the images had been accepted on enrolment and reflected practice.

- Image evaluation

Evaluation of pre-treatment MRI images (Paper I and III): Blinded to clinical data and each other's ratings, all three observers evaluated the 170 pre-treatment MRI examinations in random order over three to four months. Blinded to and more than 3 months after their first rating two observers (A and B) re-rated 126 examinations (those of the n = 126 sample) in a new random order. These examinations were selected because the re-ratings were needed for comparison purposes in the 2-year follow-up studies of these patients [128], including the study in paper II. These 126 patients were similar to the rest (n = 44) of the 170 patients in gender (*P* = 0.938; chi-squared test) and ODI (*P* = 0.278; t-test, normal distribution) and were only slightly older (mean age 41.6 years vs. 38.9 years in the n = 44 group; *P* = 0.027; t-test, normal distribution).

Conclusive pre-treatment MRI findings in the n = 170 sample (paper III) were based on simple majority, median rating, or on a fourth radiologist's rating when observers A, B and C disagreed completely on type of Modic changes.

Evaluation of change in MRI findings at 2-year follow-up (paper II): Blinded to clinical data, observers A and B had independently interpreted the pre-treatment MRI images from the n = 126 sample over three to four months as part of their first rating of all 170 pre-treatment examinations. Blinded to and more than 3 months after this first rating, they rated the 2-year follow-up MRI images from the n = 126 sample in a new random order. Any difference between these two ratings signified a change in MRI finding by non-comparison of images. Still blinded to the pre-treatment ratings, the observers then directly compared the follow-up and the pre-treatment MRIs and reported any progress or regress of at least one grading category for each MRI finding by comparison of images. They rated Modic changes, posterior HIZ in the disc, nucleus pulposus signal, disc height, disc contour, and FA. They rated each finding at each of the disc levels L3/L4, L4/L5 and L5/S1, but rated only FA at levels with disc prosthesis.

3.3 Clinical measures (paper III)

At baseline, all patients reported that LBP had been their main symptom for at least a year. They filled in a set of questionnaires [42], including the validated version of the ODI 2.0 for pain-related disability with a score range of 0 (low) to 100 (high) [129, 130]. This instrument contains one pain-item and nine items regarding activities of daily living, and patients are asked to describe their current problem. The ODI score for pain and disability is calculated as a percentage ranging from 0 to 100 with higher scores indicating lower functional capacity. Mean ODI score in asymptomatic subjects is found to be 10.2 (range 2.2-12) [129]. The ODI is recommended as a back pain-specific measure of disability [131]. Patients also scored the maximum current LBP intensity in the last week on a visual analogue scale ranging from 0 (no pain) to 100 (worst pain imaginable) [132].

3.4 Statistical analysis

Reliability of pre-treatment MRI findings in the n = 170 sample (paper I): Prior to statistical analyses all MRI findings were dichotomized into categories that reflected the inclusion criteria or that might be clinically relevant (see Results). For each observer the prevalence of each type of dichotomised MRI finding was calculated at each rated level. Each finding was further analysed at each rated level, but only findings with a mean prevalence 10-90 % across all observers at the rated level were further analysed, since very high or low prevalence can lead to very low agreement beyond chance, despite very high actual agreement [65, 104, 117]. MRI indication for prosthesis (yes/no) was noted as present when the observer reported ≥ 40% disc height decrease and / or at least two of these three findings: Modic changes type I and/or II (superior and/or inferior to disc), posterior HIZ, and dark/black nucleus pulposus. These retrospective reports were not used in the prospective trial. The MRI indication for prosthesis was analysed separately at L4/L5 and L5/S1.

Using STATA 10.0 (College Station, TX), unweighted overall kappa was computed for agreement between all observers with a 95 % bias-corrected confidence interval based on bootstrapping with 1000 repetitions. Unweighted kappa for pairwise interobserver agreement and for intraobserver agreement was calculated using SPSS 17.0 (SPSS, Chicago, IL). *P* values were computed for difference in the prevalence of findings across observers (fixed effects model, STATA 10.0). After Bonferroni adjustment for multiple (i.e. 23) comparisons, *P* < 0.002 (*P* < 0.05/23) indicated statistical significance. Kappa was interpreted as: k ≤ 0.20: poor, 0.21- 0.40: fair, 0.41- 0.60: moderate, 0.61- 0.80: good and 0.81- 1.00: very good agreement beyond chance [109].

<u>Reliability of change in MRI findings over time in the n = 126 sample (paper II):</u> We first dichotomized the rating of change into progress versus unchanged/regress and into regress versus unchanged/progress. For Modic changes only alteration from none to any type of Modic changes was noted as progress, and the reverse was noted as regress. For disc contour change from bulge to herniation was noted as progress. For each observer the prevalence of progress and the prevalence of regress for each MRI finding at each rated level were calculated separately for comparison and non-comparison of images.

Prevalence- and bias-adjusted kappa (PABAK) was calculated for progress and regress for each finding at each rated level for each image evaluation method. PABAK was also calculated separately for locations at risk and locations not at risk of being exposed to artefacts from adjacent disc prosthesis. Adjacent locations at risk was defined as the facet joints at the prosthesis level(s) and all evaluated locations at the nearest level above and the nearest level below the prosthesis level(s). Despite high actual agreement ordinary kappa values may be very low when the prevalence of positive ratings (e.g., progress or regress) is low, and in this situation PABAK is particularly useful [104, 117, 133, 134]. PABAK values were returned without confidence intervals. Similarly to other investigators, we interpreted PABAK values as ordinary kappa values, i.e. as indicating poor (≤ 0.20), fair (0.21 – 0.40), moderate (0.41 – 0.60), good (0.61 – 0.80), or very good (0.81 – 1.00) agreement beyond chance [109, 133].

Generalized estimating equations (GEE) was used to analyse 1) the impact of any adjacent prosthesis and image evaluation method on PABAK, and 2) the impact of evaluation method and observer on the rating of change (progress, unchanged, regress). In 1) we used a linear model (scale) with adjacent prosthesis and evaluation method as factors and as main effects of the model, and in 2) we used an ordinal logistic model with evaluation method and observer as factors and as main effects of the model. The Wald chi-squared test was used as a test of model effects. After Bonferroni adjustment for multiple (i.e. 22) comparisons, $P < 0.0023$ ($P < 0.05/22$) indicated statistical significance. The 22 comparisons concerned impact of evaluation method and observer on prevalence of change for 9 variables (Modic presence, Modic AP and CC extent, HIZ, nucleus pulposus signal, subjective disc height reduction and measured disc height decrease, disc contour, and FA; 18 P values), and impact of adjacent disc prosthesis and image evaluation on PABAK for progress and regress for all 36 variables together (4 P values).

The data were analysed using SPSS 17.0 (SPSS, Chicago, IL) and WINPEPI 10.9 (http://www.brixtonhealth.com/pepi4windows.html).

MRI findings in relation to disability/LBP intensity (n = 170 sample, paper III): MRI total score was defined prior to analyses and based on the following combined pre-treatment MRI findings: presence of Modic changes (primary or secondary type I and/or II, larger than small dots), posterior HIZ, dark/black nucleus pulposus, and ≥ 40% disc height decrease at levels L4/L5 and L5/S1. At each of these two lumbar levels presence of Modic changes (if present at one or both endplates) was assigned 2 points and presence of each of the three other findings was assigned 1 point. Thus, 0 to 5 points could be assigned at each level and the MRI total score could range from 0 to 10 points. Modic changes were more closely related to the presence of LBP in prior studies and thus they were given more weight than the other MRI findings in the MRI total score [18, 47, 89]. Our *a priori* hypothesis was that a higher MRI total score was related to worse disability and pain.

Normality plots indicated that the ODI and LBP intensity scores as well as the MRI total score had approximate normal distribution (Figure 11, 12, and 13). We calculated Pearson correlation coefficient (r) for the MRI total score and for each dichotomized MRI finding relevant to our *a priori* aims (i.e. presence or not of Modic changes, HIZ, dark/black nucleus pulposus, and ≥ 40% disc height decrease) at each disc level, against ODI and LBP intensity. We used scatter plots and LOWESS

(locally weighted scatterplot smoothing) to check for non-linear trends. Simple and multiple regression analyses (General Linear Model) were carried out in two models with ODI and LBP intensity as dependent variables; in model A the MRI total score was a covariate and in model B the individual dichotomized MRI findings at each disc level were covariates. In multiple regression analyses, we adjusted for gender (male/female), age (continuous), BMI (continuous), smoking ("do you smoke?" yes/no), and anxiety/depression (no versus some/much anxiety and/or depression on EuroQol-5D: www.euroqol.org).

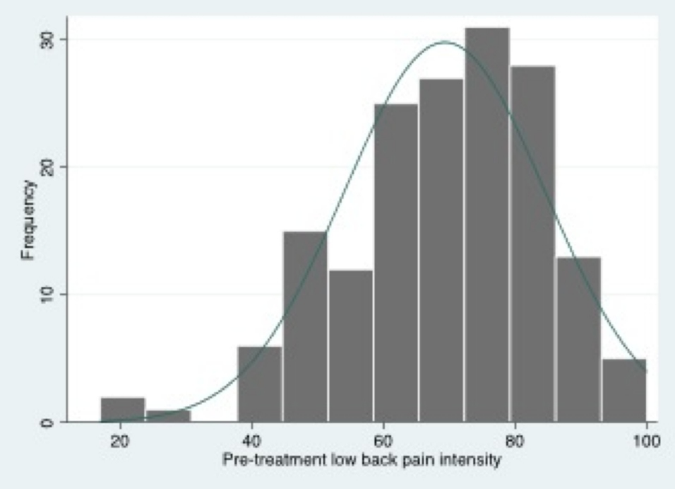


Figure 11 Frequency of pre-treatment LBP intensity scores (visual analogue scale 0-100, 0 indicates no pain and 100 indicates worst pain) in 170 disc prosthesis candidates. LBP intensity has approximate normal distribution (the line indicates normal distribution).

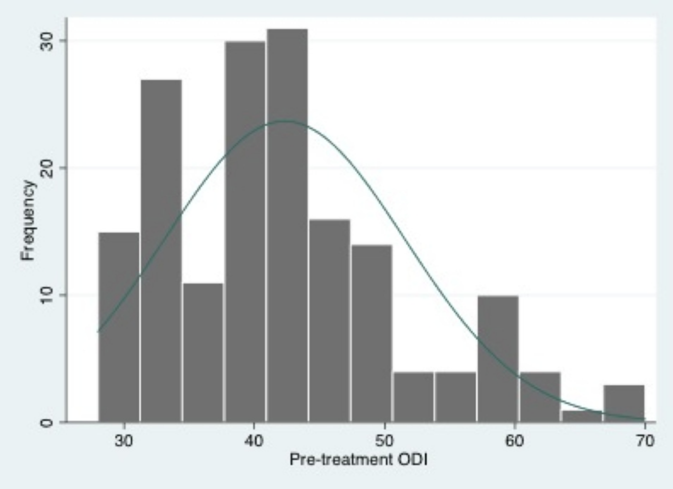


Figure 12 Frequency of pre-treatment disability scores (Oswestry disability score, scale 0-100, 0 denotes best and 100 denotes worst disability) in 170 disc prosthesis candidates. The disability score has approximate normal distribution (the line indicates normal distribution).

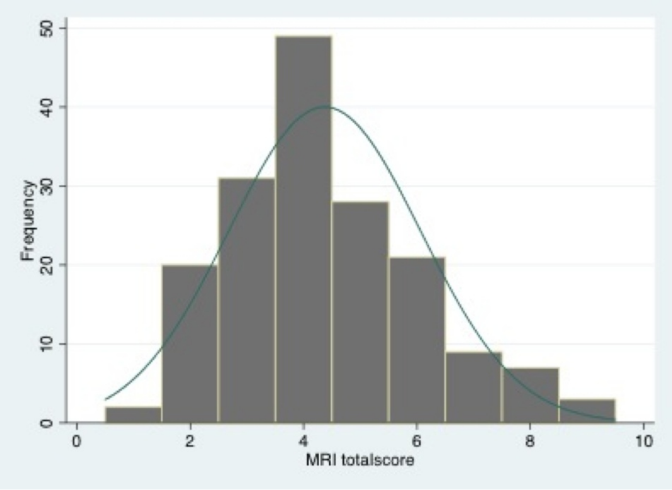


Figure 13 Frequency of pre-treatment MRI total scores in the n = 170 sample.
The MRI total score has approximate normal distribution (the line indicates normal distribution).

Post hoc the following analyses were carried out to assess the robustness of the results, in response to comments from reviewers on an early version of paper III:
a) the 1.5 T MRI sub-group was analysed separately,
b) we assigned 1 point (not 2 points) to Modic changes in the MRI total score,
c) we adjusted also for the two following relevant variables with less complete data: physical work load (0-10; 0 = no heavy load, 10 = very heavy load) and physical leisure-time activity (1-4; 1 = sedentary, 4 = hard training) [135], and
d) we added moderate/severe FA at L4/L5 and L5/S1 in the MRI total score (1 point per level) and as a covariate.
In the *post-hoc* analyses (a-d), we repeated all relevant correlation-, scatter plot/LOWESS-, and regression analyses.

The residuals in the regression models were checked for normal distribution. The collinearity of independent variables was examined by 1) comparing unadjusted and adjusted regression coefficients (β) for each variable and by 2) performing bivariate correlation analyses of all independent variables with each other.

Data were analysed using STATA (StataCorp, 2011, Stata statistical software: Release 12.0, Stata Corporation, College Station, TX, USA). $P \leq 0.05$ indicated statistical significance.

<u>Sample size:</u> The sample size was fixed and was based on power calculations for the main trial. In paper I, for each comparison, if the true kappa is 0.60 and the prevalence 30 %, 191 paired observations provide 80 % power to give a significant result at the 5 % level in a two-sided test of k = 0.40 [104]. Three observers were used in order to improve the power in this study with a fixed sample size n = 170 [104]. In paper III, if the true absolute correlation is 0.25 and the sample size is 164 the chance is 90 % to get a significant correlation at the 5 % level [136]. Therefore, our fixed sample size of 165-170 allowed detection of fairly weak correlations.

4.0 Results

4.1 Reliability of pre-treatment MRI findings (paper I)

All observers rated all findings at L3-S1 in all 170 pre-treatment examinations except for type of any Modic changes in the two examinations lacking T1-weighted images. Observers A and B rated all findings twice in 126 cases for intraobserver analysis. Due to a mean prevalence across observers of < 10 % in the n = 170 sample, we did not further analyse any finding at L3/L4 or FA at L5/S1.

Differences in prevalence of findings across observers: The prevalence at each rated level differed significantly ($P < 0.002$) but slightly across observers for most findings (Table 2 in paper I). Prevalence differed up to twofold between observers for presence of Modic changes and up to threefold for posterior HIZ and for disc height judged severely reduced. The prevalence at each rated level differed less between observers for extent of Modic changes (except for > 50 % CC extent at L5/S1 inferior to disc), dark/black nucleus pulposus signal, ≥ 40 % disc height decrease, and abnormal disc contour. The reported prevalence of FA (at L4/L5) ranged from 5.9 % to 14.1 %; it did not differ significantly across observers after Bonferroni adjustment for multiple comparisons ($P > 0.002$).

The difference in prevalence across observers took a different direction for different findings (Table 2 in paper I). Thus, the overall MRI indication for prosthesis did not differ significantly in prevalence across observers, neither at disc level L4/L5 nor at disc level L5/S1 (Table 2 in paper I).

A further analysis (not included in paper I) showed no significant difference in prevalence between each observer's first and second interpretations (n= 126) for any individual finding ($P > 0.003$) or for the MRI indication for prosthesis ($P > 0.267$), except in one observer for presence of abnormal disc contour at L5/S1 (prevalence 64.3 % versus 74.6 %, $P < 0.001$). In this analysis with multiple comparisons (as in paper I), $P < 0.002$ indicated statistical significance.

Interobserver agreement: Overall agreement was moderate or good (k = 0.56-0.77) for presence and extent of Modic changes, except inferior CC extent at L5/S1 (k = 0.40, mean prevalence across observers 14.7 %) (Table 3 in paper I). Overall agreement was moderate (k = 0.46-0.58) for HIZ and moderate or good (k = 0.50-0.72) for dark/black nucleus pulposus signal, severely reduced disc height, ≥ 40 % measured disc height decrease, and abnormal disc contour (Table 3 in paper I). Moderate/severe FA at L4/L5 had a mean prevalence across observers of 11.4 % and showed fair overall agreement based on kappa (k = 0.24); PABAK values (not included in paper I) suggested good interobserver agreement for FA (0.74 at L4/L5 and 0.66 at L5/S1).

The MRI indication for disc prosthesis showed good overall agreement both at L4/L5 (k = 0.70) and at L5/S1 (k = 0.66).

Pairwise agreement was fair in one pair at L5/S1 for inferior AP and CC extent of Modic changes, superior AP extent, posterior HIZ and disc contour, and in all pairs for FA at L4/L5. It was otherwise moderate to very good (Table 3 in paper I).

Intraobserver agreement: Intraobserver agreement on individual findings was good or very good (k = 0.61-1.00) except in one observer at L5/S1 for inferior AP and CC extent of Modic changes (k = 0.38-0.55) and for HIZ (k = 0.60) (Table 4 in paper I).

Intraobserver agreement on the MRI indication for prosthesis was also good or very good (k = 0.67-0.87) (Table 4 in paper I).

4.2 Reliability of change in MRI findings over time (paper II)

Both observers (A and B) rated changes in FA at L3/L4, L4/L5 and L5/S1 in all 126 patients. Both rated changes in the other MRI findings at 125−126 of 126 L3/L4 levels, 87−88 of 89 L4/L5 levels and 72−73 of 74 L5/S1 levels without prosthesis according to at least one observer. Missing data in one to two cases were due to prosthesis artefacts or disagreement on level.

Reliability of change by comparison of images: By comparison with the initial images, 0.0–29.4 % of follow-up images showed progress and 0.0–14.9 % showed regress, depending on finding/level and observer (Table 1 in paper II; the 29.4 % value is corrected from the 20.5 % value given in the text of paper I). The interobserver agreement on *progress* by comparison of images (Table 2 in paper II) was good or very good (PABAK 0.63–1.00) for Modic changes, HIZ, disc height, and disc contour and for nucleus pulposus signal and FA at L3/L4; and moderate (PABAK 0.46–0.59) for nucleus pulposus signal and FA at L4/L5 and L5/S1. The agreement on *regress* by comparison of images (Table 3 in paper II) was good or very good (PABAK 0.70–1.00) for all findings.

Impact of adjacent disc prosthesis: An adjacent prosthesis did not influence interobserver agreement on progress or regress across all 36 variables ($P \geq 0.22$, adjusted for image evaluation method). In locations adjacent to a prosthesis, image comparison showed only fair agreement (PABAK 0.29) for increasing FA at L5/S1 and otherwise better agreement (PABAK 0.44-1.00) (Table 2 and Table 3 in paper II). By comparison of images, the groups with and without adjacent prosthesis had equal mean PABAK (0.83 for progress and 0.89 for regress).

Non-comparison versus comparison of images: Image evaluation method significantly influenced interobserver agreement across all 36 variables for both progress and regress ($P < 0.001$). PABAK was higher by comparison vs. non-comparison for 29 (progress) and 24 (regress) of the 36 variables (Tables 2 and 3 in paper II). Mean PABAK for progress/regress was 0.82/0.89 by comparison vs. 0.74/0.82 by non-comparison.

Progress was more prevalent by non-comparison vs. comparison of images for most variables (observer A, 33 of 36 variables; observer B, 29 of 36 variables); the same applied to regress (observer A, 20 of 36 variables; observer B, 22 of 36 variables) (Table 1 in paper II). Image evaluation method significantly affected the rating of change in AP extent of Modic changes (*P* < 0.001). Observer had an impact on the rating of change for FA (*P* < 0.001). Image evaluation method tended to influence the rating of change in whether Modic changes were present or not (P= 0.003), but there were no other significant influences of image evaluation method or observer on the rating of change in MRI findings (P ≥ 0.116).

4.3 MRI findings in relation to degree of disability and pain (paper III)

Table 4 shows clinical data and **Table 5** shows the conclusive pre-treatment MRI findings. The 170 patients (mean age 41 years; 82 men, 88 women) had a mean (range) pre-treatment MRI total score of 4.4 (1-9) points: 1.6 (0-5) points from L4/L5 and 2.8 (0-5) points from L5/S1. The mean ODI score was 42.3 (standard deviation (SD) 9.3, n = 170), mean LBP intensity was 69.3 (SD 15.3, n = 165), and mean BMI was 27.0 (SD 13.3, n = 168); 46 % of the patients were smokers (n = 169); 58 % reported some (51 %) or much (7 %) anxiety/depression (n = 169).

