Towards a uniform definition for the centralisation phenomenon

By

Angeliki Georgiou Chorti PT MSc MCSP SRP

A thesis submitted in accordance with the requirements for the degree of

Doctor of Philosophy in Health Sciences

University of Warwick, Warwick Medical School

September 2009

Στον Τασούλη και τον κύριο Γιάννη

"When you can measure what you are speaking about, and express it in numbers, you know something about it; but when you can not measure it, when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind: it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the state of science."

Lord Kelvin 1824-1907

TABLE OF CONTENTS

SECTION 2 - ESTABLISHING CONSENSUS

CHAPTER 4 – A Delphi study on the centralisation phenomenon.

SECTION 4 – THESIS CONCLUSIONS AND FUTURE RESEARCH RECOMMENDATIONS

LIST OF ABBREVIATIONS

ADL	Activities of Daily Living
ADTO	Assessment – Diagnosis – Treatment – Outcome
AGREE	Appraisal of Guidelines for Research and Evaluation
AHCPR	Agency for HealthCare Policy and Research
ANR	Adherent Nerve Root
APTA	American Physical Therapy Association
AROM	Active Range of Motion
Ca	Categorical
CBI	Canadian Back Institute
Chir	Chiropractor
CHQ12	12-item Chinese Health Questionnaire
CI	Confidence Interval
Co	Continuous
COM	Combination of Strategies
ConV	Consecutive Visits
CP	Centralisation Phenomenon
CROM	Cervical Range of Motion
CSAG	Clinical Standards Advisory Group
CSP	Chartered Society of Physiotherapy
CSQ	Coping Strategies Questionnaire
Cx	Cervical
DASH	Disability of the Arm, Shoulder and Hand
DASH – W	Work Component of the DASH Questionnaire

DC	Doctor of Chiropractic
DoH	Department of Health
FAB	Fear Avoidance Beliefs
FABQ	Fear Avoidance Beliefs Questionnaire
FABQPA	Fear Avoidance Beliefs Questionnaire Physical Activity Subscale
FABQW	Fear Avoidance Beliefs Questionnaire Work Subscale
GM	Gross Movement
GP	General Practice
GPES	Global Perceived Effect Scale
GROC	Global Rating of Change
HLQoL	Health Related Quality of Life
HPC	Health Professions Council
HR	Hazards Ratio
IASP	International Association for the Study of Pain
ICC	Intraclass Correlation Coefficient
ICD	International Classification of Diseases
IDD	Internal Disc Disruption
Int	Reduction in Intensity of the Most Distal Symptoms
QTF	Quebec Task Force
K	Kappa Coefficient
K_W	Weighted Kappa Coefficient
LBP	Low Back Pain
LBPRS	Low Back Pain Rating Scale
Loc	Abolition of Distal Pain
LOE	Levels of Evidence

LR	Likelihood Ratios
Lx	Lumbar
MD	Medical Doctor
MDT	Mechanical Diagnosis and Therapy
MeSH	Medical Subject Headings
M/F	Male / Female
MSPQ	Modified Somatic Perception Questionnaire
MVA	Motor Vehicle Accident
N/A	Not Applicable
NASS	North American Spine Society
NDI	Neck Disability Index
Neuro	Improvement of Neurological Signs and Symptoms
NHP	Nottingham Health Profile
NIOSH	National Institute for Occupational Safety and Health
NOS	Non-Organic Signs
NP	Neck Pain
NR	Not Reported
NRS	Numerical Rating Scale
NRS-11	11-point Numerical Rating Scale
NSLBP	Nonspecific Low Back Pain
NT	Neural Testing
ODQ	Oswestry Disability Questionnaire
OR	Odds Ratio
ORs	Odds Ratios
OSW	Modified Oswestry Disability Questionnaire

OT	Occupational Therapist
PSEQ	Pain Self-Efficacy Questionnaire
P	Peripheralisation
PA	Palpation
PPI	Present Pain Intensity
PSFS	Patient Specific Functional Scale
Pts	Patients
PT	Physical Therapy
PTs	Physiotherapists
Q	Quality
R	Reliability
RCT	Randomised Controlled Trial
RMDQ	Ronald Morris Disability Questionnaire
RMov	Repeated Movements
RMS	Root Mean Square
ROC	Receiver Operating Characteristics
ROM	Range of Motion
RR	Relative Risk (or Risk Ratio)
RRs	Relative Risks
RTW	Return to Work
SB	Side Bend
SD	Standard Deviation
SI	Sacro-Iliac
SIJ	Sacro-Iliac Joint
SingV	Single Visit

SLR	Straight Leg Raise
SM	Segmental Movement
SMov	Single Movements
SnNout	Sensitivity Negative Out
SOHDs	Safety and Occupational Health Departments
SP	Special Testing
SPl	Shilpa Patel
SpP	Sports Related Population
SpPin	Specificity Positive In
SPSS	Statistical Package for Social Sciences
ST	Static Testing
TBC	Treatment-Based Classification
TENS	Transcutaneous Electrical Nerve Stimulation
ThTech	Therapist Generated Techniques
TSK	Tampa Scale of Kinesiophobia
Tx	Thoracic
UK	United Kingdom
US	United States
VAS	Visual Analogue Scale
WADs	Whiplash Associated Disorders
WMA	World Medical Association
WHO	World Health Organisation
WhP	Whiplash Population
WP	Working Population

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

ACKNOWLEDGEMENTS

I would like to thank:

My supervisors, Professors Sallie Lamb and Martin Underwood as well as Dr Chris McCarthy; each has contributed to this thesis in an important way and invested time and effort in guiding this work. In so many ways the thesis would not have seen the light of day without them.

Anastasios Chortis and Nikolaos Strimpakos, for their assistance in the systematic reviews – they spent so much time reviewing and rating papers!

Professor Jane Barlow and Krysia Saul for their support and advice on ethical issues; their patience with my enquiries and research study refinements exceeded any scope of duty.

Andy Kay, Jayne Watkin (University of Warwick), Bill Lesley and Karen Lawrence (Coventry University) for promoting the Neck Pain Assessment study and participating in the screening and referrals of individuals with neck pain.

Special thanks should also be extended to Kelly Parkes, communication and advertising officer, for advertising the Neck Pain Assessment study at the University of Warwick.

Cheryl Ritchie, Janet Lowe, Mark Williams and Sheila Leddington Wright for all the time they spent in assisting with the data collection of the Neck Pain Assessment study. I am also indebted to them for the valuable feedback they provided in the

design stage of this study; by offering a fresh eye and critical perspective on the study material, meaningful refinements could be made to it.

Chris Parkin, David Adey, Parm Jootla and Stuart Mathieson, for IT support and their kindness on my numerous requests and enquiries! Many thanks also to Bal Dhesi, from IT services for his support on video editing and for providing video editing software that I could actually use!

Dr. Ranjit Lall and especially, Dr. Tim Friede, statisticians; they selflessly spent time advising me in many of my statistics queries. Special thanks should also be given to Alan Donner and Harald Janson, for providing advice on reliability issues and the software for the calculation of the multivariate kappa statistic, respectively.

Dr. Stephen May, for his feedback on the PhD project plans and comments on the prognostic systematic review paper. And also, thanks to the peer reviewers of Spine and the Society for Research in Rehabilitation, for their comments on the systematic reviews.

A special thanks to all my participants: the Delphi panel for providing their suggestions on the definition and operational criteria of centralisation, and the participants with neck pain who took the time and effort to help with my research.

I am also indebted to the Hellenic State Scholarships Foundation and the University of Warwick, for the generous funding of my studentship and for bearing a proportion of the research costs. Without this funding, it would have been impossible to do this PhD.

Finally, words of gratitude do not go far enough to friends and immediate family, all of whom have endured more than their fair share of the ordeals associated with the sometimes painful and often testing process of dealing with a large volume of work such as this. A special thanks to Joaquim Magalhaes, for offering a shoulder to cry when things got tough and for always being ‘there’; and friends and colleagues from the Medical School at the University of Warwick; especially, Shilpa Patel, for her friendship and support with my research and Tarla Patel for her patience …

DECLARATION AND INCLUSION OF MATERIAL FROM A PRIOR THESIS

Degree for which thesis is submitted: PhD in Health Sciences

Title of thesis: Towards a uniform definition for the centralisation phenomenon

Date of submission: September 2009

Original registration date: September 2005

Name of candidate: Angeliki Chorti

Research institute: Warwick Clinical Trials Unit

Name of Lead Supervisor: Professor Sarah Elizabeth Lamb

I certify that:

This thesis has not been submitted for a degree at another university. However, the following material in the thesis has been used before or has been published:

Peer reviewed papers

Williams, M., McCarthy, C., Chorti, A., Cooke, M., Gates, S. (2010) A systematic review of reliability and validity studies of methods for measuring active and passive Cervical Range of Motion. *Journal of Manipulative and Physiological Therapeutics*, 33, 138-155.

Chorti, A. Chortis, A. Strimpakos, N. McCarthy, C. Lamb, SE. (2009) The prognostic value of symptom response assessment in the conservative management of spinal pain: a systematic review. *Spine*, 34, 2686-2689.***Peer reviewed abstracts***

Chorti, A. Chortis, A. Strimpakos, N. McCarthy, C. Lamb, S. The interrater reliability of the rating of clinically induced symptom responses in the physical examination of spinal pain: a systematic review. *Accepted: Clinical Rehabilitation*

Chorti, A. Chortis, A. Strimpakos, N. McCarthy, C. Lamb, S. The interrater reliability of the rating of clinically induced symptom responses in the physical examination of spinal pain: a systematic review. (Poster presentation, Society for Research in Rehabilitation, July 1st- 2nd 2009, Bath, UK).

Chorti, A. McCarthy, C. Lamb, S. Symptom response assessment: how confident can we be about its importance in reliably predicting back pain outcomes? (Poster presentation, Boston International Forum X: Primary care research on low back pain, June 14-17 2009, Boston, USA)

Chorti, A. McCarthy, C.J. Lamb, S. Toward a uniform definition for the centralization phenomenon: a Delphi study (VIPER presentation, CSP Congress, October 17-18 2008, Manchester, UK)

Chorti, A. McCarthy, C.J. Lamb, S. The centralization phenomenon: results from a recent international Delphi consensus survey (oral presentation, 1st Panhellenic MDT Conference, May 23-25 2008, Athens, Greece).

Internal presentations

Chorti, A. Using qualitative methods in healthcare research: the example of the Delphi technique. (Research degree seminar, University of Warwick, Coventry, UK).

Chorti, A. Using qualitative methods in healthcare research: the example of the Delphi technique. (PhD club, year 4, University of Warwick, Coventry, UK).

Chorti, A. A Delphi study on the centralisation phenomenon: preliminary results. (PhD club, year 2, University of Warwick, Coventry, UK).

Chorti, A. Evaluation of assessment procedures for the spine and classification systems (PhD club, year 1, University of Warwick, Coventry, UK).

Contributions

The work reported in this thesis is the candidate's own work; however, it could not have been properly accomplished as a solo effort. Several colleagues were involved in important parts of the research process. The nature and extent of these contributions are detailed below:

Mr Anastasios Chortis: assistance in the development of the protocol design (systematic reviews), data analysis of included studies (systematic reviews), critical revision of produced drafts and manuscripts (systematic reviews)

Dr Tim Friede: assistance in data analysis and interpretation, review of chapter (Neck Pain Assessment study)

Dr Ranjit Lall: assistance in sample size calculation (Neck Pain Assessment & VideoNeck studies)

Professor Sarah Elizabeth Lamb: assistance in the development of the protocol design (systematic reviews, Delphi study, Neck Pain Assessment & VideoNeck studies); critical revision of produced drafts and manuscripts (systematic reviews, Delphi study, Neck Pain Assessment & VideoNeck studies)

Ms Janet Lowe: data collection (Neck Pain Assessment study); critical revision of produced drafts, manuals and manuscripts (Neck Pain Assessment & VideoNeck studies)

Ms Sheila Leddington Wright: assistance in the development of the protocol design (Neck Pain Assessment & VideoNeck studies); data collection (Neck Pain Assessment study); critical revision of produced drafts, manuals and manuscripts (Neck Pain Assessment & VideoNeck studies)

Dr Christopher McCarthy: assistance in the development of the protocol design (systematic reviews, Delphi study, Neck Pain Assessment & VideoNeck studies);

data analysis (Delphi study); critical revision of produced drafts and manuscripts (systematic reviews, Delphi study, Neck Pain Assessment & VideoNeck studies); scientific advisor

Dr Shilpa Patel: assistance in recruitment and data cleaning (Neck Pain Assessment & VideoNeck study)

Ms Cheryl Ritchie: data collection (Neck Pain Assessment study); critical revision of produced drafts, manuals and manuscripts (Neck Pain Assessment & VideoNeck studies)

Dr Nikolaos Strimpakos: assistance in the development of the protocol design (systematic reviews), data analysis of included studies (systematic reviews), critical revision of produced drafts and manuscripts (systematic reviews)

Professor Martin Underwood: critical revision of produced drafts and manuscripts (systematic reviews, Delphi study, Neck Pain Assessment & VideoNeck studies)

Mr Mark Williams: assistance in the development of the protocol design (Neck Pain Assessment & VideoNeck studies); data collection (Neck Pain Assessment study); critical revision of produced drafts, manuals and manuscripts (Neck Pain Assessment & VideoNeck studies)

Total words in submitted thesis: 58139.

ABSTRACT

The focus of this PhD project is on symptom centralisation. Its primary objectives were to establish a standard operational definition of centralisation and evaluate its inter-rater reliability in neck pain.

Two systematic reviews on the reliability and prognostic value of clinically induced symptom responses in spinal pain (Chapters 2 and 3) showed that although the potential usefulness of centralisation has been demonstrated in low back pain, concern has been expressed about the observed inconsistency in reported definitions, and the scarcity of studies in neck pain.

A Delphi survey of experts (Chapter 4) assisted in the development of a uniform operational definition for centralisation and the identification of future research questions. Centralisation was generally defined as the progressive and stable reduction of the most distal presenting pain towards the spine midline in response to standardised spinal loading strategies.

The support by the Delphi panel of a broader definition allowed for a multitude of different ways of testing to be included in the assessment procedure which may offer some flexibility to clinicians assessing, classifying and managing different spinal pain presentations across different countries. Although the reliability of identifying centralisation was acceptable, high levels of reliability were not demonstrated (Chapter 5). However, reliability was greater for the pair of physiotherapists with prior experience and formal extensive training in symptom response assessment. Therefore, the optimal type and amount of training for eliciting and interpreting centralisation and the effect of procedural variations on testing the outcomes of individuals who centralise require urgent investigation. The prognostic and management value of centralisation are also unknown.

This investigation may pave the way for the standardisation of centralisation as a physical sign and stimulate interest for further study of potential sub-groups and classification of spinal syndromes.

SECTION 1

BACKGROUND AND LITERATURE REVIEWS

Chapter 1

General introduction

1.1 AIMS OF CHAPTER

The aims of this chapter are to:

- Present the reader with the background of this thesis by:
    - Introducing neck pain, its course, impact, risk and prognostic factors, and current assessment and treatment approaches;
    - Providing an overview of classification systems and clinically induced symptom responses, and introducing the centralisation phenomenon in spinal pain;
- Present the general objectives of the thesis and the approach followed;
- Outline the structure and content of the thesis.

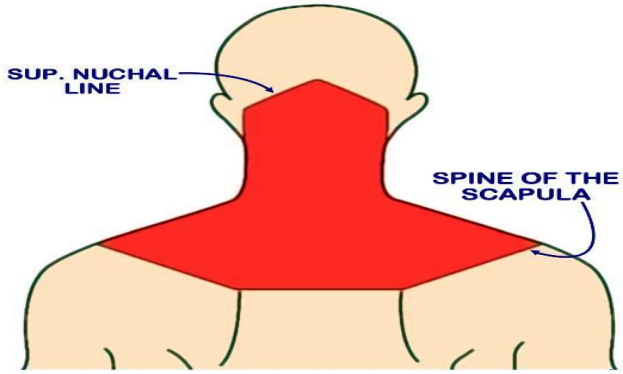
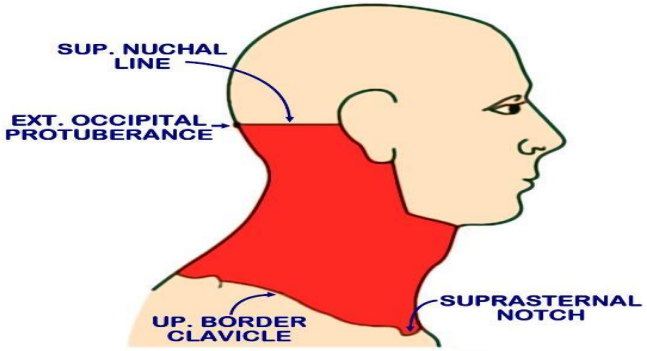
1.2 BACKGROUND

1.2.1 The problem of neck pain

1.2.1.1 Case definition

Great variation exists in how neck cases are described and considered; more than 300 case definitions have been reported by the Task Force on Neck Pain (Guzman et al., 2008). This variability, attributed to the lack of sufficient knowledge about musculoskeletal symptoms (Kuorinka et al., 1987), is similar to the inconsistency in reports of low back pain (LBP) (de Vet et al., 2002; Dionne et al., 2008). Three anatomic neck pain (NP) case definitions are recommended by national and international guideline groups (Figure 1.1). With anatomic definitions providing no clinically relevant information, new ways of labelling and classifying NP have been proposed (see section 1.2.2, "The challenge of spinal diagnosis").

Figure 1.1 Anatomic neck pain definitions proposed by national and international guideline groups.

GROUP AND DEFINITION		AUTHOR'S COMMENTS
Kuorinka and associates (1987) Pain in the cervical and upper thoracic spine area		Produced for the analysis of musculoskeletal symptoms in an ergonomic or occupational health context.
Philadelphia Panel (2001) Pain in the neck area, with or without radiation to the extremities		Introduced the term "non-specific" in neck pain.
Task Force on Neck Pain (Guzman et al., 2008). Pain in the neck area, with or without radiation to the head, trunk, and upper limbs		The most recent definition of neck cases.

The anatomic region of the neck. Adapted from: Guzman et al. (2008), with permission.

1.2.1.2 Prevalence and impact

NP is a common symptom among the general population and employees of various professions (Cote et al., 2008; Hogg – Johnson et al., 2008). In the world population, NP has been found to have mean point, one-week, one-month, six month, one-year, and lifetime prevalence rates of 7.6%, 12.5%, 23.3%, 29.8%, 37.2% and 48.5% respectively (Fejer et al., 2006). In the UK, NP has been cited as one of the four most commonly reported musculoskeletal symptoms (Urwin et al., 1998), with about one – fifth of adults reporting a new episode within the previous 12 months (Croft et al., 2001).

NP is an important source of burden to society. It is considered a frequent cause of work absence (Borghouts et al., 1999a). In the Netherlands, the estimated direct economic cost of NP was US $686 million in 1996 (Borghouts et al., 1999b). This represented approximately 0.1% of the entire gross domestic product of the Netherlands (Tuchin, 2008). The indirect costs included days off work and loss of productivity, estimated at US $185.4 million (approximately 1.4 million days), and disability pensions of US $341 million (Borghouts et al., 1999b).

1.2.1.3 Treatment

NP cases are very common in the healthcare services (Hurwitz et al., 1998; Riddle and Schappert, 2007). Most treatment and consultation choices for NP involve the GP (Lock et al., 1999; Tuchin, 2008) with the most commonly prescribed interventions for acute NP being a "wait and see policy" and / or advice, relative rest and analgesics (Borghouts et al., 1999a; Vos et al., 2007).

NP is one of the most common conditions for referral to a physiotherapist (Philadelphia Panel, 2001). In a recent survey of physiotherapy clinics in the UK, approximately 23% of musculoskeletal patients had NP (May, 2003). Not all people with NP seek care for their problem; only 25% have been reported to visit a health provider (Cote et al., 2001a). A wide range of non-invasive (or conservative) interventions are available to individuals with NP (Table 1.1). Good quality evidence supporting effectiveness of the most commonly used interventions in reducing the incidence and course of NP is scarce (Cassidy and Cote, 2008; Hurwitz et al., 2008). Less than half of the literature on NP interventions has been of acceptable scientific quality to inform clinical practice, policy making and future research (Carroll et al., 2008d). This may result in substantial disagreement on the best treatment methods for NP.

Table 1.1 Non-invasive interventions for neck pain (Philadelphia Panel, 2001; Cassidy and Cote, 2008; Hurwitz et al., 2008).

INDIVIDUAL LEVEL	Advice Education Acupuncture Electrical stimulation Exercise Laser Manual therapy Massage Thermotherapy Traction Transcutaneous Electrical Nerve Stimulation (TENS) Ultrasound
POPULATION LEVEL	Modification of the community environment through physical, psychosocial, economic and legal / regulatory interventions

1.2.1.4 Course

Most people with NP do not experience a complete resolution of symptoms (Haldeman, 2008). The most dramatic improvements are expected to occur within the first 3 months (Kamper et al., 2008; May et al., 2008a); after this period, patients with acute spinal symptoms pass to the chronic stage (van Tulder et al., 2003; Koes et al., 2006) or changes are expected to be small (Borghouts et al., 1998; Vernon et al., 2006; Kamper et al., 2008) and similar regardless of the treatment offered (Hoving et al., 2006). Between 50% and 85% of those with NP will report recurrence within 1 to 5 years (Carroll et al., 2008a; Carroll et al., 2008b; Carroll et al., 2008c). This pattern of symptom presentation is similar in the general population, workers or whiplash injuries and comparable to the reported pattern in LBP (Hestbaek et al., 2003; Pengel et al., 2003).

For most NP sufferers, interference with activities is minimal (Enthoven et al., 2004; Fejer and Hartvigsen, 2008). In most cases, neck complaints involve a mild discomfort which does not require treatment, or has any major impact on work (Haldeman, 2008). Some people develop prolonged or repetitive episodes or more serious symptoms (Haldeman, 2008). Nevertheless, factors that predict future course or outcomes are required to guide patients' and health professionals' expectations, and for more efficient targeting of resources (Carroll et al., 2008a).

1.2.1.5 Risk and prognostic factors

Risk factors are associated with the onset of NP (Borghouts et al., 1998) (Table 1.2). Prognostic factors affect the outcome once a NP episode has started (Borghouts et al., 1998) (Table 1.3). Factors may be modifiable or non-modifiable, depending on the feasibility of changing (Guzman et al., 2008).

Very few individual factors can direct towards definite prognostic decisions and relevant clinical interventions because of their non-modifiable nature, small / unknown associations with treatment or most outcomes (Carroll et al., 2008a; Carroll et al., 2008b; Carroll et al., 2008c; Cote et al., 2008; Hogg-Johnson et al., 2008; Holm et al., 2008) (Table 1.2 and 1.3). There is little information on important biological or clinical factors (Johnston et al., 2008; Jull and Sterling, 2009). Like LBP (Pincus et al., 2008), few studies are of sufficient size and methodologic rigour to produce conclusive findings (Carroll et al., 2008b).

Table 1.2 Statistically significant risk factors for the onset of neck pain*.

Risk factor	Population	Case definitions	Comments
Age	WhP	Whiplash defined by questionnaire as presence of neck / shoulder pain caused by collision and reduced or painful neck movement ever since[1]	Cohort (natural experiment) IRR $_{18\text{-}23}$ = 4.6, IRR $_{24\text{-}29}$ = 3.3, IRR $_{30\text{-}39}$ = 2.4, IRR $_{40\text{-}49}$ = 1.9, all compared to age over 55
Age	SpP	Any neck injury occurring during a match and requiring medical attention[2]	Phase I cohort Increasing age was associated with increased risk of injury
Gender	SpP	Neck injury requiring medical attention and resulting in modification of participation[3]	Phase I cohort IRR = 3.9 (1.1 - 20.7)
History of neck pain	GP	Neck pain lasting > 1 day over the past 12 months[4]	Phase II cohort RR = 1.7 (1.2 - 2.5)
History of neck pain	WhP	Being hit from behind in a rear-end collision[5]	Phase I cohort No values reported
History of low back pain	GP	Neck pain lasting > 1 day over the past 12 months[4]	Phase II cohort RR = 1.7 (1.3 - 2.1)
History of headaches	WP	Neck pain in the past 7 days and 12 months[18]	Phase II cohort OR = 2.1 (1.1 - 3.9)
History of neck injury via motor vehicle accident	GP	Neck pain often or always in the past 3 months[10]	Phase I cohort RR = 2.7 (2.1 - 3.5)
History of neck injury	SpP	Acute neck injuries resulting in complete or partial time loss, or any concussion or transient neck neurological injury[11]	Phase II cohort Adjusted rate ratio = 5.0 (3.1 - 8.2)
Healthcare visits	GP	Neck pain for at least 1 day over the past month[6]	Phase III case control study OR = 1.1 - 1.4
Physical capacity	WP	Localised musculoskeletal discomfort < 4[17]	Phase III cohort Isometric lifting strength neck / shoulder muscles: HRR $_{high}$ = 1.0, HRR $_{moderate}$ = 1.2 (0.9 – 1.6), HRR $_{low}$ = 1.3 (1.0 – 1.7); Static endurance of neck flexors: HRR $_{high}$ = 1.0, HRR $_{moderate}$ = 1.2 (0.9 - 1.4), HRR $_{low}$ = 1.2 (1.0 – 1.5)

Rule changes in sports	SpP	Neck injuries requiring attention as a result of incident during game[12]	Phase I cohort IRR = 0.2 (0.1 - 1.1)
Ethnicity	WP	Neck / shoulder symptoms ≥ 6/10 in the past week or used medication in the past week[15]	Phase II cohort $HRR_{neck/shoulder\ symptoms}$ = 0.7 (0.5 - 1.0)
Country of origin	WP	Neck pain experienced often or all the time in the past 12 months[16]	Phase II cohort OR_{men} = 1.6 (1.0 - 2.6) OR_{women} = 1.8 (1.2 - 2.9)
Poor psychological status	GP	Neck pain lasting > 1 day over the past 12 months[4]	Phase II cohort RR = 1.1 - 1.7 (0.8 - 2.3)
Psychosomatic symptoms	GP	Occasional or weekly neck / shoulder pain over the past 6 months[7]	Phase I cohort OR = 1.0 (1.0- 1.1)
Depressive / emotional symptoms	WP	Neck pain in the past 7 days and 12 months[18]	Phase II cohort OR_{medium} = 1.0, OR_{low} = 5.6 (2.0 – 15.3), OR_{high} = 4.7 (1.7 – 13.0)
Depressive / emotional symptoms	WP	NP ≥ moderate severity in the last 7 days quite bothersome in the last 12 months[19]	Phase II cohort $RR_{negative\ affectivity}$ = 1.3 (0.7 - 2.5)
Depressive / emotional symptoms	WP	Neck / shoulder pain ≥ 6 months with functional limitations[20]	Phase II cohort OR_{men} = 1.3 (1.0 – 1.8), OR_{women} = 1.5 (1.2 – 1.9)
Personality type	WP	NP ≥ moderate severity in the last 7 days quite bothersome in the last 12 months[19]	Phase II cohort $RR_{type\ A\ personality}$ = 1.7 (0.9 – 3.1)
Influence on own work	WP	Neck pain in the past 7 days and 12 months[18]	Phase II cohort OR_{great} = 1.0, OR_{some} = 1.7 (0.8 – 3.3), $OR_{little/very\ little}$ = 2.9 (1.2 – 6.7)
Physical environment	WP	Neck pain in the past 12 months[13]	Phase II cohort OR = 1.7 (1.2 – 2.5)
Physical environment	WP	Local or radiating neck pain ≥ 8 days in the past 12 months[14]	Phase II cohort OR = 2.1 (0.9 – 4.9)
Cigarette smoking	GP	Surgical, probable and possible neck cases[8]	Phase II case control OR = 2.1 (0.9 - 5.0)
Exposure to environmental tobacco	GP	Absence from work because of neck pain > 14 days in the past 12 months[9]	Phase III cohort study OR = 1.4 (1.0 - 1.8)

Head posture	WP	Neck / shoulder symptoms or disorders[21]	Phase III cohort $HRR_{symptoms} = 1.5\ (0.9 – 2.6)$ $HRR_{disorders} = 1.6\ (0.8 – 3.3)$
Work posture (flexion > 70% time)	WP	Neck pain in the last 12 months[22]	Phase III cohort RR = 1.6 (0.7 – 3.8)
Upper extremity posture	WP	Neck / shoulder symptoms or disorders[21]	Phase III cohort $HRR_{keyboard\ inner\ elbow\ angle} = 0.2\ (0.0 – 0.6)$ $HRR_{distance\ from\ table\ to\ J\ key\ >\ 17\ cm} = 0.7\ (0.5 – 1.2)$ $HRR_{mouse\ shoulder\ flexion\ angle \leq 25°} = 1.0$ $HRR_{mouse\ shoulder\ flexion\ angle\ 26\text{-}34°} = 1.3\ (0.8 – 2.2)$ $HRR_{mouse\ shoulder\ flexion\ angle\ 35\text{-}44°} = 1.7\ (1.0 – 3.0)$ $HRR_{mouse\ shoulder\ flexion\ angle\ >44°} = 1.3\ (0.7 – 2.3)$
Awkward postures or work	WP	Neck pain in the past 12 months[13]	OR = 1.0- 1.8 (0.9 – 3.2)
Awkward postures or work	WP	Neck pain in the past 12 months[23, 24]	Phase II cohort OR = 1.8 (1.1 – 2.8)
Awkward postures or work	WP	Neck / shoulder pain ≥ 6 months with functional limitations[20]	Phase II cohort $OR_{men} = 1.0 – 1.3\ (1.0 – 1.8)$
Keyboard / mouse position	WP	Neck / shoulder symptoms or disorders[21]	Phase III cohort HRR = 1.3 (0.8 – 2.1)
Keyboard / mouse position	WP	Local or radiating neck pain ≥ 8 days in the past 12 moths[14]	OR = 2.1 (1.0 – 4.5)
Telephone headrest	WP	Neck / shoulder symptoms or disorders[21]	Phase III cohort HRR 1.7 (1.0 – 3.1)
Frequently experiencing technical problems with a computer	WP	Neck symptoms > 7 days in the past 12 months[25]	Phase II cohort $RR_{< 1\ month} = 1.0$, $RR_{weekly} = 1.3\ (0.9 – 1.8)$
Driver v. Passenger / Direction of collision	WhP	Whiplash injury defined as neck injury without fracture, luxation or damage of neural tissues[26]	Phase II cohort Drivers RR = 1.8 (1.6 - 2.0), front seat passenger RR 1.4 (1.3 - 1.6). Rear-end impact compared to side impact RR = 1.8 (1.7 - 2.0), frontal impact compared to side impact RR = 1.3 (1.2 - 1.4), and other impacts RR = 1.2 (1.1 - 1.3).

Use of head restraint	WhP	Neck injury insurance claim[27]	Phase I cohort $OR_{women} = 0.6$
Use of head restraint	WhP	Neck injury insurance claim[28]	Phase I cohort Redesign of head restraints and/or seats resulted in a decrease in neck injury OR = 0.6

* Values are rounded

References:
[1] Cassidy et al., 2000; [2] Lorish et al., 1992; [3] Hinton et al., 2005; [4] Croft et al., 2001; [5] Obelieniene et al., 1999; [6] Croft et al., 2003; [7] Siivola et al., 2004; [8] Kelsey et al., 1984; [9] Eriksen, 2004; [10] Berglund et al., 2000; [11] Hagel et al., 2003; [12] Watson et al., 1996; [13] Viikari-Juntura et al., 1994; [14] Korhonen et al., 2003; [15] Gerr et al., 2002; [16] Ostergren et al., 2005; [17] Hamberg - van Reenen et al., 2006; [18] Eriksen et al., 1999; [19] Brandt et al., 2004; [20] Cassou et al., 2002; [21] Marcus et al., 2002; [22] Ariens et al., 2001; [23] Luime et al., 2004; [24] Luime et al., 2005b; [25] Jensen, 2003; [26] Berglund et al., 2003; [27] Farmer et al., 1999; [28] Farmer et al., 2002.

Abbreviations:
GP, General Population; HRR, Hazard Rate Ratio; IRR, Incidence Rate Ratio; OR, Odds Ratio; RR, Risk Ratio (or Relative Risk); SpP, Sports Related Population; WP, Working Population; WhP, Whiplash Population

Table 1.3 Statistically significant prognostic factors and neck pain outcomes.