Table 4 Pre-treatment characteristics, n = 170 sample (paper I and III)

Variables used in analyses for paper III		**n with data**
Age, *mean (SD)*	41.0 (7.1)	170
Gender women, *n (%)*	88 (51.8)	170
ODI score, 0-100, *mean (SD)*	42.3 (9.3)	170
Low back pain intensity, 0-100, *mean (SD)*	69.3 (15.3)	165
Body mass index, *mean (SD)*	27.0 (13.3)	168
Current smoker, *n (%)*	78 (46.1)	169
Anxiety/depression (EuroQol-5D), *n (%)*		169
- No	71 (42.0)	
- Some	87 (51.5)	
- Much	11 (6.5)	
Physical work load, 0-10, *mean (SD)*	5.4 (3.1)	138
Physical leisure time activity, *n (%)*		149
- sedentary	32 (21.5)	
- light activity (e.g. walking or biking) ≥ 4 hours/week	109 (73.5)	
- moderate activity, recreational sport ≥ 4 hours/week	8 (5.4)	
- hard training	0 (0.0)	

Abbreviations: SD = standard deviation, ODI = Oswestry Disability Score, MRI = magnetic resonance imaging; MRI total score was based on a sum of MRI findings (primary or secondary type I and/or II Modic changes larger than small dots, posterior HIZ, dark/black nucleus pulposus, and ≥ 40% disc height decrease) at levels L4/L5 and L5/S1. At each level, presence of Modic changes superior and/or inferior to disc gave 2 points and the other findings gave 1 point each.

Table 5 Prevalence of pre-treatment MRI findings, n = 170 sample (based on the conclusive findings used in paper III)

MRI finding	**Prevalence, %**
Modic type I and/or II more than small dots at superior and/or inferior endplate	
- L4/L5	32.3
- L5/S1	68.2
Posterior high-intensity zone (HIZ) present	
- L4/L5	25.9
- L5/S1	18.8
Nucleus pulposus signal dark or black	
- L4/L5	51.2
- L5/S1	66.5
Disc height decrease ≥ 40 %	
- L4/L5	14.1
- L5/S1	59.4
Facet arthropathy (FA), moderate or severe	
- L4/L5	5.3
- L5/S1	4.1
MRI total score, *mean (range)*	4.4 (1-9)
Score contribution from each level	
- L4/L5*, mean (range)*	1.6 (0-5)
- L5/S1*, mean (range)*	2.8 (0-5)

Primary analyses for the *a priori* aims: Neither MRI total score (Fig. 2A-B in paper III) nor any individual MRI finding was significantly correlated with ODI or LBP intensity (r ranged from -0.15 to 0.11). Simple linear regression showed similar results (Table 2 in paper III). Scatter plots and LOWESS (locally weighted scatterplot smoothing) indicated no extra-linear relationship between MRI total score and ODI or LBP intensity.

Similarly, in multiple linear regression analyses (Table 3), neither MRI total score (model A) nor any individual MRI finding (model B) was related to ODI or LBP intensity. The only exception was a weak negative relationship between a posterior HIZ at L5/S1 and ODI (β = -4.7, *P* = 0.02, model B). Anxiety/depression was related to a slightly more intense LBP (model A: β = 5.7, *P* = 0.02; model B: β = 6.6, *P* =

0.01). No other independent variable (age, gender, BMI, smoking, or MRI variable) was significant in any of the regression models in the primary analyses.

Post-hoc additional analyses: The MRI total score was still not related to ODI or LBP intensity (model A: $P \geq 0.31$) when we:
a) analysed the 1.5 T sub-group separately (n = 146-151) *or*
b) assigned 1 point and not 2 points to Modic changes *or*
c) adjusted also for physical work load and physical leisure-time activity (n = 124-128) *or*
d) added moderate/severe FA in the MRI total score (whether we assigned 1 or 2 points to Modic changes, or adjusted or not for physical load/activity). At L4/L5 and L5/S1, only 5.3 % and 4.1 %, respectively, of the 170 patients had conclusive pre-treatment findings of moderate/severe FA (**Table 5**).

Regarding individual MRI findings, HIZ at L5/S1 remained weakly and negatively related to ODI (model B) in the 1.5 T sub-group (ß = -4.1, $P = 0.05$, n = 151), and when we adjusted also for physical workload and physical leisure-time activity (ß = -5.6, $P = 0.02$, n = 128). After this adjustment, women had 4.7 point higher ODI score than men (model B: ß = 4.7, $P = 0.01$). Anxiety/depression was related to a slightly higher LBP intensity also in the 1.5 T sub-group (model A: $\beta = 5.9$, $P = 0.03$; model B: $\beta = 7.1$, $P = 0.01$). No other independent variable (such as FA or physical load/activity) was significant in any of the *post-hoc* multiple regression models ($P > 0.05$).

Evaluation of collinearity: The independent variables showed no clear collinearity that needed to be accounted for in the analyses. The β for each variable changed little from unadjusted to adjusted regression (Tables 2 and 3 in paper III). Also, bivariate correlation analyses of all independent variables revealed only a few weak correlations (r ranged from -0.32 to 0.51).

5.0 Discussion

5.1 Methodological considerations

- Study design and patients

This PhD project includes two reliability studies and one cross sectional study on associations in a case series of patients with long-lasting chronic LBP, localized degenerative lumbar MRI findings, and ODI ≥ 30%. The patients were not drawn from the general population, from LBP patients in general or from chronic LBP patients in general, and accordingly extrapolation of the results to these populations was not possible. However, the sample was representative of patients accepted for surgery with lumbar disc prosthesis. In this sample, we analysed associations between MRI score and ODI or LBP score. A case-control study or a cohort study would have been necessary to analyse the sequence of events and establish a causal association, but an association found in a cross sectional study may serve as a first step to establish hypotheses for cohort studies or a case-control studies. A cross sectional study is usually easier and cheaper to conduct, and may provide advantages such as reduced recall bias and no loss to follow up [109].

In general the choice of sample may have a strong influence on the association between variables [18]. If associations between MRI score and ODI or LBP scores exist, we would expect that the associations may be easier to find in patients with well defined, localized MRI findings at one or two levels than in patients with widespread degenerative MRI findings at different stages throughout the entire lumbar spine.

One of the strengths was that included patients were referred from clinical practice throughout Norway without any special recruitment attempt. These patients were representative not only of patients with chronic LBP accepted for disc prosthesis surgery but also of similar patients accepted for surgery with lumbar fusion, since the indications for the two surgical treatments are similar [42, 137]. However, the included patients were not representative of patients initially considered (but not necessarily accepted) for surgery.

A further strength was the large patient sample (n = 170) in paper I compared to prior reliability studies (**Table 1 in appendix**). In addition, the inclusion of both operated and non-operated patients in paper II allowed us to study how prosthesis artefacts might influence reliability. The fixed sample size of n = 170 (based on power calculations for the main trial) was also sufficient for studying relationships in paper III.

Patient age might influence results for observer agreement on lumbar MRI findings, for several reasons. First, some findings may be more prevalent in older patients [66] and prevalence affects kappa values [104]. Second, having more findings to evaluate may increase observer fatigue [105, 118] and thus reduce observer agreement. Third, the appearance of the evaluated structures such as the vertebral body marrow may vary with age [84] and this might affect the difficulty and reliability of the evaluations.

The 126 patients used for intraobserver reliability analysis in paper I were on average 2.7 years older than the rest (n = 44) of the 170 patients used for reliability analysis. This age difference hardly had any important impact on agreement. The difference was small and MRI findings that were inclusion criteria differed little in prevalence with age. For example, the mean prevalence of Modic changes across the three observers in the n = 126 sample versus the rest (n = 44) was 35.2% vs. 34.9% at L4/L5 superior to disc, 38.6% vs. 39.4% at L4/L5 inferior to disc, 73.8% vs. 75.8% at L5/S1 superior to disc, and 69.9% vs. 71.2% at L5/S1 inferior to disc.

- Image evaluation: observers, equipment and approach

An additional strength was that the observers were three radiologists from three different institutions who were experienced in reporting lumbar spine MRI and were not trained together, except in a small pilot study. I consider these observers representative of the neuro- and musculoskeletal radiologists who interpret such MRI examinations pre-operatively in clinical practice.

A limitation of the study was that the observers knew the patients were accepted for disc prosthesis surgery due to localized degeneration. How this may have affected their MRI ratings and agreement is not clear.

Observers A and C rated all images on a PACS unit whereas observer B used a high quality personal computer with dedicated software for the first rating of the pre-treatment examinations. I cannot rule out that this may have had a slight impact on the ratings. The screen resolution may be higher on a PACS unit than on a personal computer, but the importance of this difference is uncertain.

Different from some other researches [69, 138], we did not attempt to limit observer fatigue by restricting the number of images rated each day. Highly trained experts who work together and rate a few findings on a limited number of standardized images per day may achieve better agreement than reported in our study. As discussed by Peterson et al the rating of both normal disc levels and all grades of many degenerative findings may increase observer fatigue and reduce the observer's confidence and reliability compared to the rating of one or a few definite abnormalities at a single level [118]. However, the evaluation of a complexity of findings and grades of pathology at different levels is more similar to daily clinical practice [139]. Nevertheless, any study with structured ratings may overestimate the reliability in day-to-day work.

One strength of the image evaluation approach was the blinding of the observers to clinical data, each other's ratings and their own prior ratings. Another strength was the long time lag (> 3months) and altered random examination order between the first and the second rating to reduce recall bias [104, 140, 141]. A further strength was that change in MRI findings was rated both by comparing initial and follow-up images and by not comparing them. This was important, since both approaches are used in research [80, 83, 121, 123].

In an experimental setting readers perhaps interpret examinations with more care because they know that their performance will be measured; this fact may also result in more doubt in rating of findings than in ordinary clinical work [141]. After rating many images consecutively observers may tune themselves to a "standard" they can use for difficult cases and/or when they are in doubt. This standard develops during the experiment and accordingly the rating may change slightly from the beginning to the end of the experiment. We minimized any bias due to such change in rating by presenting the images in a random order (not consecutively).

The second rating of the pre-treatment images used in the intraobserver reliability analysis in paper I was the rating made with follow-up images available (for use in paper II). Thus, in paper I, the setting of the second rating differed slightly from that of the first rating, where only pre-treatment images were available. This may have caused some variation in the second rating that was not due to the observer. Thus, the true intraobserver reliability may be slightly better than we reported.

- Pilot study

To pilot the rating approach without rating any main study images (which could cause bias), all observers independently rated 6 examinations from another project in the pilot study. Ratings that observers A and B disagreed on were discussed at the pilot study meeting. Observer C did not attend the meeting (had not yet rated the pilot images), but later compared ratings with observers A and B and discussed with the last author of paper I, who had attended. Whether a short meeting like this can influence agreement is uncertain, but the mean kappa for interobserver agreement across all variables (Table 3 in paper I) was higher for observers A and B (0.66) than for observers C and A (0.63) and for observers C and B (0.54). A pilot study to agree on the protocol and interpretation of definitions and criteria is quite common in research. Many such pilot studies include more patients (sometimes from the main study) than our pilot study, but may not include independent image interpretations; to achieve a consensus the observers often both interpret and discuss the pilot images in common [65, 71, 73, 80].

- MRI images

The MRI images in our study reflected clinical practice and had been accepted on enrolment. It therefore strengthened the study that the observers were asked to also rate findings on images of sub-optimal quality, and actually rated all cases. Jarvik and Deyo expressed concern about the fact that 7.5 % of imaging cases were not evaluable in the reliability study by Carrino et al [65] despite "deemed eligible for a randomized clinical trial in which cross-sectional imaging findings were part of the inclusion criteria" [102]. We found similar agreement (paper I) as reported by Carrino et al [65], even though we did not exclude any cases. In retrospect, it would have been interesting to analyse whether the image quality influenced the reliability, but we did not formally assess image quality.

Park et al have reported mean 9.9 % diurnal variation in measured total disc height and mean 20 % measured reduction in signal intensity in the anterior part of the disc from morning to afternoon at L1-S1, but they also reported that the blinded observers found no changes judged subjectively [142]. To prevent bias from such diurnal variation, some researchers, different from us, standardized the time of the day when MRI was performed and/or let the patient rest in supine position for 30-45 minutes prior to the examination [85, 87, 94]. We cannot rule out that diurnal variation may explain a few changes in MRI findings in paper II, or may have led to a slight misclassification of some MRI findings and thus a slight underestimation of their association to disability/LBP in paper III. However, diurnal variation was unlikely to cause important bias.

The MRI technique (pulse sequences and slice thickness) was generally consistent with guidelines for performing MRI of the adult spine [55]. Only guidelines and no absolute rules exist for MRI techniques, and there is room for considerable variation in MRI technique within the guidelines. Accordingly variation will exist both in how

MRI examinations should be done based on the guidelines (e.g. depending on the available MRI equipment) and in how the MRI examinations actually are done (e.g. due to habits or available time). The varied MRI scanning may have caused some unwanted variation in the rating of MRI findings. Therefore, some aspects of the MRI technique will be discussed in the following.

Parameters affecting image quality. The most important parameters influencing MRI image quality are signal to noise ratio (SNR), contrast to noise ratio (CNR), spatial resolution, scan time and artefacts. In an ideal situation SNR and CNR is high, spatial resolution good, scan time short and there are no artefacts, but in real life these parameters interact, and optimizing one of them has a negative effect on one or more of the others [143]. The interactions between these parameters are quite complex. SNR is affected by field of view (FOV), matrix, slice thickness, number of excitations (NEX), receiver bandwidth, TR and TE. CNR is influenced by TR, TE, T1, T2, proton density (PD), inversion time, flip angle, and flow and turbo factor in FSE sequences. Spatial resolution is influenced by FOV, matrix and slice thickness. Scan time is affected by TR, phase encodings, NEX and slice number in volume imaging [143].

Magnetic field strength. The magnetic field strength varied from 0.2 T to 3.0 T (**Tables 2 and 3, section 3.2**). However, the use of 1.5 T MRI in most pre-treatment (91.2 %, 155/170) and 2-year follow-up examinations (94.4 %, 119/126) gave better image quality than the low magnetic field MRI used in many prior studies [27]. Increasing magnetic field strength is almost linearly correlated to increasing SNR, but the SNR for each individual image depends also on many other factors, as previously mentioned. The general increase in SNR obtained by increasing the magnetic field strength can be partly sacrificed to increase spatial resolution and/or reduce scan time [144]. Quantitative and qualitative image analyses have shown that diagnostic images quality can be obtained at 0.5 T, 1.0 T and 1.5 T, but the quality was higher at 1.0 T and 1.5 T than at 0.5 T; SNR and CNR were highest at 1.5 T and lowest at 0.5 T [145]. Higher magnetic field strength increases the risk of artificial thickening of endplates/cortical bones due to chemical shift artefacts, but this is unlikely to be misinterpreted as pathology [146]. In summary, higher field strength improves the SNR, CNR and/or spatial resolution [144, 145].

In paper II, the magnetic field strength differed between pre-treatment and follow-up MRI in 17 of 126 cases. It changed from 1.0 T at pre-treatment to 1.5 T at 2-year follow-up in 7 cases; from 1.5 T to 3.0 T in 6 cases; and from 3.0 T to 1.5 T, 0.2 T to 1.5 T, 0.5 T to 1.5 T, and 1.5 T to 3.0 T in 1 case each. This may have influenced some of the ratings of change in MRI findings over time [147]. In paper III (relationship of MRI findings to disability/LBP scores), the use also of other magnetic field strengths than 1.5 T did not affect our main results, since these were unchanged when analysing the 1.5 T sub-group. This is important, since the appearance of Modic changes has been shown to differ between magnetic field strengths (0.2 T vs. 1.5 T) [148].

TR and TE. TR determines the amount of T1 and proton density weighting of the images. TR influences the CNR on T1-weighted images, which depends on differences in longitudinal magnetic relaxation times (T1). TE controls the T2 weighting of the images, and influences the CNR on T2-weighted images, which depends on differences in transverse magnetic relaxation times (T2). In SE sequences, T1-weighted images typically have short TR (250-700 ms) and short TE (10-25 ms), and T2-weighted images long TR (> 2,000 ms) and long TE (> 60 ms) [149]. Typically FSE T2-weighted images have long TR (> 4000 ms) and long TE (about 100 ms) [143]. At higher magnetic field strengths T1 increases and T2 decreases, and it is recommended to increase TR for T1-weigthed images and reduce TE and TR for T2-weighted images although scan time increases [144].

T1 weighted SE sequences. In our study most T1-weighted sagittal SE sequences were obtained with the advised TR of 250-700 ms, but TR > 700 ms was used in 4 pre-treatment examinations (TR 750-911 ms) and 3 follow-up examinations (TR 705-

724 ms). Longer TR means less T1 weighting, but even a TR of 911 ms is quite short and it is difficult to see how this could have influenced the results.

T1 weighted FLAIR sequences. In 9 of 170 pre-treatment examinations (but no follow-up examinations), the only available sagittal T1 weighted sequence was T1 FLAIR (fluid attenuated inversion recovery or long tau inversion recovery). This may have had a slight impact on the rating of MRI findings. FLAIR has the advantage of shorter scanning time and T1 weighting with suppressed CSF signal and accordingly better CNR than T1 FSE (T1 FSE has poorer CNR than T1 SE used in our study though) [55]. Compared to T1 SE sequences (used in our study), T1 FLAIR may provide higher spatial resolution, better delineation of fluid from nerve roots, and improved ability to depict oedema and metastatic lesions in the fatty bone marrow [150, 151]. FLAIR may be less sensitive to very small fat concentrations, but may have similar sensitivity to Modic type II as T1 SE sequences [152].

T2 weighted FSE sequences. The sagittal T2 FSE sequences in our study had the advised long TR (often > 4000 ms) and long TE (about 100 ms) [143]. Compared to T2 SE, T2 FSE provides shorter scanning time (and thus less motion artefacts) and better SNR, but poorer contrast between fluid and fat and lower signal from a normal disc [69]. The higher fat signal on T2 weighted FSE images compared SE images may blur slight oedema such as in Modic type I changes [148], but T2 weighted SE images are seldom used in clinical practice today.

T2 weighted DRIVE sequences. When evaluating sagittal pre-treatment images (papers I and III), we did not use T2 sequences with a short TR (Restore / Fast Recovery / DRIVE) that could have obscured a HIZ and changed the signal intensity of the nucleus pulposus. However, in 12 of 126 follow-up examinations, the only available sagittal T2-weighted sequence was DRIVE. This is a FSE sequence with 90-degree flip back pulse that converts residual transverse magnetization to longitudinal magnetization, compensating for loss of water/CSF signal when TR is shortened to reduce scan time. Compared with conventional T2 FSE, T2 DRIVE provides reduced scan time and higher CSF signal with lower TR, but reduced ability to visualise intrinsic cord lesions (to compensate for this, TR has to be > 2000 ms) [153].

It is not clear how the use of DRIVE at follow up but not at pre-treatment in these few cases (12/126) may have affected the evaluation of change in MRI findings (paper II). To the best of my knowledge, no study has been published on the use of DRIVE in lumbar spine imaging, but there are a few other studies on its use [154, 155]. In a study of the cervical spine, mean SNR for disc and bone marrow signal decreased slightly after application of Driven Equilibrium to the 3D FSE sequence (TR/TE= 211/60 ms) [155]. In our study, the DRIVE sequences had TR/TE = 700/135-140 ms (longer TR and TE than in the referred studies), but no study on visualisation of degenerative MRI findings in the spine was found for these TR and TE values. Annulus fibrosus has a very short T2 (< 1 ms) and nucleus pulposus a long T2, about 100 ms [156]. Water and CSF has a long T2 about 200-250 ms [143]. Based on T1 and T2 for these different tissues, it seems unlikely that the moderate T1 weighting (TR 700) and good T2 weighting (TE 135-140) of the DRIVE sequences used in our study have influenced the rating of Modic changes, HIZ, nucleus pulposus signal, disc contour, or disc height. Accordingly, it is also unlikely that this difference in pulse sequence between pre-treatment and follow-up examinations in 12 cases has influenced reliability of change in MRI findings over time (paper II). I have reviewed the three examinations that included both sagittal T2 DRIVE and sagittal T2 FSE, and the MRI findings appeared to be similar on both sequences.

Fat suppressed sequences. Only 5 of 170 pre-treatment examinations and 4 of 126 follow-up examinations included fat-suppressed/water-sensitive sequences, which might help to distinguish Modic type I vs. II. This probably had limited impact on our results, since Modic type I and type II changes were combined as inclusion criteria and in the analyses. Furthermore, Carrino et al found no difference in reliability for type of Modic changes between examinations with vs. without fat-suppressed sagittal T2 weighted images [65].

Axial images. The angle of the axial slices in relation to the intervertebral disc and facet joints may vary between examinations. We did not record this angle. It may have differed between pre-treatment and follow-up examinations in some cases. This may have reduced the reliability when rating changes in FA (paper II), since this rating relies mostly on the axial slices and changed angle of these slices may affect the appearance of the facet joints. Changed angulation of the axial slices may have affected the ratings of HIZ and disc contour as well to some degree.