Prognostic factor	Population	Outcome	Follow-up duration	Comments
Age	GP	Reduction in pain intensity[3]	12 months	Phase II cohort OR = 0.9
Age	GP	Resolution of pain[2]	6, 12 months	Phase I cohort IRR = 0.8
Age	GP	Persistence of pain[2, 4]	6[2], 12[2,4] months	Phase I cohort[2] IRR = 1.3 Phase II cohort[4] OR = 1.7, 3.9 and 2.0 for ages 30-44, 45-59, 60-75 respectively
Age	GP	Reduction in disability[1]	12 months	Phase II cohort Beta = - 0.1 per year
Prior neck / shoulder symptoms	GP	Reduction in pain intensity[1]	12 months	Phase II cohort HRR = 0.6
Prior shoulder symptoms	WP	12-month prevalence[8]	4 years	Phase II cohort OR = 2.6 (1.4 – 4.6)
Prior neck pain > 3 months in previous year	WP	Recurrence of pain[11]	2 years	Phase II cohort OR = 1.7 (1.2 - 2.4)
No prior musculoskeletal disease	WP	Improvement in chronic pain[10]	5 years	Phase II cohort OR_{men}= 0.4 (0.3 - 0.6); OR_{women} = 0.6 (0.5 - 0.8)
Previous neck injury	GP	Pain > 1 day[4]	12 months	Phase II cohort OR = 1.5
Prior sick leave	WP	Worsening of symptoms[7]	1 year	Phase II cohort No effect sizes reported
Prior sick leave	WP	Days of sick leave > 3[9]	60 days	Phase II cohort OR = 6.5 (2.1 – 20.4)
Prior treatment and requiring additional surgery	WP	Referral for medical disability (medical unfit for duty)[12]	10-48 months	Phase I Cohort Effect sizes not reported

Comorbid low back pain	GP	Pain > 1 day[4]	12 months	Phase II cohort OR = 1.6
General health	GP	Reduction in pain intensity[1]	12 months	Phase II cohort Beta = 0.5
General health	GP	Reduction in disability[5]	6 months	Phase III cohort OR_{61-75} = 2.6; OR_{76-100} = 2.1
General exercise and sporting activities	WP	Improvement of symptoms[7]	1 year	Phase II cohort No effect sizes reported
Duration of current episode (> 6 months)	GP	Self –reported recovery[1]	12 months	Phase II cohort HRR = 0.5
Duration of current episode (> 6 months)	GP	Reduction in pain intensity[1] Reduction in disability[1]	12 months	Phase II cohort Beta = - 0.6 (pain intensity) Beta = 10.6 (disability)
Whiplash grade	WhP	Intensity of pain[19]	2 years	Phase II cohort OR_{WADII} = 1.5 (1.1 - 1.9); OR_{WADIII} = 2.4 (1.8 - 3.2)
Whiplash grade	WhP	Recovery of neck pain[15]	20 - 25 months	Phase I cohort RR = 1.8, 1.1, 2.1 between group1 / 2, group 2/3 and group 1 / 3
Whiplash grade	WhP	Time to recovery[16]	Up to 56 days	Phase I cohort Symptoms lasted longer for Grade II > Grade I (19.7 v. 6.4 days)
Whiplash grade	WhP	Disability[17]	16 months	Phase II cohort OR = 2.0 (1.1 - 3.9)
Whiplash grade	WhP	Change in health[18]	3 years	Phase I cohort RR = 3.3 (1.1 - 10.0)
Initial pain intensity	GP	Reduction in pain intensity[1] Reduction in disability[1] Self reported recovery[1]	12 months	Phase II cohort No effect sizes reported
Post-injury symptom severity	WhP	Intensity of pain[19]	2 years	Phase II cohort $OR_{\text{severe pain}}$ = 8.4 (6.5 - 10.9)

Post-injury symptom severity	WhP	Regular or moderate pain or occasional, regular or daily severe pain[23]	6 months	Phase II cohort OR $_{\text{more initial symptoms}}$ = 6.7 (5.2 - 18.8); OR $_{\text{upper extremity numbness/ weakness}}$ = 2.2 (1.2 - 3.9); OR $_{\text{disturbance in vision}}$ = 2.0 (1.0 - 3.9)
Post-injury symptom severity	WhP	Disability[24,26]	24 weeks	Phase I cohort No effect sizes reported
Post-injury symptom severity	WhP	Handicap (self-report of reduced hours and work capacity, job dismissal or change in job due to injury or application/receiving disability pension[21]	1,3, 6, 12 months	Phase II cohort More intense pain showed a trend toward predicting handicap but precision was poor
Post-injury symptom severity	WhP	Time (days) between collision and last date of compensation[22,25]	7 years	Phase II cohort RR $_{\text{neck pain on palpation}}$ = 0.9 (0.8 - 1.0); RR $_{\text{muscle pain}}$ = 0.9 (0.7 – 1.0); RR $_{\text{pain or numbness radiating to arms/hands}}$ = 0.6 (0.6 -0.8); RR $_{\text{pain or numbness radiating to shoulders}}$ = 0.8 (0.7 - 1.0); RR $_{\text{pain or numbness radiating to head}}$ = 0.8 (0.7 - 0.9) (adjusted for age and gender)
Post-injury symptom severity	WhP	Time to claim closure (self-reported recovery in depression, neck pain and physical functioning)[20]	up to 12 months	Phase II cohort HRR $_{\text{neck pain}}$ = 0.6 (0.5 - 0.8) HRR $_{\text{\%bodily pain}}$ = 0.6
Disability due to pain	GP	Reduction in pain intensity[1] Reduction in disability[1] Self reported recovery[1]	12 months	Phase II cohort No effect sizes reported
Type and intensity of initial post-injury health care	WhP	Time to claim closure (self-reported recovery in depression, neck pain and physical functioning)[20]	Up to 12 months	Phase II cohort HRR = 0.6 (MD + chiropractor in tort system, chiropractor only in fault system)
Type and intensity of initial post-injury health care	WhP	Time to claim closure (self-reported recovery in depression, neck pain and physical functioning)[30,31]	Follow up to claim closure	Phase III cohort GP (1-2 days) = 1.0; GP > 2 visits = 0.7 (0.6 - 0.9); DC (> 6visits 0.6 (0.5 - 0.8); GP and Specialist = 0.7 (0.6 -0.9); General Medicine = 0.8 (0.6 – 1.0)

Type and intensity of initial post-injury health care	WhP	Global recovery[32]	6 weeks, 3, 6, 9 and 12 months	Phase III cohort $HRR_{fitness}$ = 0.7 (0.5 - 0.9) $HRR_{outpatient}$ = 0.5 (0.3 - 0.8)
Optimism	GP	Reduction in pain intensity[3]	12 months	Phase II cohort OR = 3.0
Coping style (external locus of control)	GP	Reduction in headache frequency[6]	7 weeks, 3, 12 months	Phase III cohort OR = 1.3, 1.3, 1.2 for 3 weeks, 3 and 12 months respectively
Helplessness in controlling pain	WhP	Intensity of pain[19]	2 years	Phase II cohort OR = 2.7 (2.1 - 3.4)
Fear of movement	WhP	Disability[24,26]	24 weeks	Phase I cohort No effect sizes reported
Catastrophising	WhP	Disability[24,26]	24 weeks	Phase I cohort No effect sizes reported
Initial anxiety	WhP	Duration and severity of pain[29]	6 months	Phase II cohort No effect sizes reported
Job type (blue collar)	WP	Days of sick leave > 3[9]	60 days	Phase II cohort OR = 6.8 (2.1 – 22.4)
Job type (metal worker)	WP	Frequency and duration of sickness absence[13]	2 years	Phase II cohort RR = 2.1 (1.1 - 4.2)
Job type (enlisted personnel)	WP	Referral for medical disability (medical unfit for duty)[12]	10-48 months	Phase I Cohort RR = 2.9, p = 0.002
Shorter duration of service	WP	Referral for medical disability (medical unfit for duty)[12]	10-48 months	Phase I Cohort RR = 0.9, p = 0.02
Changing employment (previous physically heavy job)	WP	Worsening of symptoms[7]	1 year	Phase II cohort No effect sizes reported
Changing employment	WP	Disappearance of long lasting symptoms[14]	6 years	Phase II cohort OR = 4.5 (1.4 - 14.8)
High job demands	WP	Improvement in chronic pain[10]	5 years	Phase II cohort OR_{men}= 1.2 (1.0 - 1.4); OR_{women} = 1.2 (1.0 - 1.4)

Repetitive work	WP	Improvement in chronic pain[10]	5 years	Phase II cohort OR $_{women}$ = 1.3 (1.0 - 1.6)
Little influence over work	WP	12-month prevalence[8]	4 years	Phase II cohort OR = 2.5 (1.2 – 5.5)
Use of tow bar	WhP	Pain, function and / or mental dysfunction[27]	1 year	Phase I cohort RR = 1.2
Acceleration during crash	WhP	Pain, function and/or mental dysfunction[28]	≥ 6 months	Phase I cohort Symptoms had longer duration with greater mean acceleration
Insurance system	WhP	Time to claim closure (self-reported recovery in depression, neck pain and physical functioning)[20]	Up to 12 months	Phase II cohort Longer time to claim closure in tort than no fault system
Seeking legal advice	WhP	Time to claim closure (self-reported recovery in depression, neck pain and physical functioning)[20]	Up to 12 months	Phase II cohort HRR = 0.6 (0.5 - 0.7)

References:
[1] Bot et al., 2005; [2] Cote et al., 2004; [3] Michaelson et al., 2004; [4] Hill et al., 2004; [5] Hurwitz et al., 2006; [6] Stanton and Jull, 2003; [7] Jonsson et al., 1988; [8] Eriksen et al., 1999; [9] Viikari-Juntura et al., 2000; [10] Cassou et al., 2002; [11] Luime et al., 2004; [12] Kaptain et al., 1999; [13] Burdorf et al., 1998; [14] Schibye et al., 1995; [15] Norris and Watt, 1983; [16] Boyd et al., 2002; [17] Sterner et al., 2003; [18] Miettinen et al., 2004; [19] Berglund et al., 2006; [20] Cassidy et al., 2000; [21] Kasch et al., 2001; [22] Suissa et al., 2001; [23] Hartling et al., 2002; [24] Nederhand et al., 2003; [25] Suissa et al., 2003; [26] Nederhand et al., 2004; [27] Kraft et al., 2000; [28] Kraft et al., 2002; [29] Richter et al., 2004; [30] Cote et al., 2005; [31] Cote et al., 2007; [32] Cassidy et al., 2007

Abbreviations:
DC, Doctor of Chiropractice; GP, General Population; HRR, Hazard Rate Ratio; IRR, Incidence Rate Ratio; MD, Medical Doctor; OR, Odds Ratio; RR, Risk Ratio (or Relative Risk); WP, Working Population; WhP, Whiplash Population.

1.2.2 The challenge of spinal diagnosis

The usual management of spinal disorders starts with a history and physical examination (CSAG, 1994; Nachemson and Vingard, 2000). Through the history, clinicians systematically collect a verbal account from the patient covering past and present (APTA, 2001). Information on demographic, medical, social and occupation, and the current and previous episodes are some examples of the data gathered during history-taking (APTA, 2001). The physical examination may include various elements e.g. posture, muscle bulk, range of motion measurements, movement testing, palpation, neurological testing or other special tests relevant to the suspected area (Ombregt, 1995; Magee, 2002; Petty, 2006). Appendix 1.1 gives examples of physical examination procedures and their purpose.

History taking is an important part of the spinal assessment. Information from the history is primarily used to establish a patient profile and to develop hypotheses about potential and existing problems in the patient's condition (APTA, 2001). However, it is the findings from the physical examination that confirm or rule out hypotheses from the history in order to make prognostic and management decisions about the patient (APTA, 2001; Rubinstein and van Tulder, 2008). Further testing (e.g. imaging or more sophisticated diagnostic techniques) may also be used, but only to exclude structural or serious pathologies (Nachemson and Vingard, 2000; Magee, 2002).

NP may encompass a variety of conditions and causes (Ferrari and Russell, 2003; Yin and Bogduk, 2008). Historically, spine assessments relied on clinical factors to formulate a diagnosis and prognosis of treatment outcome, with history and physical

examination being evaluated for identifying the underlying pathology, but rarely for predicting outcomes (Borge et al., 2001; Varamini and Jam 2006a). It is now suggested that the patho-anatomic approach is not useful because a precise patho-anatomical diagnosis is usually impossible (Spitzer et al., 1987; CSAG, 1994; Hancock et al., 2007; Childs et al., 2008; Nordin et al., 2008). Definitions of the patho-anatomical as well as other classification approaches in spinal pain are given in Table 1.4. Thus, the term *non-specific spinal pain* has been introduced, defined as pain in the spinal area, with or without radiation to the extremities (Spitzer et al., 1987; AHCPR, 1994; CSAG, 1994; Philadelphia Panel, 2001).

Table 1.4 Definitions of classification approaches in spinal pain (Pinto et al., 2007).

APPROACH	
Patho-anatomical	Attempts to identify the nociceptive source of the patient's symptoms
Treatment-based	Uses clusters of signs and symptoms to match patients into subgroups with specific management implications
Prognostic	Based on the potential future outcome of a patient
Mechanism-based	Considers impairments identified during examination as the cause of musculoskeletal pain and dysfunction

The term 'non-specific', first introduced in LBP, was used synonymously with 'simple' LBP or 'mechanical LBP' (McCarthy et al., 2004). Several national and international guidelines recommended the term non-specific LBP (NSLBP) should encompass nerve root problems and serious pathology in an initial diagnostic triage process (Koes et al., 2001; van Tulder et al., 2006; Chou et al., 2007). Similar guidance and terminology has been given in NP (Philadelphia Panel, 2001; Guzman et al., 2008).

The non-specific label is unhelpful in characterising presentations of spinal patients. However, the non-specific category is reported to correspond to approximately 85% of all spinal cases (Abraham and Killackey-Jones, 2002). Recently, it has been suggested that spinal patients referred for conservative treatment form a heterogeneous group with prognosis and optimal treatment varying immensely (McCarthy et al., 2004; Sterling, 2004; Guzman et al., 2008). In other words, each subject is more likely to have a different prognosis and respond to a type of treatment unique to a classification (Borkan et al., 1998; Guzman et al., 2008). Heterogeneity of included study populations may be one of the reasons for the relatively small effects in clinical trials of interventions (Hay et al., 2008; Schellingerhout et al., 2008).

The potential importance of patient subgrouping has recently received widespread attention. Recommendations to establish reliable and valid classification approaches that can assist with making a prognosis and allow refinement of treatment selection have been made (Leboeuf-Yde et al., 1997; Borkan et al., 1998; Bouter et al., 1998; Koes et al., 2006).

1.2.3 Classification systems

Classification breaks down ‘a larger entity into more homogeneous subgroups of patients’ (Childs et al., 2004b, p.686). This can be accomplished by means of physical examination and presentation criteria, specific questionnaires, or other diagnostic procedures (Petersen et al., 1999). Several spinal classification approaches have been developed over the years. These are divided into one-dimensional and multidimensional (Pinto et al., 2007). Most classification systems have a biomedical nature, but some have also utilised psychosocial or biopsychosocial approaches (McCarthy et al., 2004; Billis et al., 2007).

1.2.3.1 Classification systems for spinal pain

Details on published spinal classification systems can be found in Table 1.5. Classifications have traditionally involved the use of one paradigm, and their development has been based on a judgemental or a statistical cluster analysis approach (McCarthy et al., 2004; Billis et al., 2007). Both approaches have advantages and disadvantages, with a synthesis of methodologies proposed for the development of an optimal classification system (McCarthy et al., 2004).

Some classification systems have received more attention in the literature; these are the Quebec Task Force (QTF) classification, the McKenzie classification approach, and the Treatment-Based (Delitto) Classification system (Petersen et al., 2003; Billis et al., 2007). Further information on these classification systems is provided below.

Table 1.5 Examples of spinal pain classifications.

Primary author / Country	Purpose	Method of development / type of research	Domain of interest	Categories / additional axis	Profession / setting	Criteria
Biomedical approach						
Heinrich (1985) / UK	Empirically defined diagnostic classification system	Statistical / prospective clinical study	NSLBP	7	Medical / back pain clinic (hospital)	History and clinical presentation
Barker (1990) / UK	Classification in GP practice	Judgement / prospective clinical study	LBP	6	Medical / GP practice	History, clinical presentation and investigations
Langworthy (1997) / UK	Experimental generation of clinical subgroups	Statistical / clinical trial	LBP	2	Chiropractic / chiropractic and orthopaedic clinics	History and clinical presentation
Coste (1991, 1992b) / France	Classification into clinical subgroups or syndromes	Statistical / clinical trial	NSLBP without psychiatric disorder	7	Medical / outpatient clinic	History, clinical presentation and psychiatric interview
Petersen (2003, 2004) / Denmark	Primarily patho-anatomic based classification	Judgement / literature review and clinical study	NSLBP	13 / 3	Physiotherapy / not applicable	History and clinical presentation
Sikorski (1985) / Australia	Categorisation into diagnostic groups for treatment	Judgement / clinical trial	LBP including nerve root pain	7	Medical / not specified	History, clinical presentation and radiographs
Kent (2005) / Australia	Classification into subgroups clinicians consider representative of LBP	Judgement / postal survey	NSLBP	5	Multidisciplinary / not applicable	History and clinical presentation
Key (2008a,b) / Australia	Classification based on movement impairment	Judgement / clinical observation and literature review	Back pain	6	Not reported / not applicable	Clinical presentation
Wang (2003) / Taiwan; Australia	Classification to determine treatment	Judgement / clinical study	NP	4 / several subcategories	Physiotherapy / GP practice	History and clinical presentation
McKenzie (1981); McKenzie (2003, 2006) / New Zealand	Classification to determine treatment	Judgement / clinical study	Spinal pain	3 / 2 (original)	Physiotherapy / outpatient clinic	History and clinical presentation
Laslett (1999) / New Zealand	Primarily patho-anatomic based classification	Judgement / proposal of a model	NSLBP	12	Physiotherapy / not applicable	History and clinical presentation

Spitzer (1987) / Canada	Classification to determine clinical decision making and prognosis (QTF)	Judgement / multidisciplinary team	Spinal Disorders (primarily LBP)	11 / 2	Multidisciplinary / clinics	History, clinical presentation and medical investigations
Binkley (1993) / Canada	Patho-anatomical classification based on PTs agreement levels	Judgement / Delphi survey	LBP	19	Physiotherapy / not applicable	History, clinical presentation and medical investigations
Moffroid (1994) / Canada	Classification based on physical measures	Statistical / clinical trial	LBP	5	Physiotherapy / 7 different settings	History, clinical presentation, disability and psychological questionnaires
Wilson (1999); McIntosh (2008) / Canada	Classification system to determine diagnosis and treatment direction	Judgement / reliability study	LBP (mechanical)	6	Physiotherapy / 10 back pain clinics	History and clinical presentation
Mooney (1989) / USA	Classification system	Judgement / proposal based on literature review	LBP	3 / 9	Medical (orthopaedics) / not applicable	Clinical presentation
Humphreys (1990) / USA	Patho-anatomical classification / professional approach	Judgement / proposal based on literature review	LBP	11	Chiropractice / not applicable	History, clinical presentation and radiography
MacDonald (1990) / USA	Patho-anatomical classification / professional approach	Judgement / proposal based on literature review	LBP	10	Osteopathy / not applicable	History and clinical presentation
DeRosa (1992) / USA	Classification system to determine treatment	Judgement / proposal of a classification model based on QTF	LBP (mechanical)	7 / 3	Physiotherapy / not applicable	History and clinical presentation
Rezaian (1993) / USA	Development of a practical aetiological classification system	Judgement / clinical trial	LBP	5	Medical (orthopaedics) / not specified	History, clinical presentation and investigations (e.g. x-rays)
Delitto (1995) / USA	Classification to determine treatment	Judgement / proposal based on pilot; RCTs, clinical studies	LBP (mainly acute)	4 / 2	Physiotherapy / not specified; outpatient clinics	History and clinical presentation
Marras (1995) / USA	Classification based on trunk motion measures	Judgement / clinical trial	Chronic LBP	10 of 11 QTF categories	Medical / not reported	Clinical presentation, medical investigations and trunk motion measures (with specific apparatus)
Newton (1997) / USA	Taxonomy of LBP subtypes referred to PTs	Judgement / clinical trials	LBP (acute and subacute)	17 (12 rare)	Multidisciplinary / not specified	History and clinical presentation
O'Hearn (1997) / USA	Modified QTF classification	Judgement / clinical study	LBP	9 / 2	Physiotherapy /	History, clinical presentation and

					physiotherapy clinic	investigations
BenDebba (2000) / USA	Modification of 4 QTF classification categories	Judgement / clinical trial	Chronic (persistent) LBP	4	Medical / 8 university affiliated tertiary care clinics	Clinical presentation and questionnaires
van Dillen (1998); Sahrman (2002) / USA	Classification based on movement impairment	Judgement / clinical studies	Mechanical LBP	5	Physiotherapy / outpatient clinics	Clinical presentation (symptom behaviour)
Childs (2004b) / USA	Classification to determine treatment	Judgement / proposal based on literature review, clinical studies	NP	4	Physiotherapy / not applicable	History and clinical presentation
McCormack (1990) / USA, Canada	Case finding to determine prevalence in a manufacturing workforce	Judgement	All neck upper limb disorders, particularly tendonitis / related disorders	4	Medical / occupational medicine	Clinical presentation and other data at the discretion of physician
Waris (1979) / Finland	Case finding / screening to determine prevalence, incidence of disorders in occupational health surveys	Judgement / literature review	Upper limb and neck disorders, related to work	10	Medical, physiotherapy, ergonomics / occupational medicine	Clinical presentations and special testing
Viikari-Juntura (1983) / Finland	Case finding / screening to determine prevalence, incidence of disorders in occupational health surveys	Judgement	Upper limb and neck disorders, known or anticipated relation to work	10	Physician / occupational medicine	Clinical presentaion
Spitzer (1995) / International	Classification for WADs	Judgement / literature review	WADs	5	Not specified	Clinical presentation and special testing
ICD-10 (WHO, 2001; WHO, 2004) / International	Statistical classification for many purposes	Judgement	Neck and upper limb disorders	3	Not specified	Not specified
Guzman (2008) / International	Classification based on a conceptual model intended to link epidemiology with management and consequences	Judgement / literature review	NP and associated disorders	4	Not specified	Clinical presentation and special testing

Psychosocial approach						
Main (1992) / UK	Classification based on measures of distress	Statistical / clinical trial	LBP	4	Medical (orthopedic) / orthopedic departments	Questionnaires and clinical presentation
Coste (1992a) / France	Clinical and psychological classification	Statistical / clinical trial	LBP	4	Medical / outpatient clinic	History, clinical presentation and questionnaires
Ozguler (2002) / France	Classification based on questionnaire items	Statistical / clinical trial	Chronic LBP	4	Medical / not specified	Questionnaires measuring functional, emotional, and fear related parameters
Bergstrom (2001a,b) / Sweden	Identification of chronic LBP subgroups	Statistical / clinical study	Chronic NSLBP	4	Medical (psychology) / multi-centre clinics	Questionnaires measuring psychosocial and behavioural parameters
Strong (1994, 1995) / Australia	Integration of 6 dimensions of LBP into one (multi-dimensional)	Statistical / clinical study	Chronic LBP	3	Occupational therapy / not specified	Questionnaires (pain, function, coping, depression, illness, etc.)
Keefe (1990) / USA	Classification based on observed pain behaviour	Statistical / clinical study	Chronic LBP	4	Medical (psychiatry)	Observed pain behaviour (during specific activities)
Krause (1994) / USA	Classification system based on social factors	Judgement / review and proposal of classification	Occupational LBP	8	Epidemiology / not applicable	Working status, insurance policies (compensation), and medical status
Klapow (1993, 1995) / USA	Exploration of social variables among LBP subgroups	Statistical / clinical trial	Chronic LBP	3	Medical / primary care orthopedic clinic	Questionnaires measuring psychosocial variables (life adversity, coping, social support)
Biopsychosocial approach						
Stiefel (1999a,b); Huyse (1999) / Switzerland	Classification system based on biopsychosocial factors to establish 'case complexity'	Judgement in development, statistical when testing subgroups / Cross-sectional study	Chronic LBP	4	Medical / in and outpatients	History and specific questioning in 4 domains (biologic, psychological, social and healthcare)
Harper (1992) / Australia	Taxonomy taking into account impairment, disability, and handicap due to LBP	Judgement / clinical trial	Occupational LBP of chronic nature	2 (primary impairments) / 12 (secondary) /	Multi-disciplinary / not specified	Interviews and questionnaires

				5 (tertiary)		
O'Sullivan (2005, 2006); Dankaerts (2007) / Australia	Classification based on motor control impairment	Judgement / proposal of classification based on literature review	Chronic LBP	3	Physiotherapy / not applicable	History, clinical presentation, physical examination and questionnaires (disability and fear-avoidance)
Halpern (2001)/ USA	Taxonomy of functional assessment constructs	Judgement / data obtained from clinicians	Chronic LBP	26	Medical / not applicable	Expert consensus

Abbreviations:
GP, General Practice; LBP, Low Back Pain; NP, Neck Pain; NSLBP, Non-Specific Low Back Pain; QTF, Quebec Task Force; UK, United Kingdom; USA, United States of America; WADs, Whiplash Associated Disorders.

1.2.3.1.1 The Quebec Task Force (QTF) classification

The Quebec Task Force on Spinal Disorders (Spitzer et al., 1987) was one of the first international groups to develop a classification system for spinal pain (Figure 1.2). The QTF based their system on signs and symptoms, imaging findings and response to treatment. Several authors have adopted and / or made adjustments to this classification (Marras et al., 1995; Atlas et al., 1996; O'Hearn, 1997; BenDebba et al., 2000). Although predictive validity has been established in some categories (Atlas et al., 1996; Loisel et al., 2002), its use is still limited with regard to making specific treatment choices (Murphy and Hurwitz, 2007; Billis et al., 2007; May et al., 2008a).

Figure 1.2 The Quebec Task Force classification system (adapted from Loisel et al., 2002, with permission).

QTF category	Definition	Duration of symptoms	Work status
1	Pain without radiation		
2	Pain with proximal radiation (above the knee)	< 7 days	
3	Pain with distal radiation (below the knee)	7 days to 7 weeks	Working or not working
4	Pain with distal radiation and neurologic signs	> 7 weeks	
5	Presumptive compression of a spinal nerve root on a simple roentgenogram		
6	Compression of a spinal nerve root confirmed by specific imaging techniques		
7	Spinal stenosis		
8	Post surgical 1-6 months after the intervention		
9	Post surgical > 6 months after the intervention		Working or not working
10	Chronic pain syndrome		
11	Other diagnoses		

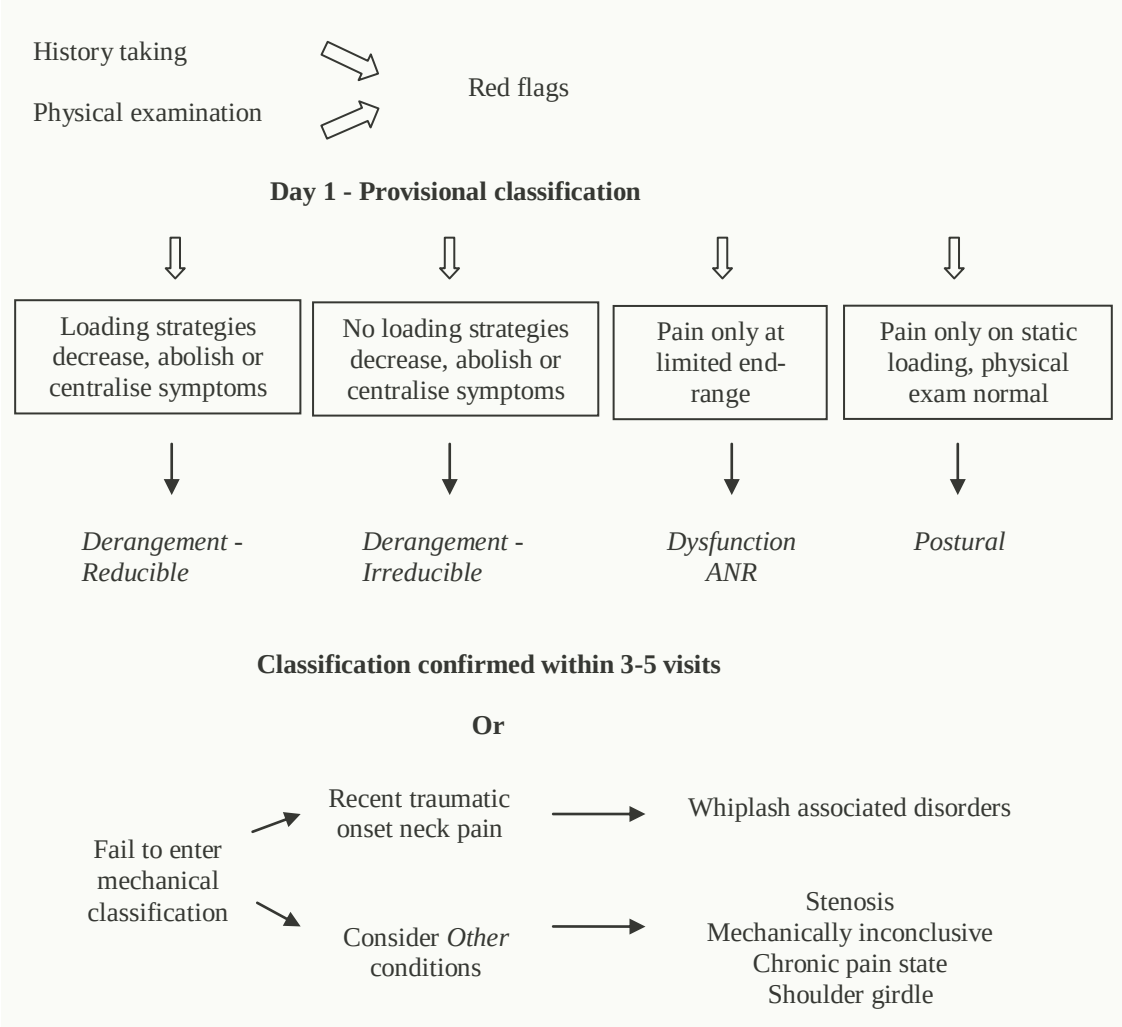
1.2.3.1.2 The McKenzie classification system

The McKenzie or Mechanical Diagnosis and Therapy (MDT) method classifies patients into treatment – based classification groups based on data from the history and physical examination (McKenzie, 1981; May and Donelson, 2008). Response to patient or therapist generated spinal loading strategies and a biomedical 'hands off' self-management orientation to classification characterise this system which was developed based on the clinical experience of a physiotherapist, named Robin McKenzie (McCarthy et al., 2004). The McKenzie system has been in common use for more than 20 years (Hefford, 2008). The original classification categorised patients with spinal pain into three main groups or syndromes: the postural, dysfunction and derangement syndrome (McKenzie 1981; McKenzie, 1990). This categorisation was criticised for not being exhaustive (Riddle, 1998; Murphy et al., 2008). The same system was later extended to include more groups that did not fit into the previous three categories following reassessment after 3-5 treatment sessions (McKenzie and May, 2003; McKenzie and May, 2006). A diagnostic algorithm (Figure 1.3) and definitions of the criteria for each category have been provided in relevant textbooks (McKenzie and May, 2003; McKenzie and May, 2006).

The McKenzie approach has been cited as an acceptable and practical approach to patient classification (McCarthy et al., 2004), and is popular for managing patients with back pain in North America (Battié et al., 1994; Li and Bombardier, 2001; Mikhail et al., 2005; Poitras et al., 2005), the UK (Foster et al., 1999; Jackson, 2001; Gracey et al., 2002) and Ireland (Byrne et al., 2006). Despite being recommended in LBP guidelines (Albright et al., 2001; Mercer et al., 2006), the assessment component of this classification (the precursor of any management option), and the

efficacy of the McKenzie approach in the treatment of spinal pain is not fully established (Aina et al., 2004; Clare et al., 2004; Machado et al., 2006). Further work is needed in the cervical or thoracic spine where evidence is still preliminary (Aina et al., 2004; Claire et al., 2004) and in patients with radicular symptoms (Busanich and Versscheure, 2006).

Figure 1.3 McKenzie classification algorithm for the cervical spine (adapted from McKenzie and May, 2006, with permission). ANR, Adherent Nerve Root.