Slice thickness and gap: The American College of Radiology guidelines recommend sagittal and axial slice thickness ≤ 4 mm and slice gap ≤ 1 mm [55]. Carrino and Morrison recommended slice thickness 3-5 mm and gap 0.3-1 mm in the sagittal plane, and slice thickness 3-4 mm and gap 0-1 mm in the axial plane (slices parallel to the intervertebral discs or as a stack through the lower spinal canal) [56].

Partial volume effect means that the signal obtained is an average of two or more tissues, and this effect results in reduction or loss of contrast between two adjacent tissues. All MRI images suffer from partial volume effect in various degrees due to insufficient spatial resolution. To compensate for this we can use thinner slices, because when the slice thickness is the same or thinner than the lesion that we want to depict, the lesion is entirely contained within the slice and only that lesion’s signal is displayed. The disadvantage of thinner slices is that SNR is reduced [143]. In order not to overlook small lesions in-between slices it is important that the slice gap is small enough, but to avoid cross excitation and hence changed image contrast it is important that the slice gap is not too small (≥ 30% of the slice thickness) [143]. The recommended slice thickness of 3-4 mm and slice gap of 0-1 mm reflect a balance between these advantages and disadvantages (partial volume effect, SNR, artefacts, and chance of overlooking small lesions).

Of the 170 pre-treatment examinations, only 1 had sagittal and 3 had axial slice thickness > 4 mm, and 7 had sagittal and 10 had axial slice gap > 1 mm. Of the 126 follow-up examinations, 0 had sagittal and 30 had axial slice thickness > 4 mm, and 1 had sagittal and 5 had axial slice gap > 1 mm. Aprill recommended slice thickness ≤ 5 mm in order to visualise HIZ [64]. Few MRI images had larger than recommended slice thickness or slice gap, and maximum slice thickness was 5 mm and slice gap 2.2 mm (**Tables 2 and 3, section 3.2**). Overall, it is unlikely that these variations in slice thickness and slice gap have had important impact on the rating of MRI findings in our studies,

MRI technique at pre-treatment vs. follow-up: In the n = 126 sample, pre-treatment and 2-year follow-up examinations had similar slice thickness, slice gap and matrix but differed in numbers with digitized printed film hard copies (7 vs. 0), T1 FLAIR (12 vs. 0), and T2 DRIVE as the only sagittal T2-weighted sequence (0 vs. 12). These differences reflected clinical practice, but may have influenced the observers' rating in various ways and consequently may have reduced reliability for change in MRI findings (**Table 3, section 3.2**).

In summary, any change in MRI technique may influence image quality and accordingly reliability of MRI findings, but SNR is probably the most important image quality factor [143]. SNR is influenced by nearly all other parameters and in clinical practice a balance between scan time, spatial resolution, SNR, CNR, anatomic area, and MRI equipment is necessary. The images in our study reflect the practice at many different radiology centres in Norway and represent the images clinicians actually use for diagnosis and treatment decisions.

- MRI ratings

We studied findings used as MRI indication for surgery because they directly affect treatment choice and are supposed to be relevant to symptoms. In addition, we studied change in degenerative MRI findings over time since such change is often evaluated after surgery (e.g. to assess adjacent level degeneration) [45, 78, 113, 128].

Posterior HIZ: The original definition of HIZ by Aprill and Bugdok used in our studies has been widely used in other studies [65, 68, 73, 116, 117, 157, 158]. This definition may have included circumferential, rim and radial fissures in the posterior outer annulus fibrosus. These three types of fissures may have different causes [157] (compressive and shearing stress in older discs, trauma, and disc degeneration with bending and compression, respectively [58]) and accordingly different associations with complaints. More detailed, quantitative criteria for HIZ (extent and intensity) may provide better agreement on the prevalence of HIZ and more valid results on relationships with LBP [159].

Nucleus pulposus signal and disc height: Pfirrmann et al's widely used system for rating disc degeneration on MRI implies a combined rating of nucleus pulposus signal, disc structure, the distinction between nucleus and annulus, and disc height [69]. We chose to rate nucleus pulposus signal separately from disc height, as these two variables were separated in the MRI indication for disc prosthesis surgery; ≥ 40% disc height decrease was a sufficient indication without any other MRI findings. It was also relevant to rate individual and not only overall degenerative disc findings because, according to Battie et al, "the determinants of disc degeneration and their effect sizes differ between specific degenerative findings. Thus, aggregating findings associated with disc degeneration into summary scores may mask relations" [24]. Measured disc height decrease ≥ 40% was one of the inclusion criteria used as MRI indication for disc prosthesis, and accordingly we analysed reliability for this variable. A strength was that we in addition analysed reliability for qualitative judgment of disc height, which is the usual approach in clinical practice.

Facet arthropathy (FA): In a systematic review on grading systems for lumbar disc and facet joint degeneration Kettler and Wilke recommended to use three to five grades and to assign the normal state "grade 0" [76]. In our study, FA was rated in accordance with this recommendation based on a combination of Fujiwara et al's criteria [77] and published illustrations used in the SPORT trial (available in an appendix to [65]). Kappa for interobserver agreement on FA was slightly higher in the SPORT trial (0.54) [65] than reported for Weishaupt et al's grading system (0.41)

[160] recommended in the systematic review [76]. In Weishaupt et al's grading system, FA is graded 0 when facet joint space is normal (2-4 mm width); grade 1 when joint space is narrowed (< 2 mm) and/or small osteophytes and/or mild hypertrophy of the articular process are present; grade 2 when narrowing of the joint space and/or moderate osteophytes and/or moderate hypertrophy of the articular process and/or mild subarticular bone erosions are present; grade 3 when narrowing of joint space and/or large osteophytes and/or severe hypertrophy of the articular process and/or severe subarticular bone erosions and/or subchondral cysts are present [160].

A potential drawback with Fujiwara et al’s criteria that we used is that they only concern joint space narrowing, osteophytes and sclerosis and do not explicitly imply an evaluation of hypertrophy and erosions as well. However, by using Fujiwara et al’s criteria in combination with the illustrations applied in the SPORT trial, we hoped to compensate for this drawback and at the same time keep the grading criteria simple. Given our results, this approach was less successful than we hoped for, and Weishaupt et al’s criteria recommended in the review might perhaps have been more useful. However, prior to the study, we felt that our approach to FA was reasonable, given the higher kappa value in the SPORT trial and discussions regarding the rating of FA during the pilot study. Fujiwara et al had concluded that the reliability for rating FA on MRI is acceptable although bony cortex margin is less well depicted and thinning of cartilage is more difficult to measure accurately on MRI than on CT, and MRI probably underestimates the severity of FA compared to CT [77]. The degree of FA may be underestimated both by CT and MRI compared to histologic grading [161].

- Time to follow-up and prevalence of change

A longer time to follow-up MRI would have improved the study of reliability of change in MRI findings (paper II) by increasing the number and spectrum of changes to rate. Some alterations in MRI findings were quite rare. For example, the observers reported new Modic changes at 0.8 % to 6.8 % (1 to 5) of endplates when comparing pre-treatment and 2-year follow-up images (Table 1 in paper II). By comparison of images, the highest numbers of changed MRI findings (22 to 37) were reported for progress of nucleus pulposus signal at L3/L4 (22, 17.5 % of 126) and FA at L4/L5 (26, 20.6 % of 126) and L5/S1 (37, 29.4 % of 126) (Table 1 in paper II). These numbers suggested an adequate range of changes to rate. Furthermore, the limited range of changes for some findings is unlikely to have biased the comparison between image evaluation methods across all findings. We could have rated the reliability of change in a constructed image sample with a higher prevalence of change, but such a sample would not have reflected practice.

- Determination of conclusive MRI findings (paper III)

We based conclusive MRI findings (paper III) on simple majority, median rating, or on a fourth radiologist's rating when majority or median was unsuitable (complete disagreement on type of Modic changes). Alternative ways to determine conclusive findings would have been consensus interpretation or single observer rating based on for example the observer with the highest intraobserver agreement. Consensus interpretation is commonly used in imaging studies, but it does not reflect clinical practice, and it may encourage pseudo-consensus because of "group-thinking" and undue influence of dominant observers [162]. A single observer is not representative of the population of radiologists, since observers differ in cognitive and perceptual abilities [141]. Furthermore, an additional reader (even a moderately experienced one) may improve the reproducibility of conclusive MRI interpretations compared to one expert alone [163]. It was therefore a strength that the conclusive MRI findings in paper III were based on multiple observers' interpretations and were therefore likely to be even more consistent than the observers' independent findings [163], which showed mostly good or moderate interobserver reliability (paper I). The high reliability further reduced the chance of underestimating the MRI findings' relationship to other variables [102].

- Disability and LBP scores (paper III)

The well-defined patient population with longstanding, non-radicular LBP, and presumed pain-relevant localized degenerative findings on MRI displayed a wide range of disability scores (**Figure 12, section 3.3**), LBP intensity scores (**Figure 11, section 3.3**), and MRI scores (**Figure 13, section 3.3**). It was therefore well suited for examining a dose-response relationship between the extent of MRI findings and the degree of disability and LBP. Importantly, our aim was to examine such a relationship and not to assess the risk of LBP according to the presence or not of different MRI findings; that would have required a control group without LBP.

A dose-response relationship might have suggested causality, even if found in a cross-sectional study [164]. Yet, the cross-sectional design limited our ability to study causal relationships.

Reliability and construct validity of the Norwegian version of the modified ODI (version 2.0) has been evaluated as acceptable for assessing chronic LBP (2-days interval between measures for reliability analysis and correlation with physical functioning scale of SF-36 for analysis of construct validity) [130]. Good reliability has been shown for LBP intensity measured on VAS [132], and construct validity is supported by results showing that pain intensity measured on VAS is positively correlated to other measures of pain intensity as well as pain behaviour [165].

The use of ODI and LBP the past week as clinical measures may be a limitation in patients with recurrent or chronic LBP, since their pain may vary considerable over time. Pain measures aggregated across time and across different measures may have better reliability and sensitivity for chronic pain than single pain items and it has been proposed that the number of days with LBP in a 6 month or 1 year period is a better measure for pain in these patients [165]. Current LBP the last week is also difficult to interpret in relation to MRI findings that develop over a long time. I would expect such MRI findings to correlate better with for example mean LBP measured over a long time (perhaps ≥ 1 year), than with current LBP defined as maximum LBP the past week.

Self-reported pain is a subjective measure that is not directly proportional to nociception (defined as nerves or receptors sending signal about tissue damage) [26, 165]. Many physiological and psychosocial factors, including chronic pain, emotional troubles, poor job satisfaction, alcohol and narcotic abuse, and compensation issues influence patients' subjective reports of disability and pain [166]. Such factors may obscure a potential relationship of MRI findings to ODI and LBP intensity scores.

- Statistical analysis

Reliability and kappa statistic: According to Ker (1991) "Reliability is the consistency with which some measure assesses a trait. Interrater reliability, then, is the consistency between raters in assessing some trait. If one rater's judgments are consistent with another, they are said to agree, agreement and reliability are often used synonymously, and as the opposite of disagreement and variability" [108]. When using the kappa statistic, one treats ratings as if the ratings are totally independent of the actual changes on the MRI images, but in real life observers apply a decision-making system to rate findings. The system may consist of a written definition describing the characteristics of the finding and/or images illustrating the findings, and of a decision-making algorithm describing a stepwise approach to decide the rating. In kappa statistic, the decision-making algorithm "is not a valid way to diagnose a case, and any agreement obtained thereby is purely chance" [108]. In the clinical setting this approach to agreement is therefore very strict, and chance agreement is probably overestimated on behalf of agreement due to real agreement on characteristics of a finding. Accordingly, real agreement may be underestimated. In addition, the magnitude of kappa indicating acceptable agreement in the clinical setting and in research is unclear [105].

Bonferroni correction: We used Bonferroni correction for multiple comparisons (see definitions) in paper I and II to reduce the risk of type I error (to find a difference when no difference really exists). This correction of the significance level is regarded as stringent and its use is debated because "when the number of comparison becomes large it may become impossible to show significant findings" [167]. According to Altman "For small numbers of comparisons (say up to five) Bonferroni correction is reasonable, but for large numbers it is highly conservative" [109]. Kent et al suggest that multiple comparisons should be permitted without Bonferroni corrections in phase 1 studies (hypothesis-setting studies), which are regarded as an exploratory phase [4]. In paper III we did not reduce the significance level to account for multiple testing and thus reduced the risk of missing important relationships.

<u>Paper I:</u> It was a strength that we analysed disagreement on prevalence of findings (bias). Such systematic disagreement is seldom reported (**Table 1 in appendix**), but have impact on the interpretation of kappa magnitude and how to improve the reliability [104]. Systematic disagreements on the rating of an MRI finding between observers indicate that improvement or clarification of the rating criteria may be needed to improve reliability.

A further strength was that we dichotomized MRI findings into categories that reflected the inclusion criteria and had been presumed to be relevant to symptoms and treatment decisions. We were less interested in minimal or borderline findings. This increased the clinical relevance of our results. It also ensured a higher prevalence of each analysed category and thus interpretable kappa values for most MRI findings (kappa is difficult to interpret when a category has prevalence < 10 %) [65, 104, 117]. We could have analysed the original categories using weighted kappa, but some of these categories had low prevalence, which might inflict the kappa values. Dichotomization into marked or less marked (rather than normal) MRI findings has also been used in some other studies [89, 152, 168].

<u>Paper II:</u> The rating of change in MRI findings was dichotomized into progress versus unchanged/regress and into regress versus unchanged/progress. We could rather have rated and analysed different steps of change to the worst or to the better. However, this would have resulted in a very low prevalence of change at each step, because any change was infrequent at the 2-year follow-up.

We defined adjacent locations at risk of artefacts from a disc prosthesis as 1) the facet joints at the prosthesis level(s) and 2) all evaluated locations at a) the nearest level above and b) the nearest level below the prosthesis level(s). We did not analyse reliability at each of the locations 1, 2a, and 2b separately. The impact of artefacts on reliability might perhaps differ between these locations, but we had no prior scientific or clinical data to support this and we did not analyse this further.

We did not find it appropriate to use ordinary kappa in paper II, due to a low prevalence of changes to rate. Low prevalence reduces kappa, while bias (disagreement on prevalence) somewhat increases kappa [104]. We therefore calculated PABAK values, which reflect a hypothetical situation without any effect of prevalence or bias [104]. Because the prevalence of change partly differed between the two image evaluation methods, we also assessed these methods' impact on agreement (regardless of prevalence) using PABAK rather than unadjusted kappa. The use of PABAK implied that - in the analysis - the prevalence of ratings was made equal for both dichotomised categories (e.g. progress or not) that the observers agreed on. This was done by reclassifying some agreements on no change to agreements on change, despite that it may be more (or less) difficult to agree on change than on no change. These adjustments reflect a hypothetical situation and not the clinical setting.

Paper III: A large patient sample with presumed pain-relevant MRI findings, the wide spread of MRI scores (**Figure 13, section 3.3**) and disability/pain scores (**Figures 11 and 12, section 3.3**), and the additional *post-hoc* analyses made it unlikely to miss important relationships between the sum of MRI findings and the degree of disability/pain. The good or moderate reliability of the MRI findings (known from paper I) also reduced the potential underestimation of their relationship to other variables [102].

There is no validated system with well-tested and adequate measurement properties for scoring a combination of lumbar MRI findings in LBP patients. We decided *a* *priori* to weight Modic changes higher than other MRI findings in the MRI total score when comparing this score to the degree of disability and LBP. The decision was made based on studies of associations between MRI findings and the presence (not degree) of LBP. It was not supported by our results for individual MRI findings. However, when we assigned the same sub-score to Modic changes as to each of the other findings, the MRI total score was still not related to ODI or LBP intensity (*P* values for the between-subjects effect were 0.96 and 0.44, respectively). Latent Class Analysis may help to identify relevant clusters of MRI findings based on for example biologically plausible pathways of degeneration [169]. Perhaps results from this type of analysis will give a better basis for choosing and combining individual MRI findings in a MRI total score. Latent Class Analysis may also be helpful in subgrouping of pain and disability characteristics for LBP patients [13].

Only Modic changes with CC extent larger than small dots were included in the MRI total score. This reduced the chance of including artefacts (for example chemical shift, susceptibility, and flow artefacts) or bone marrow inhomogeneities disguised as small Modic changes [84, 148]. Some prior studies on relationships between Modic changes and clinical variables have also excluded the smallest reported Modic changes from the analysis [83, 170].

The effect size and direction (positive or negative association) of degenerative disc findings on LBP and disability may differ between different specific findings. Consequently, a summary score of disc degeneration findings may obscure associations [24] especially if the score represents both atrophic and reparative processes with potentially opposite effect on pain and disability [72]. It is therefore a strength that we also explored associations between the individual MRI findings (and not only the MRI total score) and the degree of LBP and disability.

We adjusted for age, gender, BMI, smoking, and anxiety/depression based on prior studies [24, 99, 125, 166, 171]. Degenerative MRI findings were inclusion criteria and differed little with age in our study; yet, the relation of such findings to symptom severity may depend on age [172].

The rate of missing data was high for physical workload (18.8%), which may worsen LBP [125], and physical leisure time activity (13.4%), which may improve chronic pain [33]. Therefore, and to have ≥ 10 patients per variable in the primary analyses [173], these two variables were only included in the *post-hoc* analyses (where we had 124-128 patients and 15-17 variables). Co-existent FA was also only analysed *post-hoc*, since we focused on MRI findings used as indications for disc prosthesis surgery. Some surgeons rather consider that advanced FA at the disc level potentially suited for surgery is a contraindication and may worsen the surgical result [78, 79].

Clustering (multiple measures in each patient and at each level): MRI findings are likely to be more similar within vs. between patients (due to genetic and environmental factors) and within vs. between disc levels (due to biomechanical factors) [24, 61, 121]. The observations within one patient or within one disc level are therefore not independent but clustered, and the basic assumption of independent observations within the studied group, which is essential for most statistical analyses, is not fulfilled. To analyse clustered data Kirkwood and Sterne recommend using summary measures for each cluster, robust standard errors, random effects models or generalized estimating equations (GEE) [173]. In our study data were clustered both on patient and spinal level. To account for this, we analysed reliability separately at each rated level (paper I and II) and compared reported prevalence between observers by logistic regression for longitudinal data, fixed effects model (paper I). We used GEE to analyse the impact of an adjacent disc prosthesis and image evaluation method on PABAK, and impact of observer and image evaluation method on rating of change in MRI findings (paper II). In paper III, we accounted for clustering of MRI findings by using a summary MRI score for each patient. Only few previous studies have taken clustering into consideration (**Table 1 in appendix**).

- Summary of strengths and limitations

1. A relatively large patient sample from a well-defined, chronic LBP population (i.e. candidates for surgery with lumbar disc prosthesis)
    - with localized MRI findings presumed to be relevant to disability, pain and treatment decisions, and
    - with a wide range of MRI scores, disability scores, and pain scores.
2. The MRI images reflected clinical practice, and the three observers were from three different institutions in Norway.
3. The MRI technique (pulse sequences and slice thickness) was generally consistent with guidelines for performing MRI of the adult spine.
4. Blinding of the observers to clinical data, each other's ratings and their own prior ratings, long time lag (> 3 months) and altered random examination order between the first and the second rating to reduce recall bias.
5. Dichotomization of MRI findings into categories that reflected the criteria for including patients as candidates for surgery.
6. Analyses of disagreement on prevalence of findings (bias) were included.
7. Change in MRI findings was rated in two ways used in research: by comparing initial and follow-up images and without comparing them.
8. Longer time to follow-up MRI would have improved the study of reliability of change in MRI findings (paper II) by providing more changes to rate.
9. Conclusive MRI findings based on multiple observers' interpretations (paper III) and adequate reliability (paper I) reduced the chance of underestimating the MRI findings' relationship to LBP and disability.
10. We accounted for clustering of MRI findings in all three studies.
11. The cross sectional design did not allow for causative investigations.

5.2 Discussion of results

Here I will summarize main results and compare them to results of other studies. Further, I will propose explanations for the findings and for differences in results between studies, and suggest interpretations and implications of the research. In this section, I will also describe supplementary reviews I have conducted of selected images to explore potential sources of disagreement on the MRI findings. These reviews were not part of the aims and results of the papers.

Reliability of pre-treatment MRI findings

- Disagreement on prevalence of findings

Interobserver disagreement on prevalence: The prevalence differed most (two- to threefold) between observers for Modic changes, HIZ, severely reduced disc height and moderate/severe FA, and less for dark/black nucleus signal, measured ≥ 40 % disc height decrease and abnormal disc contour (bulge/herniation) (**paper I**). Importantly, the differences in prevalence took a different direction for different findings and did not add up to an even larger difference for the MRI indication for prosthesis. For example, observer B tended to report a lower prevalence of Modic changes and ≥ 40 % disc height decrease than observer C but a higher prevalence of HIZ and dark/black nucleus signal and thus a more similar prevalence of the MRI indication (that was based on these four findings).

Comparable data regarding interobserver disagreement on prevalence of lumbar MRI findings are scarce. Carrino et al found differences in frequency distributions (Wald test) between trained experts for disc degeneration ($P = 0.055$) and FA ($P = 0.006$) but not for Modic changes ($P = 0.52$) or HIZ ($P = 0.22$) [65]. Their results are difficult to compare with ours. They reported p values for dichotomized findings but did not report the dichotomized categories, except for HIZ (none vs. any), and they did not analyse separate lumbar levels, as we did.

Also different from us, Carrino et al graded disc signal and disc height combined as disc degeneration, according to Pfirrmann (five categories) [69]. For FA, our results seem to support those of Carrino et al. Although the difference was not significant after Bonferroni correction ($P = 0.0027$), we found that the prevalence of moderate/severe FA at L4/L5 differed more than twofold between observer B (5.9 %) and observers A and C (both 14.1 %). Similar to us, Carrino et al found disagreement on FA between observers despite common training and use of images illustrating the rating categories.

It is not clear whether Carrino et al included “small dots” as Modic changes as we did. Such small changes may be more difficult to agree on, and may have caused disagreement on prevalence of Modic changes in our study (cf. “*Summary of explanations for the results (paper I)”* on page 93). For HIZ, Carrino et al found similar frequency distribution across observers whereas we found up to three-fold differences in prevalence, perhaps partly due to short training and variable interpretation of rating criteria (cf. “*Ways of improving agreement”* on page 101). In two other studies, two observers reported rather similar prevalence of HIZ (11.1 % vs. 8.6 % and 13 % vs. 17 %, respectively; *P* value for difference not given) [68, 174]. These results did not concern individual lumbar levels. Lurie et al found similar frequencies across readers for bulges and normal discs combined [119]. Our results are not comparable; we did not combine bulges and normal discs.

To explore *potential sources* for disagreement on prevalence between observers in our study, I reviewed the pre-treatment images in all 58 discrepant cases for the three findings with most pronounced disagreement. These were 32 cases of Modic changes inferior to the disc at L4/L5 (present according to observer C only), 11 cases of severely reduced subjective disc height at L4/L5 (according to observer A and C but not observer B), and 15 cases of HIZ at L5/S1 (present according to observer B and C but not observer A).

All re-interpreted Modic changes inferior to the disc at L4/L5 (32 discrepant cases) were small, and some of the changes could be due to inhomogeneous fat in the bone marrow or signal from vertebral endplate veins [175]. All re-interpreted severely reduced discs heights at L4/L5 (11 discrepant cases) were borderline between grade 2 and 3. The distinction between grade 2 (disc lower than disc above) and grade 3 (endplates almost in contact) may be more subjective than that between grade 2 and 1 (disc as high as disc above) and between grade 1 and 0 (disc higher than disc above). For re-interpreted posterior HIZ at L5/S1 (15 discrepant cases), the brightness of the lesion was a likely source of disagreement on prevalence.

Intraobserver disagreement on prevalence: The prevalence of findings differed significantly (but slightly) only for abnormal disc contour at L5/S1 (observer B). Previous data regarding intraobserver disagreements on prevalence of lumbar MRI findings seem to be lacking.

In general, reasons for intraobserver disagreements on prevalence of findings may include *change* in the observers' tendency to prefer one or another response category [108] and adjusted interpretation of rating criteria over time. Small adjustments may perhaps be due to reflections on the rating criteria and changed beliefs during a study about the prevalence of a finding and about overrating and underrating of findings. These and other beliefs may influence the rating of findings especially when in doubt, and may partly explain differences in prevalence of findings between an observer's first and second rating.

To illuminate *sources* of intraobserver disagreement on prevalence in our study, I reviewed the pre-treatment images in all 53 discrepant cases for findings with $p < 0.05$ for difference in prevalence between the first and the second rating (although the significance level was 0.002). For observer A these were 11 cases of Modic changes inferior to the disc at L4/L5, 6 cases of > 50 % AP extent of Modic changes inferior to the disc at L5/S1, and 10 cases of abnormal disc contour at L4/L5. For observer B these were 13 cases of abnormal disc contour at L5/S1, and 13 cases of moderate/severe FA at L5/S1.

All 11 reviewed cases of Modic changes inferior to the disc at L4/L5 concerned small changes. The second but not the first rating was Modic changes in 10 cases, and vice versa in 1 case. In all 6 discrepant cases of > 50 % AP extent of Modic changes inferior to the disc at L5/S1, the AP extent was up-rated in the second rating (from < 25 % to > 50 % in 5 cases and from 25-50 % to > 50 % in 1 case).

In 9 of 10 reviewed cases of abnormal disc contour at L4/L5, a bulge had been down-rated to normal in the second rating; the last case was an up-rate from normal to herniation (which was a small lateral herniation on review). In the 13 reviewed cases of abnormal disc contour at L5/S1, the discrepancy concerned up-rating from normal to herniation (7 cases) and normal to bulge (6 cases). In the 13 reviewed cases of FA at L5/S1, the discrepancy concerned moderate vs. mild FA in 9 cases (8 were up-rates) and moderate FA was down-rated to normal in 4 cases.

In line with our observations, Brant-Zawadzki found that most variability in rating disc contour abnormalities concerned normal vs. bulging disc [138]. Similarly, Pfirrmann et al reported that most disagreement on disc degeneration concerned adjacent rating categories: “disagreement was more frequent between Grades I and II in terms of inter- and intraobserver agreement, and between Grades III and IV in terms of interobserver agreement” [69]. They stated, “this can be explained by the main discriminating features between these grades (homogeneous *versus* inhomogeneous bright nucleus for Grades I and II and the possibility of differentiating the annulus and nucleus for Grades III and IV), which are subject to a larger scope of interpretation than for the other grades” [69].

- Interobserver and intraobserver agreement

Interobserver agreement was generally moderate or good for findings included in the present indication for disc prosthesis (Modic changes, HIZ, dark/black nucleus pulposus, ≥ 40 % disc height decrease) but only fair for FA. Pairwise kappa was ≤ 0.40 (< moderate) in one observer pair for inferior AP and CC extent and superior AP extent of Modic changes, HIZ and disc contour at L5/S1, and fair for FA at L4/L5. Intraobserver agreement was mostly good or very good.

Our kappa values for interobserver and intraobserver agreement were generally similar or higher than in some prior studies for Modic changes (kappa range for interobserver/intraobserver agreement in referred studies 0.44-0.62/0.64-0.73) [65], HIZ (kappa 0.44-0.62/0.67-0.73) [65, 68, 116, 117], nucleus pulposus signal and disc height combined (kappa 0.49-0.66/0.69-0.74) [65, 117], disc signal intensity (kappa 0.59/0.87) [73], disc height reduction (kappa 0.66/0.81) [73] and abnormal disc contour (kappa 0.55/0.69) [117]. For FA, however, our kappa values were lower compared to two prior studies (kappa range 0.41-0.54/0.69-0.76 for interobserver/intraobserver agreement) [65, 160] and similar to the values in one study [120] (kappa range 0.07-0.21/0.26-0.36). This may be partly due to non-standardized images and low prevalence of moderate/severe FA in our sample (11.4 % at L4/L5). When we adjusted for low prevalence by using PABAK, interobserver agreement for moderate/severe FA appeared good (PABAK 0.74 at L4/L5 and 0.66 at L5/S1).

In three studies based on standardized MRI of 40-year-olds from the normal population, kappa values were slightly higher for Modic changes [80], HIZ [73] and abnormal disc contour [116]. Standardized MRI and common training may partly explain the better agreement for these MRI findings, as the observers in two of these studies evaluated images in a joint training session to ensure consensus in the evaluation process (50 images [73] and 15 images [80]), and training can improve agreement [63, 106].

In prior studies, lumbar spine findings had mostly moderate to good reliability at low-field MRI (< 0.2 T) [70], mid-field MRI (0.2 to 0.6 T) [73, 80, 116, 118], and high-field MRI (1.0 to 1.5 T) [65, 117, 170, 176]. It therefore seems that the improved image quality at high-field MRI may not necessarily imply improved reliability; however, the comparison of reliability between studies is difficult for several reasons, such as differences in prevalence, bias, and MRI rating criteria [65, 68, 69, 71, 77, 88, 94, 117, 138, 170, 172, 177-179].

- Summary of explanations for the results (paper I)

In general, one potential reason for interobserver disagreement on prevalence is differences in interpretation and use of rating criteria. Another possible reason is differences in the observers’ response bias, i.e. their tendency to prefer one or another response category (to rate up or down, particularly when in doubt), independently of the characteristics of the object [108]. The interobserver and intraobserver variation will probably increase if the finding itself or the criteria for grading it are subject to a large extent of subjective interpretation.

Small bone marrow changes may be difficult to classify as either Modic changes, inhomogeneous fat in the bone marrow or signal from vertebral endplate veins. The CC extent of Modic changes may be easier to evaluate than the AP extent; the CC extent is graded compared to the height of the vertebrae, which is nearly constant on most sagittal slices whereas the AP extent is graded compared to the AP diameter, which is not constant because of the ovoid form of the vertebrae.

The rating criteria defined HIZ as brighter than and separated from the nucleus pulposus. Doubt about true signal intensity of HIZ may be due to partial volume effect or signal averaging in voxels (small lesions), use of different window/level settings when viewing images and lack of a normal nucleus pulposus signal to compare with at any of the lumbar levels. It may also be difficult to determine whether the high signal really is separated from nucleus pulposus if the slice is oblique to the direction of a rupture or the nucleus is black. Agreement on HIZ might be improved by looking more closely at both axial and sagittal images and at the signal intensity compared to nucleus.

It seems that distinction between disc height reduction grade 2 (disc lower than disc above) and grade 3 (endplates almost in contact) may be more subjective and more difficult to agree on than that between grade 2 and 1 (disc as high as disc above) and between grade 1 and 0 (disc higher than disc above). Higher reliability can probably be expected for measured vs. subjectively rated disc height (higher kappa and less disagreement), although the differences are small and can be partly explained by different prevalence of reported disc height loss by the two methods.

Change in the interpretation of criteria and beliefs may have influenced the rating of disc contour as the tendency was strong for observer A to down-rate (from bulge to normal) and observer B to up-rate (from normal to herniation or bulge) from the first to the second reading. If an observer believe that he/she has over- or underestimated a finding an unconscious tendency towards the opposite may evolve, for example by changing the tendency to grade up or down when in doubt.

Change in the interpretation of criteria and beliefs may also have influenced the rating of FA as observer B had a strong tendency to up-rate FA from the first to the second reading. FA may be easier to rate on CT than on MRI [120, 160], but we did not directly compare these two imaging techniques in our study.

Reliability of change in MRI findings over time

- Reliability of change by comparison of images

Based on PABAK values, interobserver agreement was mostly good for progress and regress of Modic changes and disc findings, but was moderate for progress regarding nucleus pulposus signal and FA at L4/L5 and L5/S1 (**paper II**). These results concerned evaluation of change by comparison of pre-treatment images and follow-up images taken after disc prosthesis surgery or non-surgical treatment.

No comparable reliability data exist for change in disc findings or FA. The only prior study on reliability of change in lumbar MRI findings over time concerned change in Modic type and extent by comparison of new and old images [80]. In that study, kappa for interobserver agreement on different alterations in Modic changes over time ranged from 0.50 to 0.60 by comparison of images. Based on those results, the authors recommended non-comparison when studying the course of Modic changes [80]. In our study, PABAK for interobserver agreement on change in presence of any Modic changes by comparison of images ranged from 0.89 to 0.98 for progress and from 0.86 to 0.97 for regress. These results are difficult to compare with the results from Jensen et al because our study differs from their study in classification of change (progress and regress separately vs. no change, increase or decrease in T1 signal, T2 signal, intravertebral volume and endplate extension), prevalence of change and thus statistical approach (PABAK for each lumbar level vs. ordinary kappa and clustering of all levels with prevalence 10-90 %), sample (change in existing and new vs. only existing Modic changes), and MRI technique (high-field vs. low-field MRI, standardized vs. non-standardized images).

The use of standardized images was likely to improve agreement on findings [105, 180] and was therefore also likely to improve agreement on change in findings in the study by Jensen et al [80] compared to our study. They evaluated existing Modic changes only, and this may also have contributed to improved agreement, since it may be particularly difficult to agree whether new small bone marrow changes are Modic changes or not (overlap between normal and abnormal findings [180]).

Based on discussions with clinicians and researchers with varying MRI experience Jensen et al [63] selected CC extent as the size variable that was easiest to evaluate. In **paper II**, agreement on change in CC and AP extent was similar with good or very good PABAK values. However, CC extent may be more likely than AP extent to change over time; in our study, both observers reported a higher prevalence of progress at L4/L5 and L5/S1 for CC extent compared to AP extent. If one chooses to use one single measure for size of Modic changes (in a scientific study for instance), it seems reasonable to use CC extent rather than AP extent.

Fluid in the disc may be one source of disagreement on nucleus pulposus signal and change in such signal over time. The fluid fills vacuum clefts and may come and go [69, 81] (Conf. introduction section 1.3). This makes it difficult to rate the true nucleus pulposus signal.

To explore whether fluid in the disc may have caused disagreement in our studies, I reviewed the 8 discrepant cases for the level (L5/S1) and observer (A) with the most marked intraobserver disagreement on prevalence of nucleus pulposus signal ($p = 0.070$) in **paper I**. In 4 of these cases, the first and second rating differed two rating categories, the disc was low, and the high signal in the disc was rectangular (not ovoid as in a normal nucleus pulposus) and extended outside the nucleus pulposus. Therefore, fluid in the disc may have caused intraobserver disagreements and may thus be an important source of interobserver disagreements as well. Criteria for rating nucleus signal when fluid extends outside the nucleus would be helpful.

One reason for only moderate agreement on progress of FA at L4/L5 and L5/S1 may be that FA is more difficult to rate, and therefore more difficult to evaluate for progress in rating. Although PABAK indicated good agreement on moderate/severe FA, ordinary kappa suggested only fair agreement and the prevalence of FA tended to differ more than twofold between observers (**paper I**). Standardization of images (e.g., of slice thickness and gap, angulation of slices in relation to the facet joints, resolution, signal to noise ratio) may lead to better agreement on change in FA, since similar images are easier to compare. Better rating criteria for FA may also lead to improved rating of change. Such criteria might perhaps imply measurements of joint space narrowing and osteophytes, evaluation of joint fluid, subchondral oedema, sclerosis, cysts and erosions, articular process hypertrophy, and synovial cysts. Interobserver variation may be smaller for measurements than for categorical ratings (such as those we used for FA), especially if landmarks used to measure are clearly visible, the measurement method is well defined, and the imaging technique is standardised [106].

In summary, mostly good interobserver agreement on change in Modic and disc findings can be expected by comparison of images. However, improved criteria may be needed to improve the rating of change in nucleus pulposus signal and FA.

- Impact of adjacent disc prosthesis on the reliability of change in MRI findings

Adjacent disc prosthesis did not influence interobserver agreement on progress and regress across all MRI variables. Still, agreement on progress of FA at L5/S1 adjacent to prosthesis was only fair (PABAK 0.29 by comparison of images).

That disc prostheses cause artefacts is well known. Yet, it seems that no other study has explored the impact of adjacent disc prosthesis on the reliability of MRI findings.

Although adjacent prosthesis did not influence agreement overall, it might reduce agreement on progress of FA. The metal in prostheses can cause different types of MRI artefacts including signal loss and geometric distortion [143]. Signal loss is usually obvious (**Figure 14**), whereas small or moderate distortions may be difficult to appreciate. Such distortions may still interfere with the evaluation of for example FA by obscuring or mimicking osteophytes. Metal artefacts are likely to affect facet joints more at the disc prosthesis level than at the more distant adjacent levels superior to and inferior to the prosthesis. The observers rated FA at all adjacent levels and other findings at the more distant adjacent levels only. This may partly explain why FA differed more than most other findings in PABAK at adjacent vs. not adjacent level (0.44 vs. 0.69 at L4/L5, 0.29 vs. 0.66 at L5/S1, **paper II**).

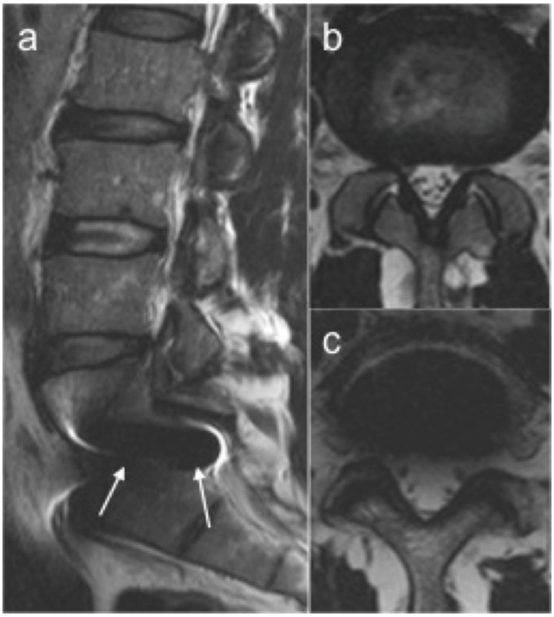


Figure 14 Sagittal T2-weighted (a) and axial T2-weighted MRI images (b and c) from one patient illustrating that the artefacts (distortion and signal loss, arrows) produced by the disc prosthesis is more pronounced at the index level (in this case L5/S1, c) than at the nearest adjacent level (in this case L4/L5, b).

It might perhaps be easier to assess change in FA after disc prosthesis surgery when more similar, early post-operative (and not only pre-treatment) images are available for comparison. However, we found that observer influenced the rating of change in FA, regardless of whether images were compared or not. This finding suggests that observers interpret criteria differently and that better rating criteria and/or a more consistent interpretation of such criteria is needed.

- Reliability of change by comparison versus non-comparison of images

For a range of existing and new findings on non-standardized clinical MRI images of the lumbar spine, we found comparison of images generally more reliable than non-comparison for evaluating change. Comparison provided significantly better interobserver agreement across all variables and a higher mean PABAK both for progress and regress **(paper II**). In addition, the observers reported fewer changes by comparison of initial and follow-up images than by non-comparison.

Our study seems to be the only study that has compared the reliability between these two methods for evaluating change in MRI findings. However, there is evidence that availability of previous images improves diagnostic accuracy [181].

Some explanations for reduced variability and lower prevalence of change by comparison of images can be considered. By non-comparison, doubt about the rating of a truly unchanged finding could result in different pre- and post-treatment ratings even if no change is evident by comparison (only doubt about the correct rating category). Comparison of images could also help to avoid variable reports of change due to differences in MRI technique or image quality. This advantage would be larger when comparing non-standardized images like in our study than when comparing standardized images such as in the study by Jensen et al [80].

Agreement on change was similar for CC extent and AP extent of Modic changes, with good or very good PABAK values. However, image evaluation method (comparison vs. non-comparison) significantly influenced the prevalence of change in AP extent $P < 0.001$) but not the prevalence of change in CC extent ($P = 0.637$) (**paper II**). Thus, the prevalence of change in CC extent seems to be less dependent on image evaluation method. This may be a further argument for using CC extent rather than AP extent when monitoring the size of Modic changes.

In summary, comparison of images provided better agreement on change in MRI findings than non-comparison. This suggests that comparison of images may be preferable for evaluating the course of Modic changes, disc findings and FA over time both in clinical practice and in research.

Implications of the reliability data and potential for improvement

The present reliability data formed a basis for further studies of these MRI findings and their relationship to clinical variables (**paper III**). Our reliability data also provided a basis for evaluating the clinical relevance of change in lumbar MRI findings over time. Modic, disc, and facet findings are not yet sufficiently documented as relevant to treatment decisions [66, 182]. However, it has been suggested that the course of such findings over time (e.g., the development of adjacent-level degeneration [44, 183, 184] and index-level FA after lumbar surgery [78, 113]) may have clinical relevance. Based on results in **paper II,** new and old images were compared to study adjacent level degeneration and index level FA at 2-years follow-up in our patient population. The study showed increased index-level FA but similar adjacent level degeneration in the surgery group vs. the non-surgery group [128]. The suboptimal reliability for rating change in FA at the prosthesis (index) level could hardly explain the marked difference in index-level FA between the groups [128]. The data in paper II still suggest a need for improved reliability for rating change in FA at the 8-year follow-up of this cohort. Our reliability data suggest that there is a room for improvement in the reliability for other MRI findings as well, but I will first discuss which level of reliability is adequate or sufficient.