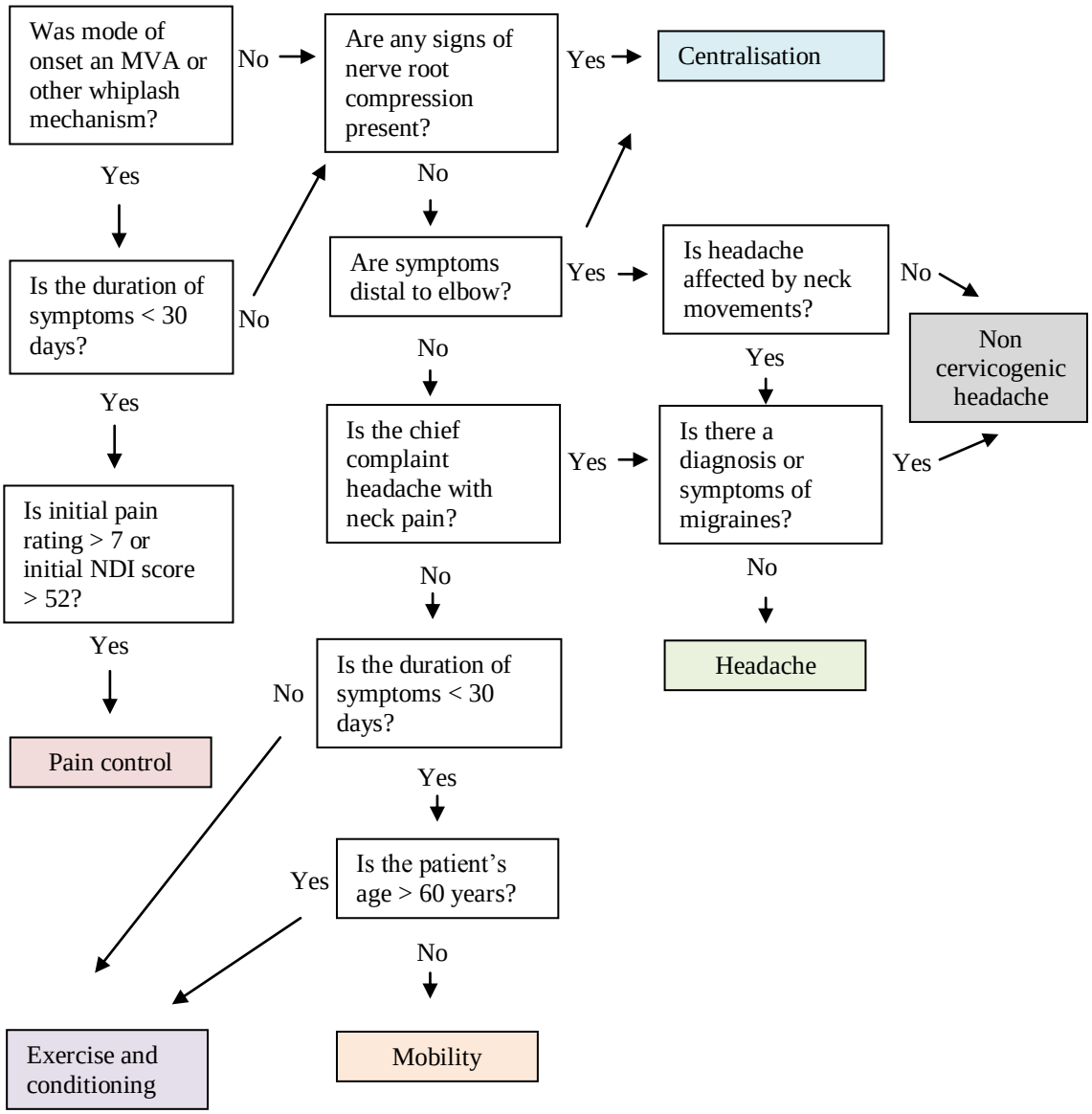
1.2.3.1.3 The Delitto classification system

Delitto et al. (1995) proposed a classification system for acute LBP. This system, often referred to as Treatment – Based Classification (TBC) system, was developed to identify subgroups of patients with similar signs and symptoms in their history and physical examination (Delitto et al., 1995). Evidence supporting the Delitto system has been reported in randomised controlled trials of LBP (Fritz et al., 2003; Brennan et al., 2006; Browder et al., 2007; Fritz et al., 2007b). However, the original system was criticised for its vagueness in the definition of acuity of symptoms and the inclusion / exclusion of other groups of patients based on arbitrary criteria (May et al., 2008b).

Although the inter-rater reliability of single categories has been questioned (Petersen, 2003), the TBC system has demonstrated moderate inter-rater reliability (kappa 0.56 to 0.60) (Fritz and George, 2000; Fritz et al., 2006). The system, developed in the US, has evolved since 1995 (Fritz et al., 2007a; Pinto et al., 2007). Clinical prediction rules have been added to identify additional LBP subgroups and clinical management pathways e.g. manipulation (Flynn et al., 2002; Childs et al., 2004a; Fritz et al., 2005; Brennan et al., 2006), stabilisation (Hicks et al., 2005), traction (Fritz et al., 2007b) and specific exercise (Parent, 2008).

A detailed description of the decision-making algorithm for classifying patients with LBP into the main subgroups was made by Pinto and associates (2007). A classification pattern (Figure 1.4) has also been described for NP (Fritz and Brennan, 2007).

Figure 1.4 Classification decision making algorithm proposed for neck pain. MVA, Motor Vehicle Accident; NDI, Neck Disability Index. Reprinted from Fritz and Brennan (2007) Preliminary examination of a proposed treatment-based classification system for patients receiving physical therapy interventions for neck pain. *Physical Therapy*, 87, 513-524 with permission of the American Physical Therapy Association. This material is copyrighted and any reproduction or distribution is prohibited..



1.2.3.2 The need for a shared paradigm in spinal pain classification

There is no widely accepted classification of spinal pain (Varamini and Jam, 2006b). Most classification systems are limited in use to the country of the system's developer (Billis et al., 2007). For example, Kent and Keating (2004) found that there was no agreement on an acceptable classification system among primary care clinicians, with 48 different assessment methods being used in LBP (Kent et al., 2009). The type of classification system and assessment methods used varied substantially by professional group (Kent et al., 2009). This has been hypothesised to account for observed treatment variations (Deyo, 1993; Cherkin et al., 1995). Further research is needed to develop new classifications and / or improve existing systems both in LBP (McCarthy et al., 2004) and NP (Buchbinder et al., 1996; Jull, 2004). The identification and inclusion of modifiable findings in classification systems, that can reliably predict patient outcomes has been identified as a high research priority in LBP and NP (Borkan et al., 1998; Haldeman, 2008; Carroll et al., 2008a).

1.2.3.3 The use of clinically induced symptom responses in spinal assessment

In this thesis, clinically induced symptom responses are defined as immediate changes in the status of symptoms resulting from physical examination or intervention strategies performed either directly on the spine or indirectly i.e. through structures that are connected or related to the spine (Borge et al., 2001). These types of responses are used frequently to assess the underlying acuity and nature of spinal pain (Tuttle, 2005; Murphy and Hurwitz, 2007), establish a diagnosis or prognosis (McKenzie, 1981; Maher and Latimer, 1992; Young et al., 2003; van Trijffel et al., 2008) or determine management strategies (McKenzie, 1981; Moffroid et al., 1994; Delitto et al., 1995; van Dillen et al., 1998; Petersen et al., 2004; Cook et al., 2005).

In a systematic review of prognostic factors of outcome in non-operative treatments of chronic LBP (Wessels et al., 2006) changes in pain intensity, sensation or unpleasantness demonstrated a stronger association with disability than changes in cognitive coping / appraisal and changes in physical performance. This finding was later confirmed in a study by the same authors (Wessels et al., 2007) where reductions in pain intensity explained the largest amount of variance in interference with daily life. Although symptom changes are important in clinical assessment (Matyas and Bach, 1985), symptom changes not associated with immediate responses to physical examination or treatment may be the product of non-specific effects and / or natural course (Bialosky et al., 2008). Reproducing or inducing immediate changes during clinical examination or treatment (i.e. clinically induced responses) has been reported as more important for accurate diagnosis and prognosis (Aina et al., 2004).

Traditionally, clinically induced symptom responses have focused on provoking or altering symptoms with a variety of spine loading strategies. Table 1.6 summarises spinal loading strategies in common practice. There are different schools of thought about which symptom responses are important. There is some favourable evidence that some clinically induced symptom responses may be reliable and prognostic of outcomes mainly in LBP (Chapters 2 and 3). However, the relative prognostic value of clinically induced symptom responses remains unclear in NP.

Table 1.6 Spinal loading strategies used to elicit clinically induced symptom responses.

Spinal loading strategy	Reference
Single trunk movements	Cyriax, 1982 Maitland and Edwards, 1986 Cailliet, 1988 Spratt et al., 1990 Moffroid et al., 1994 Delitto et al., 1995 van Dillen et al., 1998 Flynn et al., 2002 McKenzie and May, 2003 McKenzie and May, 2006 van Dillen et al., 2009
Repeated trunk movements	Spratt et al., 1990 Delitto et al., 1995 McKenzie and May, 2003; McKenzie and May, 2006 Wang et al., 2003
Combined trunk movements	Maitland, 1986 Edwards, 1994
Sustained end-range trunk positions	Moffroid et al., 1994 McKenzie and May, 2003 McKenzie and May, 2006
Segmental motion	Hubka and Phelan, 1994 Murphy and Hurwitz, 2007 Murphy et al., 2008 van Trijffel et al., 2008 Abbott et al., 2009
Muscle palpation	Travell and Simons, 1983 Murphy and Hurwitz, 2007 Murphy et al., 2008
Manual therapy (manipulation, mobilisation techniques)	Tuttle, 2005 Vernon and Humphreys, 2008
Neurodynamic testing	Petersen et al., 2004 Cleland et al., 2006b Murphy and Hurwitz, 2007 Murphy et al., 2008

1.2.3.4 The centralisation phenomenon

A commonly cited and important clinically induced symptom response for subclassifying LBP among UK physiotherapists (McCarthy et al., 2006) is the centralisation phenomenon. Centralisation was originally described by Robin McKenzie (1981; 1990) as a clinical phenomenon occurring when the patient reports that pain and referred symptoms originating from the spine move from a distal area to a location more central or near a midline position in the spine during spinal movement testing (Figures 1.5 and 1.6).

This sign is one of the core features of the McKenzie method of classification of spinal pain. Other classification approaches subsequently included centralisation in their assessment and management decisions, in LBP (Delitto et al., 1995; Laslett and van Wijmen, 1999; Wilson et al., 1999; Petersen, 2003; Petersen et al., 2003; Murphy and Hurwitz, 2007; McIntosh et al., 2008; Murphy et al., 2008) and NP (Wang et al., 2003; Fritz and Brennan, 2007).

Figure 1.5 Changes in pain location during centralisation and peripheralisation in (a) the lumbar spine (adapted from Donelson et al, 1997, with permission) and (b) the cervical spine (adapted from McKenzie and May, 2006, with permission).

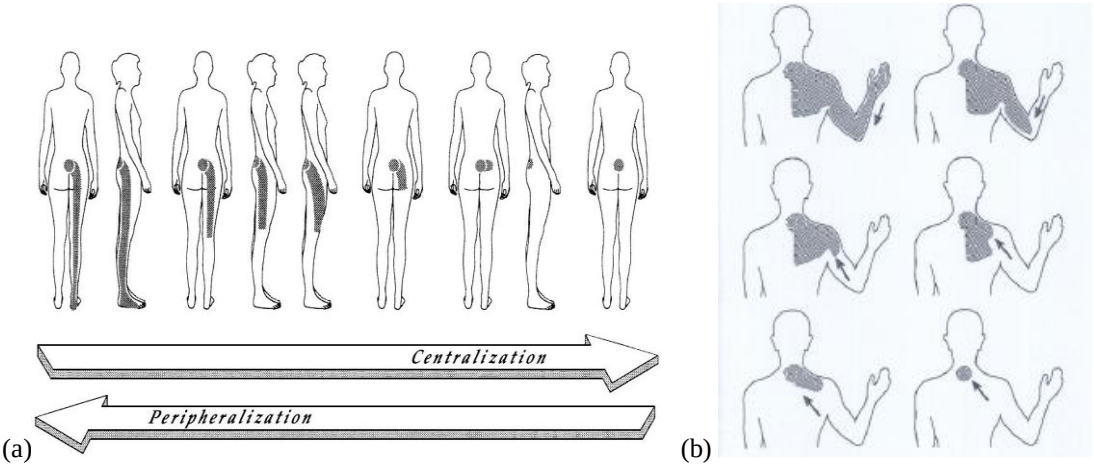
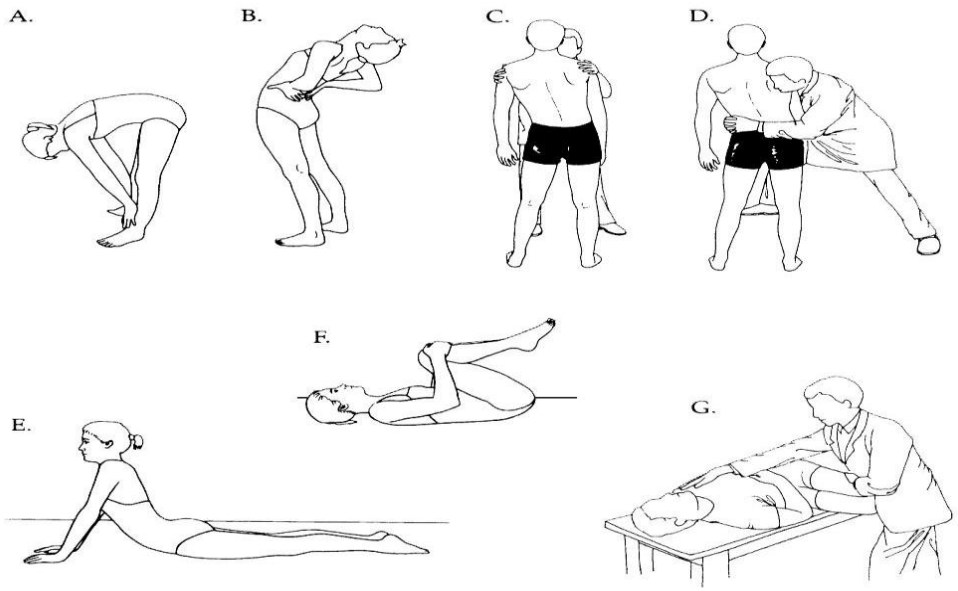


Figure 1.6 Examples of procedures commonly used to test for centralisation in the lumbar spine (adapted from Donelson et al, 1997, with permission). A. Flexion in standing. B. Extension in standing. C. Side gliding to the right. D. Manual correction of lateral shift. E. Extension in lying. F. Flexion in lying. G. Rotational mobilisation.



Centralisation can occur in the lumbar, cervical, and thoracic spine (McKenzie, 1981; McKenzie, 1990). However, most research has focused on LBP. Evidence from the lumbar spine does not necessarily transfer to the cervical spine. If this potentially important physical sign is to be used in the assessment of patients with NP, further research is needed. For centralisation to be of clinical utility, it must occur in a substantial proportion of those with NP, its definition must be consistent across groups, the phenomenon must be identified reliably, and its identification must be worthwhile for management, prognostic or diagnostic purposes (Aina et al., 2004).

1.2.3.4.1 Prevalence of centralisation

Centralisation is a frequent observation in the evaluation of non-specific spinal pain syndromes. In a meta-analysis of 1056 patients, Aina et al. (2004) found that

centralisation occurred in 70-87% of acute / sub-acute and 32-52% of chronic spinal patients. However, this conclusion was mainly based on older studies undertaken in LBP (Table 1.7). Before the start of this PhD thesis, the prevalence of centralisation in NP had been examined only in a sample (n=65) of patients with acute symptoms of NP, where 71% of patients had their symptoms centralised over consecutive visits (Werneke et al., 1999). Later studies (Cleland et al., 2007; Fritz and Brennan, 2007) reported a 35% centralisation rate in patients with NP, but these studies used a single assessment of active range of motion.

Table 1.7 Prevalence of centralisation in spinal pain. The underlined values were included in the review by Aina et al. (2004).