- “Adequate” reliability

There is no firm rule for when the reliability of a finding is adequate and the use of multiple readers, e.g. in a study, might improve the rating of a finding [106, 163]. A finding with high reliability is not necessarily valid, and a finding may be useful in research or practice despite modest reliability [102]. It is noteworthy that many tests used in daily clinical work have only moderate reliability at best [102]; kappa was 0.32 for ECG interpretation in the emergency department [185], 0.16 for clinical evaluation of lumbar lordosis [186], and 0.20-0.47 for agreement on a positive Lasegues test [187]. We still suggest that kappa ≤ 0.40 for interobserver agreement should lead to an assessment of how to improve the reliability. In a systematic review, Kettler et al defined acceptable kappa values as > 0.60 for disc degeneration and > 0.40 for FA [76]. Jarvik and Deyo proposed that it should be a goal for every radiologist to strive for substantial intraobserver agreement, and that interobserver reliability may approach intraobserver reliability with training, standardized nomenclature, and perhaps readily available standardized online examples with link to standardized nomenclature [102].

- Ways of improving agreement

There are several ways of reducing disagreements on prevalence and improving inter- and intraobserver agreements. I will summarize some ways here and provide examples of how they might be relevant to the present MRI findings.

Improved rating criteria: In general, criteria taking into account how to deal with borderline cases may reduce the observer’s tendency to interpret criteria and to change interpretation of criteria over time. Better criteria for the distinction between small vs. no Modic changes may increase agreement, since this distinction was a likely source of both inter- and intraobserver disagreement in our study. Such criteria could also indicate when to rate small bone marrow changes as Modic changes when similar bone marrow changes exist at other levels and may represent venous structures [175]. The importance of the distinction between small vs. no Modic changes is uncertain though, as minimal Modic changes may be less relevant [152].

Similarly, better criteria for brightness of HIZ may be needed, for example quantitative estimation of the brightness [159] or comparison of HIZ signal to the cerebrospinal fluid signal instead of the nucleus pulposus signal, which is more variable. Better criteria may also be needed for nucleus pulposus, especially on how to grade discs with fluid. Reliability for subjective grading of disc height reduction may be improved with better criteria for distinction between grade 2 (disc lower than disc above) and grade 3 (endplates almost in contact). Disagreement on grade of FA and a probable underestimation of FA by current radiologic modalities (both CT and MRI) [161] may also be reduced with improved rating criteria. More detailed and strict criteria might also help to reduce adjustments in interpretation of rating criteria over time due to changed beliefs and perhaps also response bias.

Improved image quality and more standardized images: Variations in rating due to partial volume effects may be reduced if slice localisation, thickness and gap are standardized, especially for HIZ (brightness and separation from nucleus pulposus) and FA (slice localisation relative to joint space). Different MRI techniques produce different image artefacts and image quality, and when MRI technique vary between baseline and follow-up images this may influence the rating of change in MRI findings.

Joint training: Common training of the observers may reduce disagreement [63, 106], for example by inducing a common understanding of rating criteria and how to rate when in doubt (reducing response bias). According to Brant-Zawadzki et al, “Variability between readers and within a single reader exists approximately 10-20 % of the time. Such variability is likely to be greater when readers from different institutions and different training backgrounds are compared” [138].

Reduced observer fatigue: Observer fatigue may reduce the quality of the image evaluation if the observer rates many images and many findings over a short time [180]. Thus, limiting the number of variables and images to rate in each session might help to improve rating quality and agreement.

- Summary of limitations in reliability and potentials for improvement

Modic changes (type/extent): Disagreement on prevalence of Modic changes. Better distinction needed between small and no Modic changes, and criteria for how to deal with small bone marrow changes. Disagreement on prevalence of Modic AP extent > 50 % may be reduced with improved rating criteria, for example measured AP extent compared to measured mid-sagittal AP diameter of the corresponding vertebrae.

Posterior HIZ: Disagreement on prevalence of HIZ. Moderate interobserver agreement. Better criteria for brightness of HIZ and more standardized images may be needed.

Nucleus pulposus signal: Disagreement on prevalence of dark/black nucleus pulposus. Better distinction between different grades of reduced signal and criteria for rating nucleus signal when fluid extends outside the nucleus may be needed. Moderate agreement on change in nucleus pulposus signal over time. Standardized images may improve agreement on change.

Disc height: Disagreement on prevalence of both severely reduced disc height (subjectively judged) and of measured ≥ 40 % disc decrease, but overall similar and good to moderate agreement. Common training of the observers and better criteria for distinction between grades of disc height reduction may reduce disagreement.

Disc contour: Disagreement on prevalence. Moderate interobserver agreement. Joint training may reduce disagreement due to different interpretation of criteria and on how to rate borderline cases for example between normal and bulging disc.

FA: Fair interobserver agreement, but good when adjusted for prevalence and bias. Better criteria for different grades of FA, joint training, and standardization of MRI images may improve agreement. Moderate interobserver agreement on change. Fair agreement on progress of FA at L5/S1 adjacent to prosthesis may be improved if early post-operative (and not only pre-treatment) images are available for comparison. In addition, modified MRI parameters (e.g. increased receive bandwidth) can reduce metal artefacts [188].

Relationship of MRI findings to disability and pain

- MRI total score in relation to degree of disability and intensity of LBP

In our study, the MRI total score was not related to ODI or LBP intensity. Adding FA to the MRI total score did not change the result. The MRI total score was based on findings used as MRI indication for lumbar disc prosthesis.

To my knowledge, no other study has examined the association between the extent of several MRI findings combined and the degree of disability and pain in patients with chronic non-radicular LBP. In LBP patients with and without radicular pain or sciatica, two studies have compared the sum of several MRI findings with the degree of disability (cf. Introduction). In one study, the extent of disc and facet findings, spinal stenosis, and degenerative spondylolisthesis was weakly related to disability ($P = 0.050$) [125]. In the other study, the extent of disc and facet findings, spinal stenosis, and other pathologies was not related to disability or LBP [126]. These results are not comparable to ours due to different patient populations and MRI variables.

- Individual MRI findings in relation to degree of disability and intensity of LBP

In our study, the only significant result for the individual MRI findings was a weak, negative association between a posterior HIZ at L5/S1 and ODI score. In particular, Modic type I and/or II changes were not related to ODI or LBP scores.

In a previous study on 53 patients with chronic nonspecific non-radicular LBP, a HIZ did not explain variation in disability or pain at baseline (neither did disc degeneration or Modic type I and type II changes) [39]. However, HIZ at baseline was associated with lower average LBP intensity at 12-month follow-up when adjusting for age, gender, and baseline symptoms ($P = 0.006$) [39]. In other studies with more heterogeneous samples (acute and chronic LBP, radicular and non-radicular LBP), HIZ was weakly related to LBP (**Table 2 in appendix**). Quantitatively estimated signal intensity of HIZ has been shown to be brighter in patients with LBP than in controls without LBP [159].

It is difficult to explain why we found a negative association whereas Kleinstuck et al [39] found no association between HIZ and disability score. The studies had similar definitions and prevalence of HIZ. However, only one spine surgeon evaluated the images in the study by Kleinstuck et al whereas we based conclusive findings on the majority rating of three radiologists. In addition, we analysed HIZ separately at L4/L5 and L5/S1 in a sample of patients with localized degenerative findings at these levels. Kleinstuck et al analysed HIZ for levels L1-S1 combined. We adjusted for age, gender, BMI, smoking, and anxiety/depression whereas Kleinstuck et al adjusted only for age and gender.

It is also possible that the weak negative relationship between a HIZ at L5/S1 and ODI in our study is spurious. The 4.7-point difference in ODI between patients with and without a HIZ at L5/S1 is in any case unlikely to be clinically useful or important [131]. This is supported by the fact that no significant associations were found between HIZ and LBP.

There could still be a relevant association between annular ruptures not visible as HIZs and ODI or LBP intensity. HIZ may represent a rupture extending to the outer part of the annulus [64]. However, according to Videman et al all current clinical methods (such as histology, discography, CT and MRI) will likely miss a portion of full annular ruptures, especially small ruptures and older ruptures with granulation tissue [189]. Milette et al also demonstrated that “loss of disc height or abnormal signal intensity in the disc was highly predictive of symptomatic tears extending into or beyond the outer annulus” (demonstrated by discography) [179]. Accordingly, an annular rupture may exist even though an annular rupture or HIZ is not visualized on

MRI, and an association between annular rupture and pain or disability may still exist.

We found no association between Modic changes and degree of LBP or disability whereas Jensen et al in their systematic review found an association between presence (not degree) of LBP and Modic changes [47]. Differences in the studied associations (associations to degree vs. presence of LBP) and samples (well-defined, chronic non-radicular LBP population in our study vs. heterogeneous LBP populations in the 10 included studies in the review by Jensen et al) may partly explain the diverging results. These 10 studies included populations with LBP and severe sciatica [190], LBP for more than 6 weeks with or without radiculopathy [191-194], discogenic chronic LBP without neurologic deficit [195], discogenic LBP [196], LBP past year [89], LBP and sciatic pain [152], and LBP 8-30 days previous year [98].

The *post-hoc* analyses of the 1.5 T subsample were important because magnetic field strength influences image resolution and may affect evaluation of Modic changes [148]. The additional analyses with changed MRI total score were also important, since there is no validated system with well-tested and adequate measurement properties for scoring a combination of lumbar MRI findings in LBP patients. Based on studies of associations between MRI findings and the presence (not degree) of LBP we decided *a priori* to weight Modic changes higher than the other findings, but this was not supported by our results for individual MRI findings. FA is a potential source of LBP [75, 197], but was not included in the primary analyses because we wanted to focus on the assumed pain relevant MRI findings used as indication for disc prostheses.

- Explanations for no clear relationship of MRI findings to disability or LBP

There are many possible explanations for detecting no or only a weak relationship between degenerative lumbar MRI findings and complaints. One potential explanation is that no genuine relationship exists. I will discuss three further types of explanations, numbered 1, 2, and 3 in the following.

1. *Most degenerative changes on MRI are a summation of slow structural changes over many years.* These MRI-verified slow structural changes may not cause pain during the whole period of their development. The lacking dose-response relationship between the sum of MRI findings and the degree of disability/pain in our study might suggest that the clinical impact of any of these MRI findings comes early and does not increase with the number or extent of findings. These MRI findings *per se* may be less relevant to the current degree of disability/pain, as measured in our study. Others have suggested and discussed briefly that disc degeneration and FA may be symptomatic only in the early phase of degeneration [118, 198].

2. Pathological degeneration may mimic normal aging. It is difficult to differentiate normal aging from pathological degeneration (osteoarthritis) both in the spine and in appendicular joints (e.g. knee, hip). As for degenerative findings in the spine, many people with degenerative findings in a joint are asymptomatic, and many patients with joint pain have no or minimal degeneration on imaging [57]. For example, in a study of asymptomatic people 82 % had signs of acromioclavicular joint osteoarthritis on MRI [199].

Normal age related changes in the intervertebral disc are proposed to be [58]:

a) reduced vascularity of the endplate and reduced cellularity of the disc as a response to mechanical stress,
b) decrease in overall proteoglycan and water content (especially in the nucleus, starting very early in childhood and not related to pain) with corresponding increase in collagen content, increased collagen cross-linking leading to reduced tissue strength (via reduced matrix turnover and repair) and *darker nucleus pulposus on MRI*,

c) microstructural clefts and tears (especially and at first in the endplate and nucleus) with little effect on internal function of the disc if they remain small, and
d) smaller and decompressed nucleus with transfer of load bearing forces from nucleus to annulus which also has become weaker and stiffer.

These age related changes make the disc more vulnerable to excessive load bearing and even to activity of daily living, and further disruption and degeneration may result (e.g., macroscopic annular rupture, abnormal disc contour, damage to the endplate, disc height reduction). Adams and Roughley stated that gross injuries to a disc never fully heal because of sparse cell population and long collagen turn-over time (probably over 100 years), although some regeneration of nucleus pulposus is possible in younger individuals because of faster a proteoglycan turnover (approximately 20 years) [58]. For lumbar Modic changes, Wang et al found that presence, larger size, and more endplates involved were related to greater age; an average increase in age of 17.4 years was associated with one more involved endplate [87]. In a population study Cheung et al reported that all studied lumbar MRI findings increased with age except Schmorl's nodes (10 %, constant) and disc herniation (most frequent at second decade), that 42 % had disc degeneration at age < 30 years and 88 % at age ≥ 50 years, and that 20 % of those with LBP had no sign of disc degeneration (based on nucleus pulposus signal and disc height) [200].

3. Pain is challenging to classify and measure. Pain in an osteoarthritic joint can vary substantially with time, even during the day and week; however, it is still believed that the degenerative imaging findings may be a source of pain even if they are not painful now and were not painful the past two weeks or the past year [57]. In accordance with this, Videman et al found that lumbar MRI findings were more clearly associated with lifetime parameters for LBP than with current LBP or LBP the past year. They also discussed that "the intermittent and variable nature of symptoms in the short-term, a high rate of forgotten symptoms in the long-term, and other issues affecting measurement of LBP would keep the unexplained portion of variability in pain high" [72].

To assess LBP-related disability, we used the recommended ODI 2.0 [131], which asks about the patient's "*current*" disability. We used a visual analogue scale to measure the maximum current LBP intensity *in the past week*. According to a review, moderate correlation can be expected between ODI scores and visual analogue scale scores for pain [129], and only a moderate correlation existed in our study (correlation coefficient 0.37, $P < 0.001$, n = 165). With a very strong correlation between ODI and LBP intensity, it would have been less important to analyse both variables for associations with MRI variables.

To provide a fair picture of their problem, the present patients with longstanding LBP perhaps also took their recent past disability and pain into account when reporting their scores. Still, results might have differed if we had measured the accumulated disability and pain over at least one year. In one study, the one-year and lifetime history of LBP was related to annular tears and reduced disc height [72]. Our focus in paper III, however, was on the patient's present situation; current disability as measured by the ODI score was actually part of the inclusion criteria used to decide on surgery with lumbar disc prosthesis.

Defining LBP and deciding what type of LBP to include in a study is a challenge [3, 201] (e.g. whether to include all LBP symptoms or only LBP of a certain duration and/or intensity). This may have a marked impact on patients' pain reports. For example, in a study from USA the point prevalence of LBP was 6.8 %, the 1-year prevalence was 10.3 %, and the lifetime prevalence was 13.8 %; these figures are considerably lower than the estimated prevalence in Europe [202]. In that study by Deyo et al the required LBP duration was *most days in at least two weeks* [202]. If the question about LBP was less restricted, the prevalence was the same as in Europe [203].

Another challenge is that subjective bias (such as the patient's individual reaction and attitude to pain and disease) may influence pain reports. Recall bias is a also a problem; when persons report pain in the past, the longer past time period they report on, the more unreliable are their reports [3, 204]. However, recall bias was likely to be small in our study when obtaining scores for “current” disability and LBP the previous week.

In addition, in patients with non-specific LBP, it is difficult to distinguish between pains originating from different tissues such as discs, vertebras and muscles [4]. This may also cloud associations between MRI findings and pain. For example, when analysing an association between LBP and disc findings, there is no obvious way of excluding from the analysis patients with LBP originating from muscles and not the disc. Modic type I changes typically have a pain pattern that can be regarded to indicate “inflammatory pain” [205]. However, it is not clear exactly which structure is painful. Unfortunately, we do not know whether a certain duration, intensity and pattern of pain can indicate that the non-specific LBP originates from a certain type of structure.

Furthermore, central and local pain sensitization may influence pain reports in patients with chronic pain. According to Woolf, “Central sensitization introduces another dimension, one where the CNS can change, distort or amplify pain, increasing its degree, duration, and spatial extent in a manner that no longer directly reflects the specific qualities of peripheral noxious stimuli, but rather the particular functional states of circuits in the CNS” [26]. In patients with knee osteoarthritis Arendt Nielsen et al found no correlation between radiological findings and pain parameters, but patients with the most severe pain had lower pain thresholds to knee pressure stimulation than controls (local sensitization). The degree of local sensitization was correlated to clinical pain ratings (pain intensity after walking, peak pain intensity previous 24 hours, and pain duration) [206]. In a recent review Roussel et al concluded that some studies have shown altered central nociceptive processing in patients with chronic LBP, and that future research should explore whether these changes are reversible and how this may influence the patient’s condition [37]. Central and/or local pain sensitization may be important to take into account in the analysis of associations between MRI findings and LBP and disability, as we can no longer expect a direct connection between the noxious stimuli (indicated by for example MRI findings) and pain when central and local sensitization is introduced [26].

Comorbidities such as sleep disturbances and osteoarthritis may also influence the reported LBP intensity in chronic LBP patients [23]. Hip and knee osteoarthritis are more prevalent in patients with disc degeneration, and are also risk factors for progression of lumbar disc degeneration [58]. Comorbidities including osteoarthritis may therefore be an additional confounder for the association between degenerative MRI findings and LBP.

In 2008, Dionne et al published consensus LBP definitions for use in prevalence studies [201]. A minimal LBP definition was based on questions about presence of LBP past 4 weeks and the influence of LBP on usual activities and daily routine. An optimal LBP definition was based on additional questions about frequency and severity of LBP (on a VAS 0-10) the past 4 weeks, time since the patient had a whole month without LBP, presence of referred pain to the lower extremity, and presence of any referred pain below the knee [201].

In future studies LBP should perhaps be better defined and specified according to both localization and character, what aggravates and alleviates it, and measured over a longer time span than present pain, pain in the past one or two week(s), and pain the past year [204]. In addition, comorbidities should be assessed and accounted for because they may confound the relation between intensity of LBP and MRI findings.

- Implications of results on associations

The new insight derived from our study in paper III is that the examined MRI findings cannot be used – in combination or individually - to explain the variance in current disability or LBP intensity (severity) in patients accepted for surgery. MRI findings may still be useful in treatment planning if they can predict treatment outcome. It seems unlikely that the sum of the localized MRI findings in our study should be more weakly related to the degree of disability and pain in candidates for surgery than in other patients with chronic non-radicular LBP and such localized MRI findings. Therefore, we would generally advise against using the extent of *localized* degenerative MRI findings to explain the *degree* of current disability and pain in patients with longstanding, non-radicular LBP. Whether the sum of more *widespread* MRI findings affecting also the upper lumbar levels might be related to the degree of disability and pain in chronic LBP patients still remains unclear.

6.0 Conclusions and future perspectives

Conclusions

I. When experienced radiologists assessed pre-treatment MRI findings in chronic LBP patients who were accepted candidates for lumbar disc prosthesis (**paper I**)

a) the prevalence differed up to threefold between observers for posterior HIZ at L5/S1 and for severely reduced subjective disc height at L4/L5; the MRI indication for prosthesis did not differ significantly in prevalence across observers,

b) interobserver agreement was generally moderate or good at L4-S1 for Modic changes and disc findings, only fair at L4/L5 for FA, and good for the MRI indication for prosthesis at L4/L5 and L5/S1; intraobserver agreement was mostly good or very good, and mostly very good for the indication for prosthesis.

Overall, moderate or good reliability can be expected for Modic changes and disc findings at L4-S1 in patients accepted for surgery with disc prosthesis or lumbar fusion. Mainly based on less disagreement on prevalence, results also indicated higher reliability for CC versus AP extent of Modic changes and for measured versus subjectively rated disc height. Efforts to reduce disagreement on prevalence may help to improve the reliability further. The high reliability of the MRI indication for prosthesis must be confirmed in unselected chronic LBP patients.

II. When experienced radiologists assessed change in lumbar MRI findings over time after disc prosthesis surgery or non-surgical treatment (**paper II**)

a) comparison of images provided moderate or good interobserver agreement,
b) an adjacent disc prosthesis did not affect agreement on change overall, but seemed to reduce the agreement for FA, and
c) overall, agreement on change was better for comparison than non-comparison of images.

These results support the practice of comparing images to assess change in MRI findings. In disc prosthesis patients, more reliable ways of evaluating change in FA should be sought.

III. When comparing pre-treatment degenerative MRI findings to pre-treatment disability and LBP in patients accepted for disc prosthesis surgery (**paper III**)

a) neither the sum of localized MRI findings used as indication for surgery
b) nor each individual MRI finding was related to degree of disability (ODI scores) or LBP intensity scores, and
c) adding FA to the sum of MRI findings and as an individual MRI finding did not affect the results.

The degenerative MRI findings used as indication for disc prosthesis surgery are unlikely to explain current pre-treatment variation in disability or pain intensity in candidates for lumbar surgery with disc prosthesis or fusion.

Future perspectives

The present results have implications for future research and clinical practice, and suggest some new directions for research:

- The reliability data for MRI findings (**paper** I) are an important foundation for further research on the MRI findings and their relation to clinical features in this patient population. These data formed the basis for the study in **paper III**. To be useful in research or clinical practice, an MRI finding must first prove reliable.

- The reliability data for change in MRI findings (**paper II**) are also an important foundation for further studies. Due to these data, results for adjacent level degeneration were based on comparison of images in a 2-year follow-up study of these patients treated with disc prosthesis or rehabilitation [128].

- Specific causes of disagreement and strategies to reduce it (including joint training and more standardized images) should be explored, especially for rating of FA in patients with disc prosthesis. One might attempt to reduce metal artefacts on MRI [188] and compare the reliability between MRI and CT

- We found no association between assumed pain-relevant MRI findings (used as indication for disc prosthesis) and the degree of disability or LBP, despite the MRI findings had reasonably good reliability and the LBP population was well defined. Further studies should include other variables that may explain differences in pain and pain-related disability among patients with chronic non-radicular LBP and degenerative MRI findings. Future studies might explore:

- Other measures of pain such as LBP intensity and disability over a longer timespan; we measured only current LBP and disability [13].