Reference	Sample Size	Patient Description	Prevalence
Kopp et al., 1986	67	Acute or exacerbated LBP with radiating symptoms, 100%	52%
Donelson et al., 1990	87	Acute 61%, subacute 17%, chronic 22% LBP	87%
Donelson et al., 1991	145	Acute 23%, subacute 38%, chronic 39%	47%
Delitto et al., 1993	24	Acute 100% LBP	61%
Erhard et al., 1994	24	Subacute 100% LBP	55%
Long, 1995	223	Chronic, not working 100% LBP	47%
Karas et al., 1997	126	Acute and chronic LBP, not working 100%	73%
Donelson et al., 1997	63	Chronic LBP 100%, not working 70%	49%
Sufka et al., 1998	36	Acute 16%, subacute 42%, chronic 42%	69%
Werneke et al., 1999	289	LBP 77%, NP 23% Acute 100%, not working 37%	77%
Flynn et al., 2002	71	Non radicular LBP, 100%	6%
Kilpikoski et al., 2002	39	Chronic 100% LBP	87%
Laslett et al., 2003	43	Chronic lumbo-pelvic pain	21%
Bybee et al., 2005	33	LBP with or without referred symptoms, 100%	91%
Laslett et al., 2005	69	Chronic LBP, 100%	32%
Laslett et al., 2006b	92	Chronic LBP, 100%	28.3%
Skytte et al., 2005	60	LBP and leg pain less than 14 weeks 100%, not working 60%	42%
Cleland et al., 2007	78	NP with or without referred symptoms, 100%	35%
Fritz and Brennan, 2007	274	NP with radicular symptoms (regardless of symptom duration)	34.7%
Schmidt et al., 2007	793	Subacute and chronic LBP with radiating symptoms, 100%	18% + 21%

Note: LBP, Low Back Pain; NP, Neck Pain.

1.2.3.4.2 Definition of centralisation

Although the core concept of centralisation has been established (Aina et al., 2004), its definition is inconsistent in the literature (Table 1.8). The standardisation of centralisation with the establishment of uniform criteria has been recommended frequently (George and Fritz, 2005; Werneke and May, 2005; Berthelot et al., 2007; Werneke et al., 2008).

Table 1.8 Examples of variations in the reports of centralisation in the literature.

Reference	**Assessment procedures**			**Criteria for CP**			**Timeframe**	
	SMov	RSMov	ThTech	Loc	Int	Neuro	SingV	ConV
Delitto et al., 1993; Delitto et al., 1995	+	+	-	+	?	+	+	-
Erhard et al., 1994	+	+	-	+	?	+	+	-
Long, 1995	-	+	-	+	-	-	+	-
Donelson et al., 1990) Donelson et al., 1997	-	+	-	+	-	-	+	-
Karas et al., 1997	-	+	-	+	+	?	-	+
Werneke et al., 1999	+	+	+	+	-	-	-	+
Fritz, 1998; Fritz et al., 2000a	+	+	-	+	-	+	+	-
Flynn et al., 2002	+	-	-	+	-	-	+	-
Petersen, 2003; Petersen et al., 2003; Petersen et al., 2004	+	+	+	+	-	?	+	-
Wang et al., 2003	-	+	-	+	-	-	+	-
McKenzie and May, 2003; McKenzie and May, 2006	+	+	+	+	-	-	-	+
Childs et al., 2004b	+	+	-	+	-	+	?	?
George et al., 2005	+	+	-	+	-	+	+	-
Tuttle, 2005; Tuttle et al., 2006	-	-	+	+	-	-	+	-
Laslett and van Wijmen, 1999; Laslett et al., 2005	-	+	-	+	+	-	+	-
McCarthy et al., 2006	+	-	-	?	?	?	?	?
Cleland et al., 2006b; Cleland et al., 2007	+	-	-	+	-	-	+	-
Fritz and Brennan, 2007	+	-	-	+	-	-	+	-
Fritz et al., 2007	+	+	-	+	-	-	+	-

Note: Loc = Abolition of distal pain; Int = Reduction in intensity of the most distal symptoms; Neuro = Improvement of neurological signs and symptoms; SMov = Single movements; RSMov=Repeated or sustained movements; ThTech = Therapist generated techniques; SingV = Single visit; ConV = Consecutive visits. (+ / -) represent the presence / absence of an item in the definition of centralisation respectively, (?) represents items whose inclusion in / exclusion from the definition of centralisation is not clear in the referenced paper.

1.2.3.4.3 Reliability of centralisation

A key aspect of any clinical test and the first property to be assessed is reliability. Reliability refers to the consistency, stability and reproducibility of a test (Sim and Wright, 2000) and represents the extent to which individuals can be distinguished from each other, despite measurement error (de Vet et al., 2003a). Two types of reliability are frequently reported: (1) intra-rater reliability and 2) inter-rater reliability (Sim and Wright, 2005). Intra-rater reliability refers to the agreement between ratings made by the same clinician in two or more occasions (Sim and Wright, 2005). Inter-rater (or inter-observer) reliability reflects the agreement between ratings made by two or more clinicians (Sim and Wright, 2005). Inter-rater reliability is often used as a measure of clinical performance (Haas, 1995) because it provides insights into the basis of management strategies and their soundness (APTA, 2001; Fritz and Wainner, 2001).

Centralisation has shown acceptable ($\kappa > 0.7$) inter-rater reliability amongst trained clinicians (Kilby et al., 1990; Werneke et al., 1999; Wilson et al., 1999; Fritz et al., 2000a; Kilpikoski et al., 2002), but not consistently (Cleland et al., 2006a; Fritz et al., 2006; Piva et al., 2006). Most studies on the reliability of centralisation have evaluated this phenomenon in patients with LBP and only a few studies (Werneke et al., 1999; Cleland et al., 2006a; Piva et al., 2006) have examined inter-rater reliability exclusively in patients with NP (Chapter 2).

1.2.3.4.4 Diagnostic implications of centralisation

Centralisation is hypothesised to indicate the intervertebral disc (McKenzie, 1981) as the source of pain. Several systematic reviews and studies have partially supported this assumption in LBP (Donelson et al., 1997; Laslett et al., 2003; Young et al., 2003; Laslett et al., 2005; Laslett et al., 2006a; Laslett et al, 2006b; Hancock et al., 2007). No diagnostic studies have been undertaken in NP.

1.2.3.4.5 Prognostic implications of centralisation

Centralisation has been associated with LBP outcomes (Wetzel et al., 2003; Aina et al., 2004; Berthelot et al., 2007). Very few studies have investigated the role of centralisation in predicting NP outcomes (Werneke et al., 1999; Tuttle, 2005; Tuttle et al., 2006; Cleland et al., 2007; May et al., 2008a).

1.2.3.4.6 Management implications of centralisation

Evidence is emerging regarding the value of centralisation in directing effective treatment in LBP (Fritz et al., 2003; Schenk et al., 2003; Long et al., 2004; Brennan et al., 2006; Browder et al., 2007). However, the use of a usual care or no treatment group to compare outcomes is not consistent across studies (Underwood et al., 2007). There are very few studies in NP and these mainly focus on the effectiveness of the treatment approach rather than the classification itself (Kjellman and Oberg, 2002; Klaber-Moffett et al., 2006).

1.3 SUMMARY

The usual management of spinal disorders starts with a history and physical examination, followed by further diagnostic tests when necessary. However, the history and physical examination have been traditionally evaluated for their ability to identify the underlying pathology and rarely for their value of predicting outcomes.

Current recommendations direct towards establishing classification approaches that distinguish patient groups suitable for treatment and identifying modifiable factors predicting spinal outcomes. Despite efforts, there is no such widely accepted spinal pain classification system or assessment procedure. Clinically induced changes in the patient's symptoms during assessment and subsequent treatment show promise, but their comparative value to other procedures remains to be tested, especially in NP.

A very commonly cited clinically induced symptom response procedure is the centralisation phenomenon. Although centralisation has been established as a useful and important physical sign in LBP, its definition, reliability, diagnostic ability and prognostic value is still unclear in NP.

1.4 OBJECTIVES OF THE THESIS

This thesis focuses on centralisation and the development of a standard definition of NP. In order to achieve this aim, the following objectives were set:

- Summarise and appraise the evidence on the reliability and prognostic value of clinically induced symptom responses in non-specific spinal pain;
- Develop and establish consensus on the operational definition of centralisation;
- Evaluate the inter-rater reliability of the identification of centralisation and related symptom response classification in patients with NP and explore sources of measurement error.

1.5 APPROACH

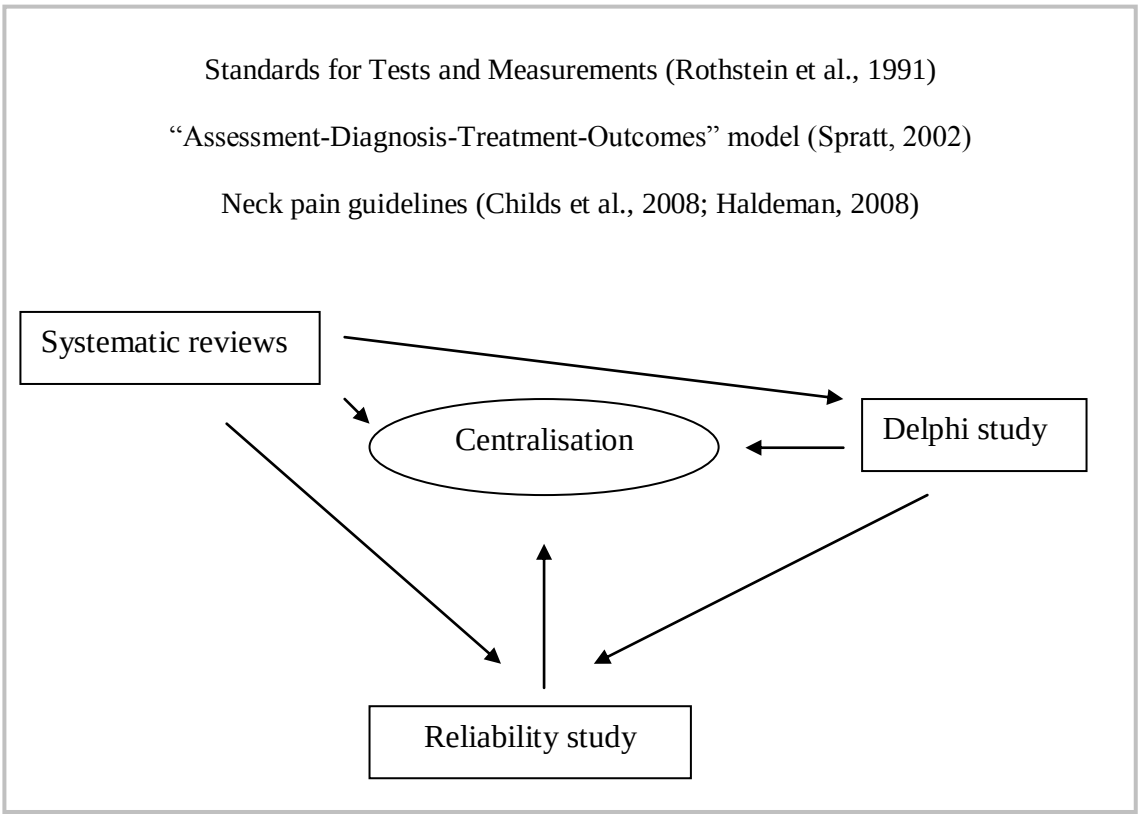
A number of principles were used to develop the operational definition of centralisation:

- Definition that was consistent with best practice for test development and research models intended to identify and validate subgroups in non-specific spinal syndromes;
- Definition that was evidence-based, i.e. considered current guidelines and literature as well as experts' opinions;
- Definition that could be delivered within the context of current practice in terms of staffing and time;
- Documentation to a standard that promoted consistency and enabled replication.

The stages used to develop the definition of centralisation and their inter-relationship are illustrated in Figure 1.7. Originally, two additional pieces of work were planned capturing both patients' as well as day to day clinician's perspectives, but these plans

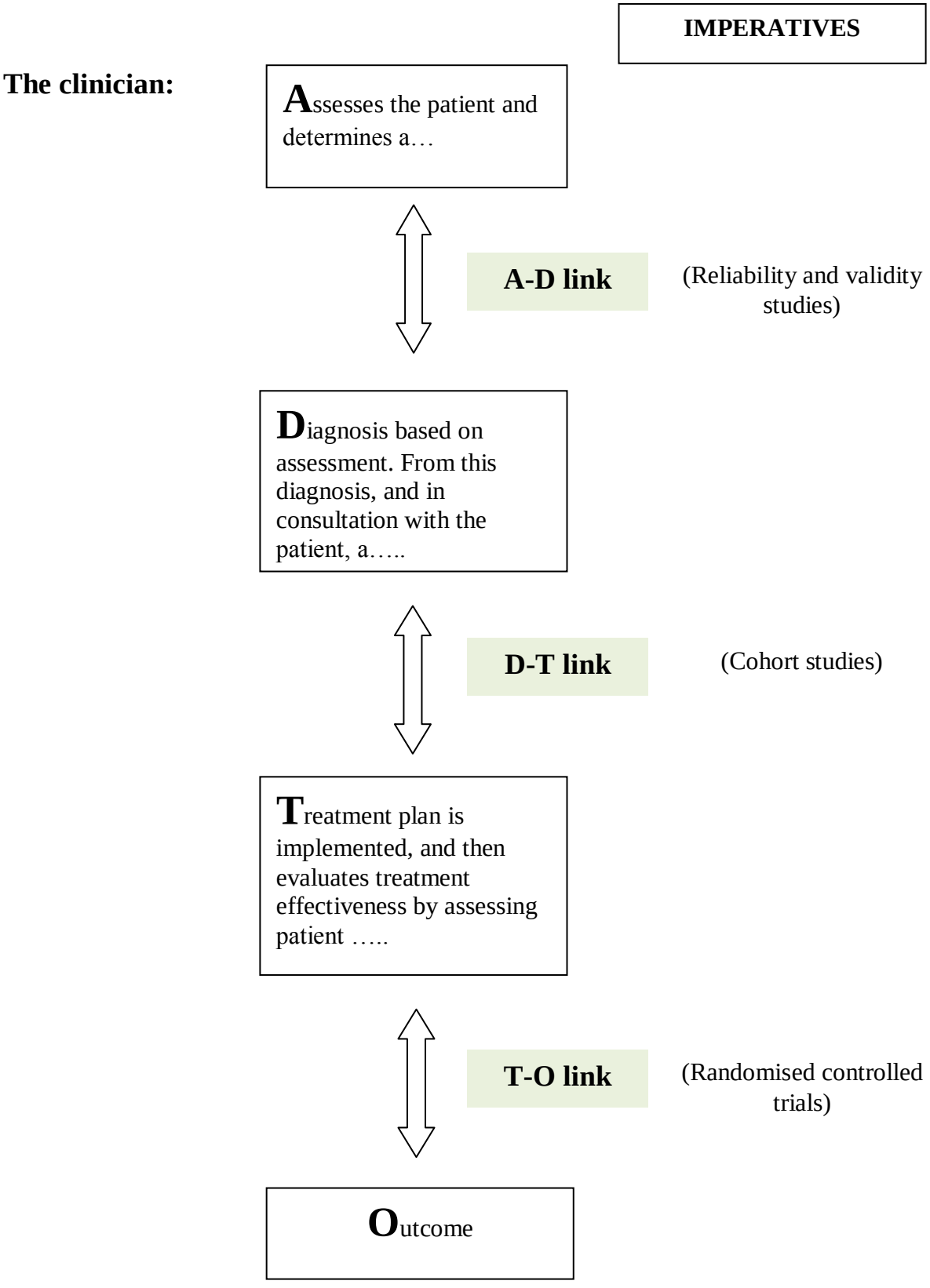
had to be abandoned due to time constraints. The first piece of work would have involved a qualitative study of the experiences of individuals with neck pain and would have been imbedded in the reliability study. The second piece of work was planned to involve the ratings and perspectives of clinicians from the UK and the international sector on the symptom response assessment of individuals with neck pain (VideoNeck study).

Figure 1.7 Stages used to develop the definition of centralisation and their inter-relationship.



Consequently, the stages within this PhD thesis were restricted to the context of a simpler model, the "Assessment-Diagnosis-Treatment-Outcomes" (ADTO) subgroup validation model (Spratt, 2002). This model argues that diagnosis (D) should be derived from a well defined assessment (A) (Spratt, 2002). For this diagnosis (D), there is an appropriate treatment (T) and specific outcomes of treatment (O) (Spratt, 2002). All the above inter-related elements should be validated through appropriate designs (Figure 1.8).

Figure 1.8 The "Assessment-Diagnosis-Treatment-Outcomes" model (Spratt, 2002). A-D, Assessment – Diagnosis link; D-T, Diagnosis – Treatment link; T-O, Treatment – Outcomes link.



1.5.1 The use of the mixed methodology approach

Traditionally, the use of quantitative data has been advocated in healthcare research (Reid, 1988). This may be understandable in the light of the dominant paradigm which favours positivistic approaches (Powell, 2003). However, attention to the use of other research paradigms in healthcare is slowly increasing. With the types of issues that this thesis was focusing on being complex and requiring different aspects and perspectives, the use of a mixed methods approach in the use of systematic reviews, a Delphi study and a reliability study was required. Further information is provided below and in the relevant chapters.

1.5.1.1 Systematic reviews of the literature

Systematic reviews can provide a comprehensive and reliable overview of available evidence by adhering closely to a scientific approach that is based on explicit, pre-specified and reproducible methods (Petticrew and Roberts, 2006; CRD, 2008). By serving as an accurate picture of past research, research plans can be placed into context, relevance can be established and the development of new or the refinement of past methodologies can be promoted (Petticrew and Roberts, 2006). The systematic reviews in Chapters 2 and 3 formed the basis on which research plans for this series of studies on centralisation were made. The reviews were intended to systematically summarise and evaluate evidence on a comprehensive range of clinically induced symptom responses as opposed to the limited number and types of responses investigated in previous reviews. This approach was challenging as it involved greater difficulty analysing and interpreting results and potentially more variable results (Dickersin and Berlin, 1992). However, it offered several advantages (Counsell, 1997; Dickersin and Berlin, 1992; CRD, 2008): (a) less risk for missing

potentially relevant studies due to narrow inclusion criteria (b) greater generalisability of the review findings, if no substantial variability was found; (c) if variability existed, exploration of what caused variations and generation of new hypotheses.

1.5.1.2 Delphi study

The aim of this study was to establish consensus on the criteria for the definition of centralisation and related symptom response groups. This is the first time a formal method has been applied to survey experts' attitudes to classifying patients to symptom response groups, and this step was intended to ensure face and content validity[1] of proposed definition. The results of this study were then used in conjunction with the results of the systematic reviews to inform further research steps.

1.5.1.3 Reliability study

After establishing consensus, a study was undertaken to assess the inter-rater reliability of centralisation and related symptom response classifications in 48 participants with NP. For the first time, this study also explored potential sources of measurement error and proposed strategies for improving reliability of these types of responses.

[1] Face validity refers to the 'perception that the people being measured, or the people administering the measures, have of the measure' (Clark-Carter, 2004). Content validity refers to the degree to which a measure covers the full range of what is being measured (Clark-Carter, 2004).

1.6 STRUCTURE OF THE THESIS

This thesis has been divided into six chapters. Chapter 1 provides a general introduction to the background and objectives of this thesis. Two systematic reviews on symptom response, one including reliability studies and the other studies of prognosis and their findings are presented in Chapters 2 and 3 respectively, justifying the rationale behind the proposed research in this thesis. The next stage involved establishing a consensus on the centralisation phenomenon. The operational definition of centralisation and future research suggestions made by experts participating in an international Delphi study are discussed and analysed in Chapter 4. Chapter 5 describes the methods and results of a reliability study in NP. Finally, Chapter 6 summarise the main findings of this thesis and discuss clinical and research implications in NP.

Chapter 2

The inter-rater reliability of the rating of clinically induced symptom responses in spinal pain: a systematic review.

2.1 AIMS OF CHAPTER

This chapter presents a systematic review of reliability studies on clinically induced symptom responses in spinal pain. The definition of clinically induced symptom response is presented in Chapter 1.The aim of this chapter was to:

- Present the range of clinically induced symptom responses that have been investigated in reliability studies and their inter-rater reliability in spinal pain;
- Discuss evidence in the context of the quality of research investigating these procedures;
- Analyse alternatives for measuring reliability and their relative strengths and weaknesses, with a view to have a point of reference for subsequent chapters (Chapter 5).

A decision to explore the influence of training and experience was made *a posteriori* following the results of the Delphi study (Chapter 4).

2.2 INTRODUCTION

Several systematic reviews of reliability of physical examination tests have been published recently. These reliability investigations include reviews of chiropractic (Hestbaek and Leboeuf-Yde, 2000) and clinical (van der Wurff et al., 2000; May et al., 2006) tests for the lumbo-pelvic region, as well as manual spinal examination procedures (Seffinger et al., 2004; van Trijffel et al., 2005; Hollerwöger, 2006; Stochkendahl et al., 2006).

Hestbaek and Leboeuf-Yde (2000) systematically reviewed the reliability of chiropractic tests for the lumbo-pelvic spine. Only pain provoked on palpation produced consistently acceptable results, although the authors acknowledged the

need for further evaluation of other tests which had not been sufficiently investigated. In the same year, a systematic review of reliability studies, this time on clinical tests for the sacro-iliac joint, was published (van der Wurff et al., 2000). In this review, sacro-iliac joint pain provocation tests showed greater promise than mobility testing, suggesting a necessity for further investigation of these procedures in the sacro-iliac area. Seffinger et al. (2004) investigated the reliability of spinal palpation in neck pain (NP) and low back pain (LBP) and found that only pain provocation and gross motion palpation procedures demonstrated reliability ≥ 0.4 in the highest quality studies. Similar to Hestbaek and Leboeuf-Yde (2000) and in keeping with Van Trijffel et al. (2005), the inter-rater reliability of passive intervertebral motion was low, with most studies characterised by poor methodological quality.

Another systematic review of 48 studies considered the inter-rater reliability of motion palpation, static palpation, palpation of osseous structures, soft tissue palpation, and global spine assessment (Stochkendahl et al., 2006). The pooled inter-rater reliability was acceptable (reliability threshold ≥ 0.4) for palpation of osseous structures ($\kappa = 0.53$) and soft tissue pain ($\kappa = 0.42$), but low for motion palpation ($\kappa = 0.17$) and soft-tissue changes ($\kappa = 0.03$). These findings were different from another review on cervical spine manual tests that did not find any substantial differences between pain-related procedures and passive motion palpation tests (Hollerwöger, 2006). However, the approach to analysis of reliability studies was different between these two reviews; one (Stochkendahl et al., 2006) pooled reliability estimates from included studies whereas the other (Hollerwöger, 2006) visually examined the range of observed values and drew conclusions based on these comparisons.

None of the above systematic reviews extended their investigation to other physical examination procedures commonly used in the day-to-day clinical assessment. In response to this knowledge gap, May and associates (2006) systematically evaluated the literature for reliability studies on examination procedures used in the assessment of non-specific LBP. Only pain changes in response to repeated movements demonstrated moderate evidence of high reliability (May et al., 2006). The authors of this review however, applied higher thresholds of reliability (≥ 0.7), resulting in most physical examination tests demonstrating conflicting results or low reliability.

Most of the above systematic reviews support the use of clinically induced symptom responses in the physical examination of patients with NP and LBP over other assessment procedures (Seffinger et al., 2004; May et al., 2006; Hestbaek and Leboeuf-Yde, 2000; van der Wurff et al, 2000; Stochkendahl et al., 2006). However, these reviews do not provide a comprehensive account of all the literature on clinically induced symptom responses, and they are restricted to one or two spinal areas. The latest year search strategies were performed is 2005 and new relevant studies have been reported since then. Some reviews recruited asymptomatic individuals; this may influence the validity of results since reliability may be inflated when asymptomatic participants are included (Haas, 1991; Lindsay et al., 1995). Different inclusion criteria, methodologies and thresholds were applied, resulting in variable conclusions about the same procedures across reviews. This raises uncertainty on the reliability of clinically induced symptom responses and the type of procedures that are most reliable in the physical examination of spinal pain.

2.3 MATERIALS AND METHODS

2.3.1 Selection criteria

The selection of studies was according to the criteria listed in Table 2.1. The inclusion and exclusion criteria were defined in terms of the spine population, clinically induced symptom responses, and reliability investigations. The whole range of clinically induced symptom responses rather than particular types of responses was considered. The rationale behind this approach is summarised in Chapter 1. Repeated as well as parallel inter-rater reliability study designs were acceptable for inclusion since focus was both on judgements as well as clinical performance.

Table 2.1 Eligibility criteria for the selection of reliability studies.

Inclusion criteria
1. Primary research
2. Inter-rater reliability study
3. Current episode of spinal pain with or without radiating symptoms (CSAG, 1994; Philadelphia Panel, 2001)
4. Adults (≥ 18 years)
5. Investigation of at least one clinically induced symptom response variable[2]
6. English language.
Exclusion criteria
1. Inclusion criteria are not met[3]
2. Recruitment of participants on the basis of specific patho-anatomical and / or other confirmed serious pathologic conditions
3. Investigation of pain behaviours[4]
4. Full text unavailable

[2] Studies using clinically induced symptom response categories to classify patients into diagnostic / syndrome groups were also eligible provided that the end result of the positive / negative response was directly related to patient grouping.
[3] This also refers to studies where it was impossible to determine the nature of the investigated symptom responses or no separate analysis was reported.
[4] Pain behaviours refer to changes that reflect an expression of pain such as guarding, bracing, grimacing (Jensen et al., 1989).

2.3.2 Search strategy and selection of studies

A search strategy was developed by the author of this thesis (Angeliki Chorti) to identify relevant studies. Ovid-MEDLINE, Ovid-EMBASE, Ovid-CINAHL and Ovid-AMED were searched from inception up to March 2007. These databases were selected because of their relevance to the nature of the review question (CRD, 2008). The online databases search strategy is presented in Appendix 2.1. All databases were searched using a template based on the research question being broken down into separate components (CRD, 2008). Combinations of keywords and Medical Subject Headings (MeSH) terms were used. Proposed steps involved in developing adequate search strategies and search terms for reliability studies (Murphy et al., 2003b) were followed (Figure 2.1). Where appropriate, the term / key words were modified to optimise the yield of relevant citations from the individual databases. These adjustments were made to account for any indexing variations often observed in different databases (Murphy et al, 2003a). Because reliability studies are poorly indexed in databases (van Trijffel et al., 2005), the reference lists of identified systematic reviews and articles were also searched. Finally, forward citation tracking was performed for all included articles using the Web of Science (Bakkalbasi et al., 2006).

Figure 2.1 Steps involved in constructing a search strategy (adapted from Murphy et al., 2003b with permission).

A. *Break down the research question, "What is the reliability of clinically induced symptom responses?"* to the 3 relevant components: reliability, spine, and procedure terms

B. *Identify specific Medical Subject Headings / key terms and their variations for each component*

C. *Apply Boolean operators to formulate a search strategy. For each component, expand terms using the "OR" operator. The result of each set is combined using the "AND" operator*

The details of the retrieved citations were scanned by the author of this thesis (Angeliki Chorti) to determine eligibility according to the aforementioned selection criteria. Two reviewers (Angeliki Chorti and Anastasios Chortis) then independently read the full text of eligible citations or citations where the eligibility could not be determined. Any disagreements between reviewers were explored and resolved through discussion. If further information was required, contact with the authors of the studies was sought. If disagreement persisted, the decision of another reviewer (Nikolaos Strimpakos) was used.

2.3.3 Quality assessment

There is no widely accepted and validated method for assessing the quality of reliability studies (van Trijffel et al., 2005; May et al., 2006). Therefore, the selection of a quality assessment tool rested on the careful consideration of prior instruments and guidelines.

A quality assessment instrument developed in a systematic review of physical examination procedures in LBP (May et al., 2006) was used (Table 2.2). This tool is comprehensive and relevant to the study context and was selected for its clarity, provided by the operationalisation of its quality criteria. As cited in the original publication, the selection of quality criteria was made to represent areas of external validity, internal validity and statistical methodology (May et al., 2006), important domains in other systematic reviews of reliability of physical tests (Hollerwöger, 2006; Stochkendahl et al., 2006; Myburgh et al., 2008). Modifications in the operationalisation of criteria were made to ensure relevance to this study.

Seven of the criteria originate from evidence of variation and design-related bias in diagnostic studies (Lijmer et al., 1999; Whiting et al., 2004), the standards for reporting of diagnostic studies (Bossuyt et al. 2003) and a validated tool for assessing quality of diagnostic studies (Whiting et al., 2003). There is, as yet, no evidence on the applicability of these items to the reliability context or any empirical evidence of methodological bias (van Trijffel et al., 2005). Some items were added to fit the context of reliability (May et al., 2006), based on relevant theoretical considerations on the design and conduct of reliability studies (Cohen, 1960; Maclure and Willett, 1987; Thompson and Walter, 1988; Feinstein and Cicchetti, 1990; Altman, 1991; Haas, 1991; Brennan and Silman, 1992; Byrt et al., 1993; Streiner and Norman, 2003).

Each study was independently assessed by two raters (Angeliki Chorti and Anastasios Chortis) who were not blind to authors and journal information because of their familiarity with the literature. Disagreements were resolved through discussion / consensus and if disagreement persisted, the opinion of the third rater (Nikolaos Strimpakos) was sought. The quality score of included studies was used as part of the determination of evidence (May et al., 2006). The maximum score that a study could achieve was 100. There is no consensus on the threshold that determines a high from a low quality study. However, thresholds of 50% (Stochkendahl et al., 2006) and 60% or above (May et al., 2006; Myburgh et al., 2008) have been used in previous systematic reviews.

Table 2.2 Quality assessment instrument for reliability studies (adapted from May et al., 2006 with permission).

CRITERIA (TOTAL 100)
Study population (total 25) 1 Adequate description of study population (4): there is a description of inclusion / exclusion criteria (1 point), number of participants (1 point), study participants characteristics i.e. demographic and clinical (1 point), and numbers of withdrawals and drop-outs (1 point); score 4 if all of the above are adequately described. 2 Representative of clinical practice (4): score 4 based on the description of the source population and the way patients were recruited; 0 if unlikely to be representative or uncertain. 3 Subjects selected randomly or consecutively (7): a score of 7, if the study explicitly states that subjects were selected randomly or consecutively. For random assignments, methods of allocation using date of birth, date of admission, hospital numbers or alternation should not be considered appropriate. 4 Number of subjects (10): score 10 if the study provides a justification for the sample size used; If there is no justification, the following scoring should apply: if < 25, score 0; > 25, score 3; > 50, score 6; > 75, score 10.
Study conditions (total 35) 5 Procedure clearly described and reproducible (5): for test-retest designs, score 5 if there is both a clear description of the procedure used (in the text or referenced) (4 points) and the time interval between examinations is described (1 point); for concurrent or videotaped examinations score 5 if there is a clear description of the technique used (text or reference). 6 Procedure executed in a uniform manner (5): score 5 if the same procedure has been executed among examiners. 7 Adequate measures to reduce bias (10): score 10 if examiners were blinded to the findings of other examiners e.g. examiner blinded to the other examiner's findings, results were sealed, and there was an independent adjudicator in parallel examinations. 8 Level of examiners (10): if experienced with procedure, score 10; if experienced clinicians or including a subset of experienced with procedure, score 5; if students / juniors, score 2. 9 Consensus/ training procedure prior to testing with pilot study (5): score 5 if study explicitly states consensus or training among examiners and / or pilot phase.
Study results (total 40) 10 More than one pair of examiners tested (10): score 10 if more than one pair of examiners were used and tested for agreement among patients. 11 Multiple testing between examiners (5): score 5 if multiple testing between examiners. 12 Standardised measure of test outcome: score 5 if there is a dichotomous and/or clear description of the outcome. 13 Frequencies of outcome and agreement reported (10): score 10 if both frequencies and agreement outcomes are reported; 5 for frequencies of outcome (e.g. categorical: frequency counts / continuous means, standard deviations) and 5 for agreement values. 14 Appropriate statistics (10): score 10 if appropriate statistics were used e.g. kappa for binary data / weighted kappa for ordinal data / intraclass correlation coefficient for continuous data (7 points) and measures of variance (3 points).

2.3.4 Data extraction

Data extraction is 'the process by which researchers obtain the necessary information about study characteristics and findings from the included studies' (CRD, 2008 p.28). Table 2.3 presents the information that was extracted, in keeping with recommendations of reliability systematic reviews and guidelines on diagnostic tests (Bossuyt et al., 2003, CRD, 2008; Myburgh et al., 2008). Data extraction was performed by the author of this thesis (Angeliki Chorti).

Table 2.3 Types of extracted information from included reliability studies.

Data extraction on:

- **General information** (e.g. authors, title, citation, type of publication, country of origin and source of funding)
- **Study details** (e.g. aims / objectives of study, study design, inclusion / exclusion criteria, recruitment procedures)
- **Participants** (e.g. sampling strategy, number and characteristics i.e. proportion of males / females, age, symptomatic area)
- **Examiners** (e.g. setting, inclusion / exclusion criteria, number, profession, practicing experience, experience with procedure or training)
- **Assessment procedure** (e.g. type and details of procedure eliciting symptom response, spinal area, blinding)
- **Outcome measures and judgement criteria** (e.g. definition and rationale for the units, cut-off points, categories)
- **Statistical analysis** (methods for calculating or comparing inter-rater reliability and quantifying uncertainty).