- Additional pain modifying variables such as pain sensitization and altered central pain processing [26], and comorbidities. How these variables may influence the association between MRI findings and clinical features is unclear and should be investigated [2, 28, 37], perhaps by means of a clinical total score of pain sensitization, comorbidities, psychosocial factors, insurance issues, genetics, and more thoroughly defined LBP and disability (e.g. mean duration of each LBP episode, days with LBP and minimum/maximum pain intensity and disability the past year, coincidental radicular pain).

- Further MRI findings for example fatty atrophy of muscles, and signs of instability, venous congestion and inflammation. The morphological findings we rated on MRI may not be those that are most relevant to pain and disability.

7.0 References

1. Vos, T., 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**(9859): p. 2163-96.

2. Airaksinen, O., et al., *Chapter 4. European guidelines for the management of chronic nonspecific low back pain.* Eur Spine J, 2006. **15 Suppl 2**: p. S192-300.

3. Waddell, G., *The back pain revolution*. 2004, second edition: Churchill Livingstone, Elsevier.

4. Kent, P., J.L. Keating, and C. Leboeuf-Yde, *Research methods for subgrouping low back pain.* BMC Med Res Methodol, 2010. **10**: p. 62.

5. van Tulder, M., 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**: p. S169-91.

6. FORMI, *Nasjonale kliniske retningslinjer for korsryggsmerter med og uten nerverotaffeksjon.* (Formidlingsenheten for muskel- og skjelettlidelser), 2007.

7. Walker, B.F., *The prevalence of low back pain: a systematic review of the literature from 1966 to 1998.* J Spinal Disord, 2000. **13**(3): p. 205-17.

8. Ihlebaek, C., et al., *Prevalence of low back pain and sickness absence: a "borderline" study in Norway and Sweden.* Scand J Public Health, 2006. **34**(5): p. 555-8.

9. Papageorgiou, A.C., et al., *Estimating the prevalence of low back pain in the general population. Evidence from the South Manchester Back Pain Survey.* Spine (Phila Pa 1976), 1995. **20**(17): p. 1889-94.

10. Vingard, E., et al., *Seeking care for low back pain in the general population: a two-year follow-up study: results from the MUSIC-Norrtalje Study.* Spine (Phila Pa 1976), 2002. **27**(19): p. 2159-65.

11. Croft, P.R., et al., *Outcome of low back pain in general practice: a prospective study.* BMJ, 1998. **316**(7141): p. 1356-9.

12. Hestbaek, L., C. Leboeuf-Yde, and C. Manniche, *Low back pain: what is the long-term course? A review of studies of general patient populations.* Eur Spine J, 2003. **12**(2): p. 149-65.

13. Leboeuf-Yde, C., et al., *Evidence-based classification of low back pain in the general population: one-year data collected with SMS Track.* Chiropr Man Therap, 2013. **21**(1): p. 30.

14. Kjaer, P., et al., *Modic changes and their associations with clinical findings.* Eur Spine J, 2006. **15**(9): p. 1312-9.

15. Lærum, E., Brage, S., Ihlebæk, C., Johnsen, K., Natvig, B., Aas, E., *Rapport muskel skjelett sykdommer, MST-rapport 1.* FORMI (Formidlingsenheten for muskel- og skjelettlidelser), 2013.

16. van Tulder, M.W., B.W. Koes, and L.M. Bouter, *A cost-of-illness study of back pain in The Netherlands.* Pain, 1995. **62**(2): p. 233-40.

17. Martin, B.I., et al., *Expenditures and health status among adults with back and neck problems.* JAMA, 2008. **299**(6): p. 656-64.

18. Endean, A., K.T. Palmer, and D. Coggon, *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**(2): p. 160-9.

19. Balague, F., et al., *Non-specific low back pain.* Lancet, 2012. **379**(9814): p. 482-91.

20. Costa Lda, C., et al., *Prognosis for patients with chronic low back pain: inception cohort study.* BMJ, 2009. **339**: p. b3829.

21. Brage, S., et al., *[Musculoskeletal disorders as causes of sick leave and disability benefits].* Tidsskr Nor Laegeforen, 2010. **130**(23): p. 2369-70.

22. Carnes, D., et al., *Chronic musculoskeletal pain rarely presents in a single body site: results from a UK population study.* Rheumatology (Oxford), 2007. **46**(7): p. 1168-70.

23. Hagen, E.M., et al., *Comorbid subjective health complaints in low back pain.* Spine (Phila Pa 1976), 2006. **31**(13): p. 1491-5.

24. Battie, M.C., et al., *The Twin Spine Study: contributions to a changing view of disc degeneration.* Spine J, 2009. **9**(1): p. 47-59.

25. MacGregor, A.J., et al., *Structural, psychological, and genetic influences on low back and neck pain: a study of adult female twins.* Arthritis Rheum, 2004. **51**(2): p. 160-7.

26. Woolf, C.J., *Central sensitization: implications for the diagnosis and treatment of pain.* Pain, 2010. **152**(3 Suppl): p. S2-15.

27. Chou, D., et al., *Degenerative magnetic resonance imaging changes in patients with chronic low back pain: a systematic review.* Spine (Phila Pa 1976), 2011. **36**(21 Suppl): p. S43-53.

28. Modic, M.T. and J.S. Ross, *Lumbar degenerative disk disease.* Radiology, 2007. **245**(1): p. 43-61.

29. Mulholland, R.C., *The myth of lumbar instability: the importance of abnormal loading as a cause of low back pain.* Eur Spine J, 2008. **17**(5): p. 619-25.

30. Fields, A.J., E.C. Liebenberg, and J.C. Lotz, *Innervation of pathologies in the lumbar vertebral end plate and intervertebral disc.* Spine J.

31. Iordanova, E., et al., *[Long-lasting low back pain and MRI changes in the intervertebral discs].* Tidsskr Nor Laegeforen, 2010. **130**(22): p. 2260-3.

32. Shiri, R., et al., *The association between obesity and low back pain: a meta-analysis.* Am J Epidemiol, 2010. **171**(2): p. 135-54.

33. Landmark, T., et al., *Associations between recreational exercise and chronic pain in the general population: evidence from the HUNT 3 study.* Pain, 2011. **152**(10): p. 2241-7.

34. Heneweer, H., et al., *Physical activity and low back pain: a systematic review of recent literature.* Eur Spine J, 2011. **20**(6): p. 826-45.

35. Shiri, R., et al., *The association between smoking and low back pain: a meta-analysis.* Am J Med, 2010. **123**(1): p. 87 e7-35.

36. Leboeuf-Yde, C., *Back pain--individual and genetic factors.* J Electromyogr Kinesiol, 2004. **14**(1): p. 129-33.

37. Roussel, N.A., et al., *Central sensitization and altered central pain processing in chronic low back pain: fact or myth?* Clin J Pain, 2013. **29**(7): p. 625-38.

38. Carragee, E., et al., *Are first-time episodes of serious LBP associated with new MRI findings?* Spine J, 2006. **6**(6): p. 624-35.

39. Kleinstuck, F., J. Dvorak, and A.F. Mannion, *Are "structural abnormalities" on magnetic resonance imaging a contraindication to the successful conservative treatment of chronic nonspecific low back pain?* Spine (Phila Pa 1976), 2006. **31**(19): p. 2250-7.

40. Blumenthal, S., 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**(14): p. 1565-75; discussion E387-91.

41. Berg, S., et al., *Total disc replacement compared to lumbar fusion: a randomised controlled trial with 2-year follow-up.* Eur Spine J, 2009. **18**(10): p. 1512-9.

42. Hellum, C., 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**: p. d2786.

43. Gibson, J.N. and G. Waddell, *Surgery for degenerative lumbar spondylosis: updated Cochrane Review.* Spine (Phila Pa 1976), 2005. **30**(20): p. 2312-20.

44. Harrop, J.S., et al., *Lumbar adjacent segment degeneration and disease after arthrodesis and total disc arthroplasty.* Spine (Phila Pa 1976), 2008. **33**(15): p. 1701-7.

45. Jacobs, W., et al., *Total disc replacement for chronic back pain in the presence of disc degeneration.* Cochrane Database Syst Rev, 2012. **9**: p. CD008326.

46. Mannion, A.F., J.I. Brox, and J.C. Fairbank, *Comparison of spinal fusion and nonoperative treatment in patients with chronic low back pain: long-term follow-up of three randomized controlled trials.* Spine J, 2013.

47. Jensen, T.S., et al., *Vertebral endplate signal changes (Modic change): a systematic literature review of prevalence and association with non-specific low back pain.* Eur Spine J, 2008. **17**(11): p. 1407-22.

48. 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 Orthop Suppl, 2013. **84**(349): p. 1-35.

49. Leone, A., et al., *Lumbar intervertebral instability: a review.* Radiology, 2007. **245**(1): p. 62-77.

50. Rahme R , M.R., *The modic vertebral endplate and marrow changes: pathologic significance and relation to low back pain and segmental*

instability of the lumbar spine. AJNR Am J Neuroradiol., 2008. **29**(5): p. 838-42.

51. Hellum, C., et al., *Predictors of outcome after surgery with disc prosthesis and rehabilitation in patients with chronic low back pain and degenerative disc: 2-year follow-up.* Eur Spine J, 2012.

52. Middleton, K. and D.E. Fish, *Lumbar spondylosis: clinical presentation and treatment approaches.* Curr Rev Musculoskelet Med, 2009. **2**(2): p. 94-104.

53. Chou, R., et al., *Diagnosis and treatment of low back pain: a joint clinical practice guideline from the American College of Physicians and the American Pain Society.* Ann Intern Med, 2007. **147**(7): p. 478-91.

54. Henschke, N., et al., *Prevalence of and screening for serious spinal pathology in patients presenting to primary care settings with acute low back pain.* Arthritis Rheum, 2009. **60**(10): p. 3072-80.

55. ACR-ASNR-SCBT, *MR practice guideline for the performance of magnetic resonance imaging (MRI) of the adult spine.* 2012.

56. Carrino, J.A., Morrison, W.B., *Imaging of Lumbar Degenerative Disc Disease.* Seminars in Spine Surgery, 2003. **15**(4 (December)): p. 361-383.

57. Juni, P., S. Reichenbach, and P. Dieppe, *Osteoarthritis: rational approach to treating the individual.* Best Pract Res Clin Rheumatol, 2006. **20**(4): p. 721-40.

58. Adams, M.A. and P.J. Roughley, *What is intervertebral disc degeneration, and what causes it?* Spine (Phila Pa 1976), 2006. **31**(18): p. 2151-61.

59. Adams, M.A. and P. Dolan, *Intervertebral disc degeneration: evidence for two distinct phenotypes.* J Anat, 2012. **221**(6): p. 497-506.

60. Modic, M.T., et al., *Imaging of degenerative disk disease.* Radiology, 1988. **168**(1): p. 177-86.

61. Zhang, Y.H., et al., *Modic changes: a systematic review of the literature.* Eur Spine J, 2008. **17**(10): p. 1289-99.

62. Mitra, D., V.N. Cassar-Pullicino, and I.W. McCall, *Longitudinal study of vertebral type-1 end-plate changes on MR of the lumbar spine.* Eur Radiol, 2004. **14**(9): p. 1574-81.

63. Jensen, T.S., et al., *Characteristics and natural course of vertebral endplate signal (Modic) changes in the Danish general population.* BMC Musculoskelet Disord, 2009. **10**: p. 81.

64. Aprill, C. and N. Bogduk, *High-intensity zone: a diagnostic sign of painful lumbar disc on magnetic resonance imaging.* Br J Radiol, 1992. **65**(773): p. 361-9.

65. Carrino, J.A., et al., *Lumbar spine: reliability of MR imaging findings.* Radiology, 2009. **250**(1): p. 161-70.

66. Emch, T.M. and M.T. Modic, *Imaging of lumbar degenerative disk disease: history and current state.* Skeletal Radiol, 2011. **40**(9): p. 1175-89.

67. Fardon, D.F. and P.C. Milette, *Nomenclature and classification of lumbar disc pathology. Recommendations of the Combined task Forces of the North American Spine Society, American Society of Spine Radiology, and American Society of Neuroradiology.* Spine (Phila Pa 1976), 2001. **26**(5): p. E93-E113.

68. Smith, B.M., et al., *Interobserver reliability of detecting lumbar intervertebral disc high-intensity zone on magnetic resonance imaging and association of high-intensity zone with pain and anular disruption.* Spine, 1998. **23**(19): p. 2074-80.

69. Pfirrmann, C.W., et al., *Magnetic resonance classification of lumbar intervertebral disc degeneration.* Spine (Phila Pa 1976), 2001. **26**(17): p. 1873-8.

70. Luoma, K., et al., *Low back pain in relation to lumbar disc degeneration.* Spine (Phila Pa 1976), 2000. **25**(4): p. 487-92.

71. Raininko, R., et al., *Observer variability in the assessment of disc degeneration on magnetic resonance images of the lumbar and thoracic spine.* Spine (Phila Pa 1976), 1995. **20**(9): p. 1029-35.

72. Videman, T., et al., *Associations between back pain history and lumbar MRI findings.* Spine (Phila Pa 1976), 2003. **28**(6): p. 582-8.

73. Solgaard Sorensen, J., et al., *Low-field magnetic resonance imaging of the lumbar spine: reliability of qualitative evaluation of disc and muscle parameters.* Acta Radiol, 2006. **47**(9): p. 947-53.

74. Masharawi, Y., et al., *The reproducibility of quantitative measurements in lumbar magnetic resonance imaging of children from the general population.* Spine (Phila Pa 1976), 2008. **33**(19): p. 2094-100.

75. Kalichman, L. and D.J. Hunter, *Lumbar facet joint osteoarthritis: a review.* Semin Arthritis Rheum, 2007. **37**(2): p. 69-80.

76. Kettler, A. and H.J. Wilke, *Review of existing grading systems for cervical or lumbar disc and facet joint degeneration.* Eur Spine J, 2006. **15**(6): p. 705-18.

77. Fujiwara, A., et al., *The relationship between facet joint osteoarthritis and disc degeneration of the lumbar spine: an MRI study.* Eur Spine J, 1999. **8**(5): p. 396-401.

78. Siepe, C.J., 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**(22): p. 1991-2003.

79. McAfee, P.C., *The indications for lumbar and cervical disc replacement.* Spine J, 2004. **4**(6 Suppl): p. 177S-181S.

80. Jensen, T.S., J.S. Sorensen, and P. Kjaer, *Intra- and interobserver reproducibility of vertebral endplate signal (modic) changes in the lumbar spine: the Nordic Modic Consensus Group classification.* Acta Radiol, 2007. **48**(7): p. 748-54.

81. Schweitzer, M.E. and K.I. el-Noueam, *Vacuum disc: frequency of high signal intensity on T2-weighted MR images.* Skeletal Radiol, 1998. **27**(2): p. 83-6.

82. Albert, H.B., et al., *The prevalence of MRI-defined spinal pathoanatomies and their association with modic changes in individuals seeking care for low back pain.* Eur Spine J, 2011. **20**(8): p. 1355-62.

83. Kuisma, M., et al., *A three-year follow-up of lumbar spine endplate (Modic) changes.* Spine, 2006. **31**(15): p. 1714-8.

84. Modic, M.T., et al., *Degenerative disk disease: assessment of changes in vertebral body marrow with MR imaging.* Radiology, 1988. **166**(1 Pt 1): p. 193-9.

85. Pfirrmann, C.W., et al., *Effect of aging and degeneration on disc volume and shape: A quantitative study in asymptomatic volunteers.* J Orthop Res, 2006. **24**(5): p. 1086-94.

86. Shambrook, J., et al., *Clinical presentation of low back pain and association with risk factors according to findings on magnetic resonance imaging.* Pain, 2011. **152**(7): p. 1659-65.

87. Wang, Y., T. Videman, and M.C. Battie, *Modic changes: prevalence, distribution patterns, and association with age in white men.* Spine J, 2012. **12**(5): p. 411-6.

88. Jarvik, J.J., et al., *The Longitudinal Assessment of Imaging and Disability of the Back (LAIDBack) Study: baseline data.* Spine (Phila Pa 1976), 2001. **26**(10): p. 1158-66.

89. Kjaer, P., et al., *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**(10): p. 1173-80.

90. Kanayama, M., et al., *Cross-sectional magnetic resonance imaging study of lumbar disc degeneration in 200 healthy individuals.* J Neurosurg Spine, 2009. **11**(4): p. 501-7.

91. Jensen, M.C., et al., *Magnetic resonance imaging of the lumbar spine in people without back pain.* N Engl J Med, 1994. **331**(2): p. 69-73.

92. Boden, S.D., et al., *Abnormal magnetic-resonance scans of the cervical spine in asymptomatic subjects. A prospective investigation.* J Bone Joint Surg Am, 1990. **72**(8): p. 1178-84.

93. Kaapa, E., et al., *Correlation of size and type of modic types 1 and 2 lesions with clinical symptoms: a descriptive study in a subgroup of patients with chronic low back pain on the basis of a university hospital patient sample.* Spine (Phila Pa 1976), 2012. **37**(2): p. 134-9.

94. Sambrook, P.N., A.J. MacGregor, and T.D. Spector, *Genetic influences on cervical and lumbar disc degeneration: a magnetic resonance imaging study in twins.* Arthritis Rheum, 1999. **42**(2): p. 366-72.

95. Eskola, P.J., et al., *Genetic association studies in lumbar disc degeneration: a systematic review.* PLoS One, 2012. **7**(11): p. e49995.

96. Battie, M.C., T. Videman, and E. Parent, *Lumbar disc degeneration: epidemiology and genetic influences.* Spine (Phila Pa 1976), 2004. **29**(23): p. 2679-90.

97. Kvistad, K.A. and A. Espeland, *[Diagnostic imaging in neck and low back pain].* Tidsskr Nor Laegeforen, 2010. **130**(22): p. 2256-9.

98. Schenk, P., et al., *Magnetic resonance imaging of the lumbar spine: findings in female subjects from administrative and nursing professions.* Spine (Phila Pa 1976), 2006. **31**(23): p. 2701-6.

99. Samartzis, D., et al., *The association of lumbar intervertebral disc degeneration on magnetic resonance imaging with body mass index in overweight and obese adults: a population-based study.* Arthritis Rheum, 2012. **64**(5): p. 1488-96.

100. van Tulder, M.W., et al., *Spinal radiographic findings and nonspecific low back pain. A systematic review of observational studies.* Spine (Phila Pa 1976), 1997. **22**(4): p. 427-34.

101. de Schepper, E.I., 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**(5): p. 531-6.

102. Jarvik, J.G. and R.A. Deyo, *Moderate versus mediocre: the reliability of spine MR data interpretations.* Radiology, 2009. **250**(1): p. 15-7.

103. Pereira-Maxwell, F., *A- Z of Medical statistics, a companion of critical appraisal*. 1998, reprinted 2006, London: Arnold.

104. Sim, J. and C.C. Wright, *The kappa statistic in reliability studies: use, interpretation, and sample size requirements.* Phys Ther, 2005. **85**(3): p. 257-68.

105. Berlin, L., *Accuracy of diagnostic procedures: has it improved over the past five decades?* AJR Am J Roentgenol, 2007. **188**(5): p. 1173-8.

106. Robinson, P.J., *Radiology's Achilles' heel: error and variation in the interpretation of the Rontgen image.* Br J Radiol, 1997. **70**(839): p. 1085-98.

107. Rigby, A.S., *Statistical methods in epidemiology. v. Towards an understanding of the kappa coefficient.* Disabil Rehabil, 2000. **22**(8): p. 339-44.

108. Ker, M., *Issues in the Use of Kappa.* Investigative Radiology, 1991. **26**(1): p. 78-83.

109. Altman, D., *Practical statistics for medical research*. 1991: New York: Chapman& Hall-CRC.

110. Banerjee M, C.M., McSweeney L, Sinha D *Beyond Kappa: A Review of Interrater Agreement Measures.* The Canadian Journal of Statistics, 1999. **Vol.27**(No.1): p. pp. 3-23.

111. Hutton, M.J., J.H. Bayer, and J.M. Powell, *Modic vertebral body changes: the natural history as assessed by consecutive magnetic resonance imaging.* Spine (Phila Pa 1976), 2011. **36**(26): p. 2304-7.

112. Carragee, E., *Surgical treatment of lumbar disk disorders.* JAMA, 2006. **296**(20): p. 2485-7.

113. Park, C.K., K.S. Ryu, and W.H. Jee, *Degenerative changes of discs and facet joints in lumbar total disc replacement using ProDisc II: minimum two-year follow-up.* Spine (Phila Pa 1976), 2008. **33**(16): p. 1755-61.

114. Hollingworth, W., et al., *The diagnostic and therapeutic impact of MRI: an observational multi-centre study.* Clin Radiol, 2000. **55**(11): p. 825-31.

115. Brealey, S. and M. Westwood, *Are you reading what we are reading? The effect of who interprets medical images on estimates of diagnostic test accuracy in systematic reviews.* Br J Radiol, 2007. **80**(956): p. 674-7.

116. Kovacs, F.M., et al., *Agreement in the interpretation of magnetic resonance images of the lumbar spine.* Acta Radiol, 2009. **50**(5): p. 497-506.

117. Arana, E., et al., *Lumbar spine: agreement in the interpretation of 1.5-T MR images by using the nordic modic consensus group classification form.* Radiology, 2010. **254**(3): p. 809-17.

118. Peterson, C.K., et al., *Inter- and intraexaminer reliability in identifying and classifying degenerative marrow (Modic) changes on lumbar spine magnetic resonance scans.* J Manipulative Physiol Ther, 2007. **30**(2): p. 85-90.

119. Lurie, J.D., et al., *Reliability of magnetic resonance imaging readings for lumbar disc herniation in the Spine Patient Outcomes Research Trial (SPORT).* Spine (Phila Pa 1976), 2008. **33**(9): p. 991-8.

120. Stieber, J., et al., *The reliability of computed tomography and magnetic resonance imaging grading of lumbar facet arthropathy in total disc replacement patients.* Spine (Phila Pa 1976), 2009. **34**(23): p. E833-40.

121. Luoma, K., et al., *MRI follow-up of subchondral signal abnormalities in a selected group of chronic low back pain patients.* Eur Spine J, 2008. **17**(10): p. 1300-8.

122. Luoma, K., et al., *Relationship of Modic type 1 change with disc degeneration: a prospective MRI study.* Skeletal Radiol, 2009. **38**(3): p. 237-44.

123. Sharma, A., T. Pilgram, and F.J. Wippold, 2nd, *Association between annular tears and disk degeneration: a longitudinal study.* AJNR Am J Neuroradiol, 2009. **30**(3): p. 500-6.

124. Videman, T., et al., *Magnetic resonance imaging findings and their relationships in the thoracic and lumbar spine. Insights into the etiopathogenesis of spinal degeneration.* Spine (Phila Pa 1976), 1995. **20**(8): p. 928-35.

125. Mariconda, M., et al., *Relationship between alterations of the lumbar spine, visualized with magnetic resonance imaging, and occupational variables.* Eur Spine J, 2007. **16**(2): p. 255-66.

126. Arana, E., et al., *Relationship between low back pain, disability, MR imaging findings and health care provider.* Skeletal Radiol, 2006. **35**(9): p. 641-7.

127. Hellum, C., *Patients with chronic low back pain and degenerative disc; effects of suregery with disc prosthesis versus rehabilitation, predictors of treatment outcome and health economic considerations.* PhD Thesis, Faculty of Medicine, University of Oslo and Department of Orthopaedics, Oslo University Hospital Ullevål 2013.

128. Hellum, C., 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.

129. Fairbank, J.C. and P.B. Pynsent, *The Oswestry Disability Index.* Spine (Phila Pa 1976), 2000. **25**(22): p. 2940-52; discussion 2952.

130. Grotle, M., J.I. Brox, and N.K. Vollestad, *Cross-cultural adaptation of the Norwegian versions of the Roland-Morris Disability Questionnaire and the Oswestry Disability Index.* J Rehabil Med, 2003. **35**(5): p. 241-7.

131. Smeets, R., et al., *Measures of function in low back pain/disorders: Low Back Pain Rating Scale (LBPRS), Oswestry Disability Index (ODI), Progressive Isoinertial Lifting Evaluation (PILE), Quebec Back Pain Disability Scale (QBPDS), and Roland-Morris Disability Questionnaire (RDQ).* Arthritis Care Res (Hoboken), 2011. **63 Suppl 11**: p. S158-73.

132. Chapman, J.R., et al., *Evaluating common outcomes for measuring treatment success for chronic low back pain.* Spine (Phila Pa 1976), 2011. **36**(21 Suppl): p. S54-68.

133. Jensch, S., et al., *CT colonography with limited bowel preparation: performance characteristics in an increased-risk population.* Radiology, 2008. **247**(1): p. 122-32.

134. Mak, H.K., et al., *Hypodensity of >1/3 middle cerebral artery territory versus Alberta Stroke Programme Early CT Score (ASPECTS): comparison of two methods of quantitative evaluation of early CT changes in hyperacute ischemic stroke in the community setting.* Stroke, 2003. **34**(5): p. 1194-6.

135. Leren, P., et al., *The Oslo study. Cardiovascular disease in middle-aged and young Oslo men.* Acta Med Scand Suppl, 1975. **588**: p. 1-38.

136. Browner, W., Newman, T.B., Hulley, S., *Estimating sample size and power: Applications and examples* in *Designing clinical research, Third edition*. 2007, Wolters Kluwer Health/ Lippingcott Williams & Wilkins: Philadelphia, PA 19106 USA.

137. Fritzell, P., et al., *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 (Phila Pa 1976), 2001. **26**(23): p. 2521-32; discussion 2532-4.

138. Brant-Zawadzki, M.N., et al., *Interobserver and intraobserver variability in interpretation of lumbar disc abnormalities. A comparison of two nomenclatures.* Spine (Phila Pa 1976), 1995. **20**(11): p. 1257-63; discussion 1264.

139. Jones, A., et al., *The Modic classification: inter- and intraobserver error in clinical practice.* Spine, 2005. **30**(16): p. 1867-9.

140. Lucas, N.P., et al., *The development of a quality appraisal tool for studies of diagnostic reliability (QAREL).* J Clin Epidemiol, 2010. **63**(8): p. 854-61.

141. Obuchowski, N.A., *Special Topics III: bias.* Radiology, 2003. **229**(3): p. 617-21.

142. Park, C.O., *Diurnal variation in lumbar MRI. Correlation between signal intensity, disc height, and disc bulge.* Yonsei Med J, 1997. **38**(1): p. 8-18.

143. Westbrook, C., Kaut, C., *MRI in practice*. 1998, Second edition, London, United Kingdom: Blackwell Science Ltd.

144. Kuhl, C.K., F. Traber, and H.H. Schild, *Whole-body high-field-strength (3.0-T) MR Imaging in Clinical Practice. Part I. Technical considerations and clinical applications.* Radiology, 2008. **246**(3): p. 675-96.

145. Maubon, A.J., et al., *Effect of field strength on MR images: comparison of the same subject at 0.5, 1.0, and 1.5 T.* Radiographics, 1999. **19**(4): p. 1057-67.

146. Taber, K.H., et al., *Pitfalls and artifacts encountered in clinical MR imaging of the spine.* Radiographics, 1998. **18**(6): p. 1499-521.

147. Kuhl, C.K., et al., *Whole-body high-field-strength (3.0-T) MR imaging in clinical practice. Part II. Technical considerations and clinical applications.* Radiology, 2008. **247**(1): p. 16-35.

148. Bendix, T., et al., *Lumbar modic changes-a comparison between findings at low- and high-field magnetic resonance imaging.* Spine (Phila Pa 1976), 2012. **37**(20): p. 1756-62.

149. Bitar, R., et al., *MR pulse sequences: what every radiologist wants to know but is afraid to ask.* Radiographics, 2006. **26**(2): p. 513-37.

150. Melhem, E.R., et al., *MR of the spine with a fast T1-weighted fluid-attenuated inversion recovery sequence.* AJNR Am J Neuroradiol, 1997. **18**(3): p. 447-54.

151. Lavdas, E., et al., *Comparison of T1-weighted fast spin-echo and T1-weighted fluid-attenuated inversion recovery images of the lumbar spine at 3.0 Tesla.* Acta Radiol, 2010. **51**(3): p. 290-5.

152. Kuisma, M., et al., *Modic changes in endplates of lumbar vertebral bodies: prevalence and association with low back and sciatic pain among middle-aged male workers.* Spine (Phila Pa 1976), 2007. **32**(10): p. 1116-22.

153. Lighvani, A.A.a.M., E.R., *Advances in high-field MR imaging of the spine.* Appl Radiol. , 2009. **38**(6): p. 18-27.

154. Gold, G.E., et al., *Driven equilibrium magnetic resonance imaging of articular cartilage: initial clinical experience.* J Magn Reson Imaging, 2005. **21**(4): p. 476-81.

155. Melhem, E.R., R. Itoh, and P.J. Folkers, *Cervical spine: three-dimensional fast spin-echo MR imaging--improved recovery of longitudinal magnetization with driven equilibrium pulse.* Radiology, 2001. **218**(1): p. 283-8.

156. Bae, W.C., et al., *Conventional and ultrashort time-to-echo magnetic resonance imaging of articular cartilage, meniscus, and intervertebral disk.* Top Magn Reson Imaging, 2010. **21**(5): p. 275-89.

157. Khan, I., R. Hargunani, and A. Saifuddin, *The lumbar high-intensity zone: 20 years on.* Clin Radiol, 2014. **69**(6): p. 551-558.

158. Zook, J., et al., *Inter- and intraobserver reliability in radiographic assessment of degenerative disk disease.* Orthopedics, 2011. **34**(4).

159. Liu, C., et al., *Quantitative estimation of the high-intensity zone in the lumbar spine: comparison between the symptomatic and asymptomatic population.* Spine J, 2014. **14**(3): p. 391-6.

160. Weishaupt, D., et al., *MR imaging and CT in osteoarthritis of the lumbar facet joints.* Skeletal Radiol, 1999. **28**(4): p. 215-9.

161. Lee, J.C., et al., *Radiographic grading of facet degeneration, is it reliable? - a comparison of MR or CT grading with histologic grading in lumbar fusion candidates.* Spine J, 2012. **12**(6): p. 507-14.

162. Bankier, A.A., et al., *Consensus interpretation in imaging research: is there a better way?* Radiology, 2010. **257**(1): p. 14-7.

163. Espeland, A., N. Vetti, and J. Krakenes, *Are two readers more reliable than one? A study of upper neck ligament scoring on magnetic resonance images.* BMC Med Imaging, 2013. **13**: p. 4.

164. Newman, T., Browner, W., Hulley, S., *Enhancing causal interference in observational studies*, in *Designing clinical research.* , S. Hulley, Cummings, S., Browner, W., Grady, D., Newman, T., Editor. 2007, Lippincott Williams & Wilkins: Philadelphia, USA. p. p. 127-146

165. Von Korff, M., M.P. Jensen, and P. Karoly, *Assessing global pain severity by self-report in clinical and health services research.* Spine (Phila Pa 1976), 2000. **25**(24): p. 3140-51.

166. Carragee, E.J., et al., *Discographic, MRI and psychosocial determinants of low back pain disability and remission: a prospective study in subjects with benign persistent back pain.* Spine J, 2005. **5**(1): p. 24-35.

167. Rigby, A.S., *Statistical methods in epidemiology: I. Statistical errors in hypothesis testing.* Disabil Rehabil, 1998. **20**(4): p. 121-6.

168. Weishaupt, D., et al., *MR imaging of the lumbar spine: prevalence of intervertebral disk extrusion and sequestration, nerve root compression, end plate abnormalities, and osteoarthritis of the facet joints in asymptomatic volunteers.* Radiology, 1998. **209**(3): p. 661-6.

169. Jensen, R.K., et al., *Can pathoanatomical pathways of degeneration in lumbar motion segments be identified by clustering MRI findings.* BMC Musculoskelet Disord, 2013. **14**(1): p. 198.

170. Wang, Y., et al., *Quantitative measures of modic changes in lumbar spine magnetic resonance imaging: intra- and inter-rater reliability.* Spine (Phila Pa 1976), 2011. **36**(15): p. 1236-43.

171. Battie, M.C. and T. Videman, *Lumbar disc degeneration: epidemiology and genetics.* J Bone Joint Surg Am, 2006. **88 Suppl 2**: p. 3-9.

172. Takatalo, J., et al., *Does lumbar disc degeneration on magnetic resonance imaging associate with low back symptom severity in young Finnish adults?* Spine (Phila Pa 1976), 2011. **36**(25): p. 2180-9.

173. Kirkwood, B.R., Sterne J.A.C., *Essentials Medical Statistics*. 2003, Second edition, Massachusetts, USA: Blackwell science Ltd.

174. Pande, K.C., K. Khurjekar, and V. Kanikdaley, *Correlation of low back pain to a high-intensity zone of the lumbar disc in Indian patients.* J Orthop Surg (Hong Kong), 2009. **17**(2): p. 190-3.

175. Saywell, W.R., et al., *Demonstration of vertebral body end plate veins by magnetic resonance imaging.* Br J Radiol, 1989. **62**(735): p. 290-2.

176. Fu, M.C., et al., *Interrater and intrarater agreements of magnetic resonance imaging findings in the lumbar spine: significant variability across degenerative conditions.* Spine J, 2014.

177. Boos, N., et al., *1995 Volvo Award in clinical sciences. The diagnostic accuracy of magnetic resonance imaging, work perception, and psychosocial factors in identifying symptomatic disc herniations.* Spine (Phila Pa 1976), 1995. **20**(24): p. 2613-25.

178. Mulconrey, D.S., et al., *Interobserver reliability in the interpretation of diagnostic lumbar MRI and nuclear imaging.* Spine J, 2006. **6**(2): p. 177-84.

179. Milette, P.C., et al., *Differentiating lumbar disc protrusions, disc bulges, and discs with normal contour but abnormal signal intensity. Magnetic resonance imaging with discographic correlations.* Spine (Phila Pa 1976), 1999. **24**(1): p. 44-53.

180. Goddard, P., et al., *Error in radiology.* Br J Radiol, 2001. **74**(886): p. 949-51.

181. White, K., K. Berbaum, and W.L. Smith, *The role of previous radiographs and reports in the interpretation of current radiographs.* Invest Radiol, 1994. **29**(3): p. 263-5.

182. Jensen, R.K. and C. Leboeuf-Yde, *Is the presence of Modic changes associated with the outcomes of different treatments? A systematic critical review.* BMC Musculoskelet Disord, 2011. **12**: p. 183.

183. Sears, W.R., et al., *Incidence and prevalence of surgery at segments adjacent to a previous posterior lumbar arthrodesis.* Spine J, 2010. **11**(1): p. 11-20.

184. Lund, T. and T.R. Oxland, *Adjacent Level Disk Disease-Is it Really a Fusion Disease?* Orthop Clin North Am, 2011. **42**(4): p. 529-41.

185. Snoey, E.R., et al., *Analysis of emergency department interpretation of electrocardiograms.* J Accid Emerg Med, 1994. **11**(3): p. 149-53.

186. Fedorak, C., et al., *Reliability of the visual assessment of cervical and lumbar lordosis: how good are we?* Spine (Phila Pa 1976), 2003. **28**(16): p. 1857-9.

187. Poiraudeau, S., et al., *Value of the bell test and the hyperextension test for diagnosis in sciatica associated with disc herniation: comparison with Lasegue's sign and the crossed Lasegue's sign.* Rheumatology (Oxford), 2001. **40**(4): p. 460-6.

188. Marshman, L.A., et al., *Minimizing ferromagnetic artefact with metallic lumbar total disc arthroplasty devices at adjacent segments: technical note.* Spine (Phila Pa 1976), 2010. **35**(2): p. 252-6.

189. Videman, T. and M. Nurminen, *The occurrence of anular tears and their relation to lifetime back pain history: a cadaveric study using barium sulfate discography.* Spine (Phila Pa 1976), 2004. **29**(23): p. 2668-76.

190. Albert, H.B. and C. Manniche, *Modic changes following lumbar disc herniation.* Eur Spine J, 2007. **16**(7): p. 977-82.

191. Braithwaite, I., et al., *Vertebral end-plate (Modic) changes on lumbar spine MRI: correlation with pain reproduction at lumbar discography.* Eur Spine J, 1998. **7**(5): p. 363-8.

192. Cvitanic, O.A., et al., *Subchondral marrow changes after laser diskectomy in the lumbar spine: MR imaging findings and clinical correlation.* AJR Am J Roentgenol, 2000. **174**(5): p. 1363-9.

193. Lim, C.H., et al., *Discogenic lumbar pain: association with MR imaging and CT discography.* Eur J Radiol, 2005. **54**(3): p. 431-7.

194. Weishaupt, D., et al., *Painful Lumbar Disk Derangement: Relevance of Endplate Abnormalities at MR Imaging.* Radiology, 2001. **218**(2): p. 420-7.

195. Ito, M., et al., *Predictive signs of discogenic lumbar pain on magnetic resonance imaging with discography correlation.* Spine (Phila Pa 1976), 1998. **23**(11): p. 1252-8; discussion 1259-60.

196. Kokkonen, S.M., et al., *Endplate degeneration observed on magnetic resonance imaging of the lumbar spine: correlation with pain provocation and disc changes observed on computed tomography diskography.* Spine (Phila Pa 1976), 2002. **27**(20): p. 2274-8.

197. Suri, P., et al., *Presence and extent of severe facet joint osteoarthritis are associated with back pain in older adults.* Osteoarthritis Cartilage, 2013. **21**(9): p. 1199-206.

198. Dai, L., *Disc degeneration and cervical instability. Correlation of magnetic resonance imaging with radiography.* Spine (Phila Pa 1976), 1998. **23**(16): p. 1734-8.

199. Stein, B.E., et al., *Detection of acromioclavicular joint pathology in asymptomatic shoulders with magnetic resonance imaging.* J Shoulder Elbow Surg, 2001. **10**(3): p. 204-8.

200. Cheung, K.M., 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**(9): p. 934-40.

201. Dionne, C.E., et al., *A consensus approach toward the standardization of back pain definitions for use in prevalence studies.* Spine (Phila Pa 1976), 2008. **33**(1): p. 95-103.

202. Deyo, R.A. and Y.J. Tsui-Wu, *Descriptive epidemiology of low-back pain and its related medical care in the United States.* Spine (Phila Pa 1976), 1987. **12**(3): p. 264-8.

203. Lawrence, R.C., et al., *Estimates of the prevalence of arthritis and selected musculoskeletal disorders in the United States.* Arthritis Rheum, 1998. **41**(5): p. 778-99.

204. Nystrom, B., *Spinal fusion in the treatment of chronic low back pain: rationale for improvement.* Open Orthop J, 2012. **6**: p. 478-81.

205. Bailly, F., et al., *Inflammatory pain pattern and pain with lumbar extension associated with Modic 1 changes on MRI: a prospective case-control study of 120 patients.* Eur Spine J. **23**(3): p. 493-7.

206. Arendt-Nielsen, L., et al., *Sensitization in patients with painful knee osteoarthritis.* Pain, 2010. **149**(3): p. 573-81.

207. Kovacs, F.M., et al., *Disc degeneration and chronic low back pain: an association which becomes nonsignificant when endplate changes and disc contour are taken into account.* Neuroradiology, 2013.

208. Borenstein, D.G., 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**(9): p. 1306-11.

Appendix

Appendix Table 1 Studies on reliability of lumbar degenerative magnetic resonance imaging (MRI) findings. Inter/intra = numbers of individuals or results for interobserver/intraobserver reliability. HIZ = high-intensity zone in the posterior disc, DD = disc degeneration, CC = craniocaudal, AP = anteroposterior, FA = facet arthropathy, ICC = intraclass correlation coefficient

Author et al, year	**Sample inter/intra**	**Observers inter/intra**	**MRI variables**	**Kappa inter/intra**	**Comments**
Fu 2014 [176]	75/10	4/4 2 orthopaedic spine surgeons 2 musculoskeletal radiologists	• Disc hydration • Disc space height • Disc herniation and canal compression • FA present or not	• 0.38 • 0.45 • 0.28-0.37 • 0.44	Prevalence and bias not reported. Clustering not taken into account (multiple levels evaluated in each patient). Type of kappa not reported (weighted or unweighted). 1.5 T or 3.0 T MRI
Wang 2011 [170]	83/83 (913 endplates)	2/1 Orthopaedic surgeon and spine researcher with experience in spine MRI evaluation	• Modic type Quantitative dimension measures of Modic changes • Width ratio • Height ratio • Area ratio	• 0.79/0.88 ICC: • 0.66-0.73/0.82-0.83 • 0.82-0.88/0.90 • 0.75-0.81/0.90-0.93	Prevalence and bias not reported. Clustering not taken into account (multiple levels evaluated in each patient). Type of kappa not reported (unweighted or weighted). 1.5 T MRI

Zook 2011 [158]	71/71	3/3 Fellowship-trained spine surgeons	Presence or not of: • Modic changes • HIZ • Reduced disc signal • Disc herniation Continuous measure: • Disc height	• 0.65/0.74-0.95 • 0.82/0.68-0.85 • 0.92/0.90-0.91 • 0.51/0.39-0.73 • 0.58/0.77-0.99	Prevalence and bias not reported. Number of rated levels not reported, and clustering may exist (multiple levels evaluated in each patient). Type of kappa not reported (unweighted or weighted). Magnet strength not reported
Arana 2010 [117]	53/53	5/5 Radiologists	• Modic changes • HIZ • DD (Pfirrmann grading)[a)] • Disc contour	• 0.53/0.72 • 0.60/0.73 • 0.49/0.69 • 0.55/0.69	Only finding with prevalence 10-90 % analysed with kappa. Bias not analysed or reported. Mean kappa for dichotomised finding normal/abnormal all levels together, but clustering of data taken into account. 1.5 T MRI
Kovacs 2009 [116]	50/50	7/6 Radiologists	• Modic changes • HIZ • DD (Pfirrmann grading)[a)] • Disc contour	• 0.83/0.82 • 0.62/0.72 • 0.22/0.60 • 0.73/0.71	Only finding with prevalence 10-90 % analysed with kappa. Bias not analysed or reported. Clustering of data taken into account. Mean kappa for dichotomised finding normal/abnormal at each level. 0.2 T MRI

Carrino 2009 [65]	111/40	4 3 musculoskeletal radiologists, 1 orthopaedic spine surgeon experienced in spine interpretation	• Modic changes CC extent Superior Inferior AP extent Superior Inferior • HIZ • DD (Pfirrmann grading)[a)] • FA (4 grades based on illustrations)	• 0.59/0.64 • 0.47/0.50 • 0.48/0.60 • 0.43/0.54 • 0.57/0.60 • 0.44/0.67 • 0.66/0.74 • 0.54/0.69	Prevalence, bias and clustering taken into account; conditional logistic regression for systematic differences in reported frequency between observers and bootstrapping for clustering of data. Weighted kappa (linear). 1.5 T MRI
Stieber 2009 [120]	10/10	13/13 10 fellowship-trained orthopaedic surgeons and 3 orthopaedic spine fellows	• FA (Fujiwara grading)[b)]	• 0.21/0.36 for surgeons (mean?) • 0.07/0.26 for fellows (mean)	Prevalence and bias not reported. Not reported whether three lowest lumbar levels added together in analyses (cluster). Type of kappa not reported (unweighted or weighted). Magnet strength not reported
Peterson 2007 [118]	51/51	2/2 Experienced radiologists	• Modic changes	• Presence Entire lumbar spine 0.52/0.70-0.86 L3/L4 0.66 (inter, low prevalence) L4/L5 0.81 (inter) L5/S1 0.58 (inter)	Prevalence reported and effect on kappa magnitude discussed. Bias not reported. Probably not clustering. Unweighted kappa for presence of Modic changes. 0.6 T MRI

Jensen 2007 [80]	50/50	2/1 1 radiologist, 1 chiropractor	• Modic changes per endplate (normal, type I, II or III) • CC extent of Modic changes (5 grades from normal to more than 50 % of vertebral body height)	• 0.80/0.82 • 0.80/0.83	Only finding with prevalence 10-90 % analysed with kappa. Sample chosen so that Modic prevalence was 54 %. Bias not reported. Clustering not taken into account (multiple levels from same patient added together). Linearly weighted kappa, except for Modic type. Observers trained together. 0.2 T MRI
Solgaard-Soerensen 2006 [73]	50/50	2/1 Radiologists	• HIZ • Disc signal intensity • Disc height reduction • Disc protrusion	• 0.86/0.97 (unweighted kappa) • 0.59/0.87 (linear weights) • 0.66/0.81 (linear weights) • 0.68/0.78 (linear weights)	Only finding with prevalence 10-90 % analysed. Bias not reported. Clustering not taken into account (all levels with prevalence 10-90 % added together). Common training: 50 MRIs in consensus (not used in study). 0.2 T MRI
Mulconrey 2006 [178]	17 (80 lumbar levels)	4 Orthopaedic surgery resident, radiology resident, nuclear medicine radiologist, spine surgeon	• Modic changes present or not • Disc signal intensity reduction present or not	• Inter 0.31-0.85 • Inter 0.65-0.85	Prevalence and bias not reported. Clustering not taken into account. Type of kappa not reported (unweighted or weighted). 1.5 T MRI

Jones 2005 [139]	50/50	5/5 Spine surgeons with different MRI experience	• Modic type	• 0.85 (mean?)/0.71-1.00	Prevalence not reported or taken into account. Bias not reported. No clustering; only one level evaluated in each patient. Type of kappa not reported (unweighted or weighted). Magnet strength not reported
Jarvik 2001 [88]	10 (50 discs)	2 Radiologists	• Annular tear • Disc signal intensity (normal, mild, moderate, severe) • Disc height reduction • Disc contour (normal, bulge, protrusion, extrusion)	• Inter 0.54 • Inter 0.84 • Inter 0.56 • Inter 0.68	Prevalence and bias not reported. Clustering not accounted for (5 levels added in each patient). Kappa unweighted for annular tear and disc height decrease (dichotomized to presence or not) and weighted for other MRI variables (ordinal). 1.5 T MRI
Pfirrmann 2001 [69]	60/60 (300 discs)	3/3 1 orthopaedic surgeon, 1 fellowship trained musculoskeletal radiologist and 1 musculoskeletal senior radiologist	• DD (Pfirrmann grading)[a)]	• 0.74-0.81/0.84-0.90	Prevalence and bias reported, but not discussed in relation to kappa magnitude. Clustering not taken into account. Type of kappa not reported (ordinary or weighted) 1.0 T MRI

Luoma 2000 [70]	164	3 Radiologists	• Disc signal intensity	• Inter 0.59-0.83	Prevalence and bias not reported. No clustering (individual disc levels L2-S1 analysed). Weighted kappa. 0.1 T MRI
Weishaupt 1999 [160]	50/50	2/2 Radiologists	• FA (Based on Pathria grading)[c)]	• 0.41/0.70 and 0.76	Prevalence and bias not reported. Clustering not taken into account (facet joints from L1 to S1 added together). Weighted kappa, weighting not reported. 1.0 T MRI
Sambrook 1999 [94]	65/65	3/3 2 radiologists, 1 rheumatologist	• Disc signal intensity • Disc height reduction • Disc protrusion	• 0.52/0.75 (mean?) • 0.70/0.51 (mean?) • 0.72/0.70 (mean?)	Prevalence and bias not reported. Method of analysis not well described, and probably clustering of data existed and was not accounted for. 4 point grading system, weighted kappa (weights not reported). 1.0 T or 1.5 T MRI

Milette 1999 [179]	45	2 Neuroradiologists	• Disc signal intensity - central - peripheral • Disc height • Disc contour	Unweighted kappa for normal or not: - Inter 0.81 - Inter 0.75 • Inter 0.74 • Inter 0.59 (weighted kappa 0.66 for normal, bulge, protrusion)	Prevalence reported, but effect on kappa not discussed. Bias partly reported, but not discussed. Clustering not taken into account (132 disc levels analysed in 45 patients). 1.5 T MRI
Smith 1998 [68]	72	2 Neuroradiologists	• HIZ	• Inter 0.57	Prevalence reported (8.6-11 %) and effect on kappa discussed. Bias reported in 2x2 contingency table (11 vs. 22 in cells b vs. c) but not discussed. Clustering of data not taken into account (multiple levels added). 1.5 T MRI
Boos 1995 [177]	46	2 1 radiologist, 1 neuroradiologist	• Disc signal intensity • Disc herniation	• Inter 0.85-0.93 individual levels L1-S1 • Inter 0.66-0.77 individual levels L1-S1	Prevalence reported, but effect on kappa not discussed. Bias not reported or taken into account. No clustering. Type of kappa not reported (unweighted or weighted). Common training in advance. 1.5 T MRI

Brant-Zawadzki 1995 [138]	125/88	2/2 Experienced neuroradiologists	• Disc contour (normal versus abnormal)	• 0.67/0.79 and 0.75 (unweighted kappa)	Prevalence and bias reported in contingency table, but influence on kappa not discussed. Clustering not taken into account (5 discs combined). 1.5 T MRI
Raininko 1995 [71]	122/20	3/3 1 MRI experienced orthopaedic surgeon, 1 neuroradiologist, and 1 general radiologist with less experience in spinal MRI	• Modic type II ○ Superior ○ Inferior to the disc • Disc herniation	 • 0.64/0.87 • 0.53/0.79 • 0.30/0.70 (low prevalence)	Prevalence and bias reported and taken into account. Clustering not taken into account (multiple levels added). Mean kappa for normal vs. abnormal for each variable. 1.5 T MRI

Kappa is defined as the difference between observed and expected agreement (by chance) expressed as a fraction of the maximum possible difference. Kappa = (observed agreement - expected agreement) / (1 - expected agreement) [104].

a) Pfirrmann grades of disc degeneration based on nucleus pulposus appearance and disc height combined [69]:

I: nucleus pulposus homogeneous bright white with clear distinction of the nucleus and the annulus, and normal height of the intervertebral disk.

II: nucleus pulposus inhomogeneous, with or without horizontal bands.

III: nucleus pulposus inhomogeneous with intermediate signal intensity (grey), clear distinction of the nucleus and the annulus, height of the intervertebral disk normal to slightly decreased.

IV: nucleus pulposus inhomogeneous with low to intermediate signal intensity (dark), distinction of the nucleus and the annulus lost, height of the intervertebral disk normal to moderately decreased.

V: nucleus pulposus hypointense (black) either inhomogeneous or homogeneous, distinction of the nucleus and the annulus lost, height of the intervertebral disk indicates a collapsed disk space.

b) Fujiwara grades of FA on MRI [77]:

1: normal, 2: joint space narrowing or mild osteophyte, 3: sclerosis or moderate osteophyte, 4: marked osteophyte.

c) Weishaupt [70] grading of FA on CT and MRI: 0: normal facet joint space (2-4 mm width), 1: joint space narrowing (< 2mm) and/or small osteophytes and/or mild hypertrophy of the articular process, 2: narrowing of the facet joint space and/or moderate osteophytes and/or moderate hypertrophy of the articular process and/or moderate subarticular bone erosions, 3: narrowing of the facet joint space and/or large osteophytes and/or severe hypertrophy of the articular process and/or severe subarticular bone erosions and/or subchondral cysts.

Weishaupt grading of FA is based on Pathria grades of FA on oblique conventional radiographs and CT scans [76]: 0: normal, 1: joint space narrowing (mild degenerative disease), 2: Narrowing plus sclerosis or hypertrophy (moderate degenerative disease), 3: Severe osteoarthrosis with narrowing, sclerosis, and osteophytes (severe degenerative disease).

Appendix Table 2 Studies on the association between selected degenerative magnetic resonance imaging (MRI) findings and low back pain (LBP). HIZ = high-intensity zone in the posterior disc, FA = facet arthropathy, OR = odds ratio, CI = confidence interval.

Author and year	**Study design and sample**	**LBP variables**	**MRI findings**	**Results for association between MRI findings and LBP**	**Comments**
Chou 2011 [27]	Systematic review Adults with non-radicular LBP for > 3 months and possible degenerative cause of LBP	LBP ≥ 3 months	• Modic changes • Annular tear • HIZ • Disc signal intensity • Disc height reduction • Disc protrusion • FA • Endplate irregular	OR (95 % CI): • 4.2 (2.1-9.2) • 2.0 (1.3-3.3) • 2.5 (1.5-4.0) • 1.8-2.8 (range) • 2.5 (1.6-4.0) • 1.3-3.2 (range) • 1.1 (0.7-1.7) • 0.9 (0.6-1.5)	Heterogeneous LBP definition and leg pain poorly reported. Annular tear and HIZ not defined and results reported separately for HIZ and annular tear. Cross sectional studies. Crude OR. ≤ 0.2 T MRI, but 1.5 T MRI in one study

Endean 2011 [18]	Systematic review	LBP past 12 months	• HIZ/annular tear • Disc degeneration • Disc protrusion • FA • Endplate defects	OR (95 % CI): • 2.5 (1.6-3.9) in one study and 4.6 (1.9-11.1) in another study • 2.5 (2.0-7.4) • 3.6 (1.8-7.0) • 1.1 (0.7-1.6) in one study and 4.4 (0.9-21) in another study • 0.9 (0.6-1.5)	LBP not well defined and both with and without radicular pain. Disc degeneration variably defined as reduced disc height and/or reduced disc signal intensity. OR is a meta-estimate or referred from reviewed papers if meta-estimate not possible.
Jensen 2008 [47]	Systematic review	Non-specific LBP	• Modic changes	• Median OR 3.4, range 2.0-19.9 • Positive association in 7 of 10 reviewed studies	LBP not well defined and both with and without radicular pain. Most patients had long-lasting LBP

Liu 2014 [159]	Case-control 72 cases and 79 controls Age 24-59 years	LBP > 6 months Controls had never had any relevant LBP or related complaints (e.g. no medical consultation or work absence because of LBP)	• HIZ • Quantitative dimension (area of HIZ) and signal intensity measures	• Prevalence of HIZ 45.8 % in cases and 20.2 % in controls • Mean signal intensity of HIZ higher in cases (*P* < 0.001) - No significant difference in dimension of HIZ between groups	HIZ was the only MRI finding evaluated. Standardized MRI examination. Inter- and intraobserver agreement excellent (intraclass correlation coefficient = 0.81-0.98) 1.5 T MRI
Kovacs 2013 [207]	Case-control 240 cases 64 controls referred for headache (normal MRI of the head) Age 35-40 years	LBP ≥ 90 days with or without leg pain Controls had ≤ 1 prior LBP episode for < 7 days	• Severe disc degeneration (Pfirrmann grade 4 and 5: grey or black nucleus pulposus signal with no clear distinction between nucleus and annulus, and normal to collapsed disc space)	- Prevalence of ≥ 1 disc with severe degeneration 65.8 % in cases and 46.9 % in controls • OR 2.06, 95 % CI, 1.05-4.05 • OR 1.81, 95 % CI, 0.81-4.05 adjusted for Modic changes and disc protrusion/herniation	Association between severe DD and chronic LBP not significant when Modic changes and disc contour were taken into account, but study may be underpowered. 1.5 T MRI

Kaapa 2012 [93]	Cross sectional 62 of 4380 consecutive LBP patients with a clearly detectable Modic type I Age 25-65 years	Nonspecific LBP > 3 months, Oswestry Disability Index (ODI)	• Relative sagittal area of Modic type I lesion (compared to sagittal area of corresponding vertebrae) • Mixed Modic type I/II vs. pure type I	• No correlation with pain intensity or ODI for largest single Modic type I lesion or for sum of relative sagittal areas of Modic type I at all lumbar disc levels • More pain ($P = 0.0451$) and disability ($P = 0.0156$) in patients with pure Modic type I vs. mixed Modic type I and II	Selection of patients with large Modic type I lesions (largest lesion was average 24.6 % of sagittal area of corresponding vertebrae). No controls. 1.5 T MRI
Cheung 2009 [200]	Cross sectional population study 1043 persons Age 18-55 years	Ever LBP > 2 weeks and consulted physician or received treatment	• Annular tear • Disc degeneration score (score for disc signal/height summed across all lumbar levels) • Disc herniation • Schmorl´s nodes	• OR 2.1; 95 % CI, 1.4-3.1 • OR 2.2; 95% CI, 1.4-3.4 • No significant association • No significant association	Disc degeneration score: 1-slight decrease in nucleus pulposus signal, 2-generalized hypointense nucleus, 3-generalized hypointense nucleus with disc space narrowing. 0.2 T MRI

Kjaer 2006 [14]	Cross sectional population study 412 persons, 393 had MRI Age 40 years	• LBP duration past year • LBP week, month, year • Seeking care • ≥ 1 prior episode • Non-trivial LBP	• Modic changes (any) • Disc degeneration (hypointense nuclear complex signal and/or disc height narrower than the disc above if normal) [73]	• Persons with a) Modic changes and disc degeneration had the highest prevalence of all LBP parameters; the prevalence was lower in b) those with disc degeneration only, and even lower in c) those with no disc degeneration or Modic changes (*P* range, 0.0001 to 0.0032) • The prevalence of LBP parameters was more similar in groups b and c compared to group a	73 persons had Modic changes and disc degeneration, 141 had disc degeneration only, 179 had neither disc degeneration nor Modic changes. 0.2 T MRI
Kleinstuck 2006 [39]	Cross sectional 53 patients with chronic nonspecific LBP Average age 44 years	• LBP intensity in the last 2 weeks (visual analogue scale) • Low back disability (Roland Morris questionnaire)	• Modic type I and II (grade 0, no changes to grade 3, severe changes; extent ≥ 50% of the vertebral height as measured on the midsagittal image) • HIZ (yes/no) • Disc degeneration (Pfirrmann grade 1-5) • Disc bulge (yes/no)	• None of the MRI variables contributed significantly to explaining the variance in baseline levels of either average pain, worst pain, or disability (multiple regression analyses adjusted for age and gender, *P* > 0.05)	LBP intensity: worst pain and average pain examined separately. Only sagittal T2 weighted images available for analysis and Modic type I and II assessed together. 1.5 T MRI

McGregor 2004 [25]	Cross sectional retrospective 1064 female twins Age 45-72 years	Lifetime prevalence of LBP (> 1 month duration and disability)	MRI severity sum score based on 4 point grading of • Disc signal • Disc height • Disc bulge • Anterior osteophyte	OR for LBP using lower quartile MRI score as reference group • 3.6 (95 % CI, 1.8-7.3) for upper quartile MRI score • 3.45 (95 % CI, 1.74-6.85) for third quartile MRI score • 1.25 (95 % CI, 0.58-2.70) for second quartile MRI score	MRI changes the strongest predictor of LBP (compared to age, weight, height, smoking, alcohol, weight bearing activity, psychological distress). 1.0 T MRI
Videman 2003 [72]	Cross sectional retrospective 115 monozygotic male twin pairs Age 35-69 years	- Current LBP - Ever LBP for > 1 day; intensity worst episode; number of episodes - LBP past year: frequency and intensity of episodes - Disability past year - Disability from worst lifetime LBP episode	• Annular tears • Disc height • Disc bulging • Disc herniations • Osteophytes • Spinal stenosis • Endplate irregularities	• *Annular tear* (OR range 1.5-1.9) and *disc height decrease* (OR range 4.0-5.0) associated with lifetime frequency of LBP interfering with daily activity, disability, and worst lifetime pain episode • *Annular tear* associated with LBP frequency (OR 1.8; 95 % CI, 1.1-2.9) and intensity (OR 1.8; 95 % CI 1.2-3.0) past 12 months before imaging • *Disc height narrowing* associated with LBP frequency (OR 2.2; 95 % CI, 1.4- 3.7) and intensity (OR 1.8; 95 % CI 1.1-2.9) past 12 months before imaging No other MRI finding associated with LBP variables	Adjusted for age and clustered by twin pair in multivariate model: Annular tear and disc height narrowing only significant MRI parameters associated with LBP; associations were weak but strongest for lifetime LBP and LBP past 12 months. 1.5 T MRI

Jarvik 2001 [88]	Cross-sectional 148 patients from the health care system without LBP or with less than “mildly bothersome” LBP past 4 months Age 36-71 years, mean 54 years	Number of previous LBP episodes (never, 1-5 times, more than 5 times)	• Modic changes • Annular tear (focal hyperintensity of the annulus on T2-weighted images) • A) Disc signal reduction (moderate or severe) • B) Disc height loss (any decrease compared to a normal disc) • Disc degeneration (A, B, or moderate to severe disc bulge) • Bulge, protrusion, extrusion • FA (moderate or severe hypertrophy)	• Extrusion only MRI finding significantly associated with history of LBP: OR 32.5 (95 % CI, 3.4-308.3) for > 5 times versus never previous LBP	Results are from baseline cross-sectional part of a longitudinal prospective study, and concern presence of the MRI finding at one or more levels from L1-S1. 1.5 T MRI

Borenstein 2001 [208]	Cross sectional and prospective 50 persons with no history of LBP/sciatica 31 rescanned at 7 years follow-up Average age 43.6 years at follow-up	Duration of LBP, from 0 = no pain to 5 = more than 6 weeks	• Disc degeneration based on signal intensity of nucleus pulposus and disc height decrease • Herniated nucleus	• Moderate disk degeneration correlated to duration of LBP (*P* = 0.04) • Correlated to duration of LBP (*P* = 0.01)	Results for cross-sectional follow-up part of prospective study. 1.5 T MRI
Luoma 2000 [70]	Cross sectional retrospective 164 men Age 40-45 years	LBP or sciatica past 12 months, past 4 years, lifetime	• Dark/black nucleus pulposus signal • Anterior or posterior disc bulge	• LBP past 12 months associated with numbers of discs with - dark/black nucleus pulposus (OR 2.0; 95 % CI, 1.2-3.1) - posterior bulge (OR 2.7; 95 % CI, 1.5-4.8) - anterior bulge (OR 3.4, 95 % CI, 1.4-8.2) • LBP past 4 years associated with numbers of discs with dark/black nucleus pulposus (OR 2.1, 95 % CI, 1.2-3.6)	OR for LBP per one unit change in the number of discs with the sign, adjusted for body height, smoking, overweight, and history of car driving. 0.1 T MRI