2.3.5 Overview of inter-rater reliability statistics

The issue of the most appropriate reliability statistic is unclear (Ludbrook, 2002; Streiner and Norman, 2003). Nevertheless, the way that a variable is measured determines which statistical methods are appropriate (Agresti, 2002). Considerations of the strengths and limitations of the most common approaches are presented below by type of data / level of measurement.

2.3.5.1 Categorical data

Categorical data represent types of data which can be divided into a set of categories or groups (Agresti, 2002). Categorical data can be classified into nominal or ordinal level data (Agresti, 2002; Sim and Wright, 2005). Nominal level data do not have a natural ordering and refer to judgements in relation to discrete categories e.g. “yes” or “no” , “male” or “female” (Agresti, 2002; Jill and Blackman, 2004; Sim and Wright, 2005). Ordinal level data, have ordered categories e.g. “mild”, “moderate”, “severe”, but distances between categories are unknown (Agresti, 2002).

Analysis of categorical data in inter-rater reliability studies is generally based on the use of tables, where the number of observations falling into each category is presented and divided into rows and columns (an example is given in Table 2.4). The totals for each category, known as marginal distributions (grey cells in Table 2.4), represent the number of individuals in each row or column, without accounting for the effect of the other rater.

Table 2.4 Table for categorical data in a reliability study of assessments of 85 xeromammograms by two radiologists (adapted from Boyd et al., 1982, with permission). Numbers in each cell represent classifications made by the two radiologists.

	Radiologist B				
Radiologist A	Normal	Benign	Suspected cancer	Cancer	Total
Normal	21	12	0	0	33
Benign	4	17	1	0	22
Suspected cancer	3	9	15	2	29
Cancer	0	0	0	1	1
Total	28	38	16	3	85

2.3.5.1.1 Percentage or proportion agreement

The investigation of inter-rater reliability started historically with the calculation of percentage (or proportion) agreement (Haas, 1991; Jill and Blackman, 2004). The ratio of the number of agreements between observers to the total number of comparisons made (i.e. overall agreement) was most commonly used over other approaches (Haas, 1991).

Exclusive reporting of percentage agreement was soon abandoned because this approach did not account for chance agreement (Thompson and Walter, 1988; Haas, 1991; Banerjee et al., 1999). In other words, if examiners' agreement was due to chance, it was impossible to establish whether and to what degree they really agreed or not (Sim and Wright, 2005).

2.3.5.1.2 Measures of association

Another approach followed in early inter-rater reliability studies of categorical data was based on tests of association (strength of association and statistical significance tests). The chi-square (χ^2) and Cochran`s Q were commonly used for nominal data

(Haas, 1991; Banerjee et al., 1999), whereas Kendall's coefficient of concordance was used for ordinal data (Armstrong, 1981). Measures of association were considered inappropriate for reliability reporting because they do not necessarily equate with agreement, especially when outcomes have more than two categories (Light, 1971; Banerjee et al., 1999). An example is given in Table 2.5.

Table 2.5 Example illustrating the difference between agreement and association (Light, 1971).

Suppose for a certain data set, cells B and C are zero and cells A and D are non-zero. A second data set has zeros in cells A and D and non-zero values in B and C. Both tables can give the same χ^2 even though one represents maximal agreement and the other maximal disagreement.		Rater 1	
		Presence	Absence
Rater 2	Presence	A	B
	Absence	C	D

2.3.5.1.3 Scott's π

Scott (1955) was one of the first to introduce a chance-corrected measure of inter-rater reliability for nominal scale categories, known as pi (π). Although this measure targeted research where subjective ratings were involved (Scott, 1955), the marginal probabilities of a positive finding between raters were not allowed to differ i.e. each rater should have the same probability of finding a positive finding.

2.3.5.1.4 The kappa statistic

Scott's π was extended by Cohen (1960) who proposed kappa (κ) as an alternative chance – corrected measure of reliability (Landis and Koch, 1977). In contrast to Scott's π, Cohen's κ (1960) allowed for the marginal probabilities of success

associated with the raters to differ (Light, 1971; Banerjee et al., 1999). Several expansions of kappa have been developed, although some are actually extensions of Scott's π rather than Cohen's κ (Jill and Blackman, 2004). These adaptations include intraclass kappas (Kraemer and Bloch, 1988), differential weighing (Cohen, 1968), conditional agreement (Light, 1971), as well as multiple response categories (Kraemer, 1980) and raters (Fleiss, 1971; Fleiss and Cuzick, 1979).

2.3.5.1.5 Modelling agreement for categorical data

The above approaches focus primarily on summary measures of agreement whose purpose is to indicate the *degree* of reliability between raters (Agresti, 1992; Gwet, 2008). Statistical modelling is proposed to investigate the *pattern* of agreement between raters. Some examples are (Agresti, 1992): (1) log-linear models for square tables (e.g. quasi-independence and quasi-symmetry models) (Agresti 1988; Becker and Agresti, 1992; Valet et al., 2007); (2) latent class models reflecting the joint distribution between ratings as a mixture of clusters for homogeneous participants, each cluster having the same 'true rating' (Tanner and Young, 1985; Guggenmoos-Holzmann and Vonk, 1998); and (3) Rasch models decomposing participant by rater rating distributions using rater and participant main effects. Modelling is particularly recommended when subjective ratings are involved (Becker and Agresti, 1992; Roberts, 2008) or when the purpose is to detect rater bias[5] (Ludbrook, 2002). Unfortunately, developed models seem to be more appropriate for ordinal rather than nominal data, and more importantly, they are difficult to understand and interpret (Ludbrook, 2002).

[5] Rater bias is the extent to which raters disagree on the proportion of positive (or negative) cases and is reflected in the difference between cells b and c in Table 2.5 (Sim and Wright, 2005).

2.3.5.2. Continuous data

Reliability and agreement studies may involve continuous data when the ratings in question are on a continuous scale, e.g. height, weight, or range of motion (Landis and Koch, 1977; Jill and Blackman, 2004; Sim and Wright, 2005).

2.3.5.2.1 Comparison of means

One of the first methods of assessing inter-rater agreement in continuous data was by comparing means of each rater. Agreement was then determined by whether the two raters gave the same mean measurement, through a statistical significance test (Altman and Bland, 1983). Despite the appeal of this approach, little information was conveyed on the agreement of compared methods.

2.3.5.2.2 Measures of association

An early approach using association or correlation coefficients was to calculate Pearson's product-moment correlation coefficient, *r*, between the two methods of measurement, but this approach suffered from the same limitations discussed above for association measures in categorical data (Altman and Bland, 1983). Pearson's *r* was later substituted by the intraclass correlation coefficient (ICC). Six versions of the ICC are reported as measures of inter-rater reliability depending on the research question being asked, but not all are considered appropriate by some (Armstrong, 1981).

2.3.5.2.3 The Bland-Altman method

Bland and Altman (1983, 1986) proposed graphical techniques as a first step to investigating agreement between two methods of clinical measurement (Table 2.6). For example, for measurements A and B, their suggestion was to plot the difference between the methods (A-B) against the (A+B) / 2 average (Figure 2.2) instead of plotting with a regression line drawn through the data (Figure 2.3) (Altman and Bland, 1983). The advantage of this approach was that systematic differences between investigated methods or random variation according to the mean value could be clarified, and if observed, summarised with appropriate measures (Altman, 1991) e.g. analysis of differences after a logarithmic transformation (Altman and Bland, 1983).

Table 2.6 PEFR measured with Wright and mini Wright peak flow meter. PEFR, Peak expiratory flow rate. Please note that only the first measurement of each method was used for illustrative purposes (adapted from Bland and Altman, 1986 with permission).

Subject	**Wright peak flow meter**		**Mini Wright peak flow meter**	
	First PEFR (l / min)	**Second PEFR (l / min)**	**First PEFR (l / min)**	**Second PEFR (l / min)**
1	494	490	512	525
2	395	397	430	415
3	516	512	520	508
4	434	401	428	444
5	476	470	500	500
6	557	611	600	425
7	413	415	364	460
8	442	431	360	390
9	650	638	658	642
10	433	429	445	432
11	417	420	432	420
12	656	633	626	605
13	267	275	260	227
14	478	492	477	467
15	178	165	259	268
16	423	372	350	370
17	427	421	451	443

Figure 2.2 Example of data plotting in the Bland – Altman method (adapted from Altman and Bland, 1983, with permission).

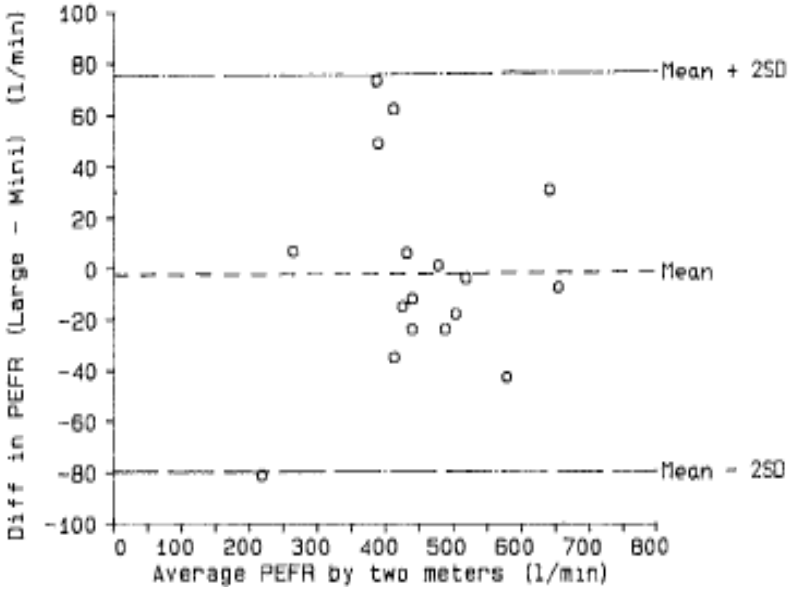
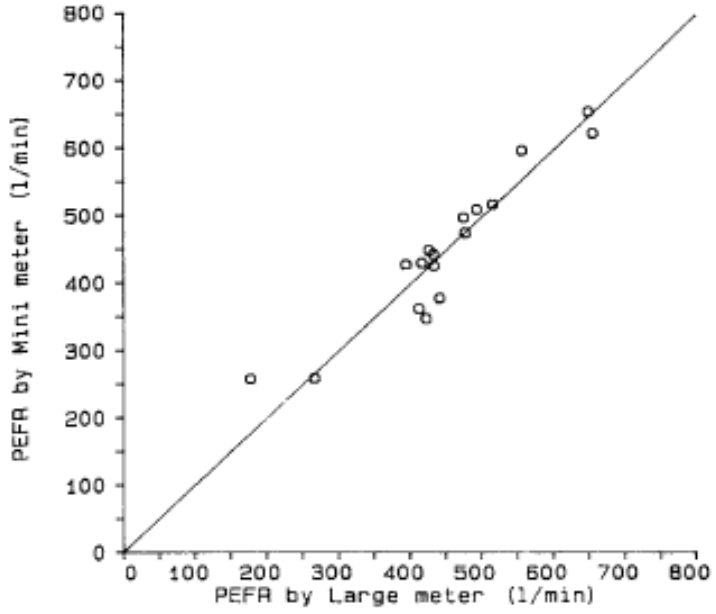


Figure 2.3 Example of data plotting using the regression line (adapted from Altman and Bland, 1983, with permission).



2.3.5.2.4 Modelling agreement for continuous data

Modelling approaches have also been proposed for inter-rater reliability analyses using continuous scales. Some examples (Altman and Bland, 1983) are: (1) least squares regression, (2) principal component analysis. Again, the considerable complexity of such analyses makes their use impractical for use by clinicians (Altman and Bland, 1983).

2.3.6 Investigation of heterogeneity

In the context of this reliability review, heterogeneity refers to the variability among reliability studies and is generally divided into clinical, methodological and statistical diversity (Deeks et al., 2008). Clinical diversity in reliability studies may arise from differences in the participant or rater population, investigated clinically induced symptom responses, and outcomes between studies, whereas methodological heterogeneity relates to study design and risk of bias (Deeks et al., 2008). Statistical heterogeneity (Deeks et al., 2008), represents differences in results from studies in terms of the degree of agreement, or the direction of agreement (Dickersin and Berlin, 1992). Uncertainty exists on the methods for identifying heterogeneity in systematic reviews of tests (Dinnes et al., 2005). Some methods (Donner and Klar, 1996) have been developed for reliability data, but their use is still questionable when there is clinical and methodological heterogeneity. Thus, heterogeneity was first explored through a visual inspection of the clinical and methodological study characteristics and the results of the individual studies. A visual inspection of raw data of included studies was also attempted (Brennan and Silman, 1992; Roberts, 2008), but this was not always feasible because of poor reporting quality.

2.3.7 Data synthesis considerations

Extracted data were combined into descriptive data and presented in summary tables, one describing study characteristics and two reporting on results of reliability studies for each spinal area. Clinically induced symptom responses were collated into types of responses based on the characteristics in included papers. Results from included studies were drawn or calculated from the original articles in the form of point and variance estimates (Sim and Wright, 2005). Where possible, the reliability statistic or a range of values for the statistic was presented with 95% confidence intervals (95% CI) for included studies.

Synthesis of reliability studies in systematic reviews is a challenging task. Problems in reviews of reliability studies arise from reliability not being an established property that a particular test does or does not have, but a reflection of the performance of a test when applied to a certain population under certain conditions (Streiner and Norman, 2003). Reliability measures are unstable and influenced by prevalence and rater bias (Feinstein and Cicchetti, 1990; Brennan and Silman, 1992; Byrt et al., 1993; Schuster, 2004). This makes comparisons of reliability coefficients across studies difficult to interpret (Armitage and Berry, 2002) and synthesis of reliability data inappropriate unless the above issues are addressed (Thompson and Walter, 1988; Altman, 1991).

There is no established method of meta-analysis for reliability studies. Some authors provided formulae for combining reliability coefficients (Charter, 2003) whereas others have suggested methods of comparing kappa statistics in multiple independent samples (Donner and Klar, 1996). Both approaches suffer from methodological

limitations, and problems of prevalence and bias are only partly addressed. It was decided *a priori* that a meta-analysis would be performed if the comparability between study characteristics and findings allowed data to be combined, there was an adequate amount of data (≥ 2 studies) and study quality and reporting was satisfactory (Tierney et al., 2007; CRD, 2008). If these requirements were not met, a qualitative approach to the synthesis of data would be followed.

2.3.8 Interpretation of reliability statistics

Reliability values normally range from -1 (perfect disagreement or less than chance agreement) to 1 (perfect agreement) (Fleiss, 2003). Proposed classifications for the kappa statistic (Landis and Koch, 1977; Cicchetti and Sparrow, 1981; Altman, 1991) weighted kappa (Cicchetti and Sparrow, 1981; Fleiss, 1986; Fleiss, 2003) and the ICC (Shrout and Fleiss, 1979) are presented in Tables 2.7 and 2.8. These classifications are arbitrary, and some have questioned their appropriateness (Brennan and Silman, 1992; Ludbrook, 2002). There is no consensus about what constitutes a clinically acceptable level of reliability in studies investigating physical examination procedures (May et al., 2006). Some advocate the 0.7 threshold for a test to be useful clinically (Hripcsak and Heitjan 2002; van Trijffel et al., 2005; May et al., 2006). Others propose 0.4 as the minimum (Seffinger et al., 2004; Stochkendahl et al., 2006; Myburgh et al., 2008). Confusion exists because the impact of reliability on prognostic or treatment performance is unclear.

Table 2.7 Examples of empirical approaches to evaluating the level of agreement in categorical variables (Landis and Koch, 1977; Altman, 1991). *κ*, kappa statistic.

		Landis & Koch (1977)	Altman (1991)
Levels of agreement *κ*	Poor	< 0.00	< 0.20
	Slight	0.00 - 0.20	
	Fair	0.21 - 0.40	0.21 - 0.40
	Moderate	0.41 - 0.60	0.41 - 0.60
	Good / Substantial	0.61 - 0.80	0.61 - 0.80
	Very good / Almost perfect	0.81 - 1.00	0.81 - 1.00

Table 2.8 Examples of an empirical approach to evaluating the level of agreement in continuous variables (Shrout and Fleiss, 1979). *ICC*, intraclass correlation coefficient.

		Shrout & Fleiss (1979)
Levels of agreement *ICC*	None	< 0.10
	Slight	0.11 - 0.40
	Fair	0.41 - 0.60
	Moderate	0.61 - 0.80
	Substantial	0.81 - 1.00

2.3.9 Sensitivity analysis

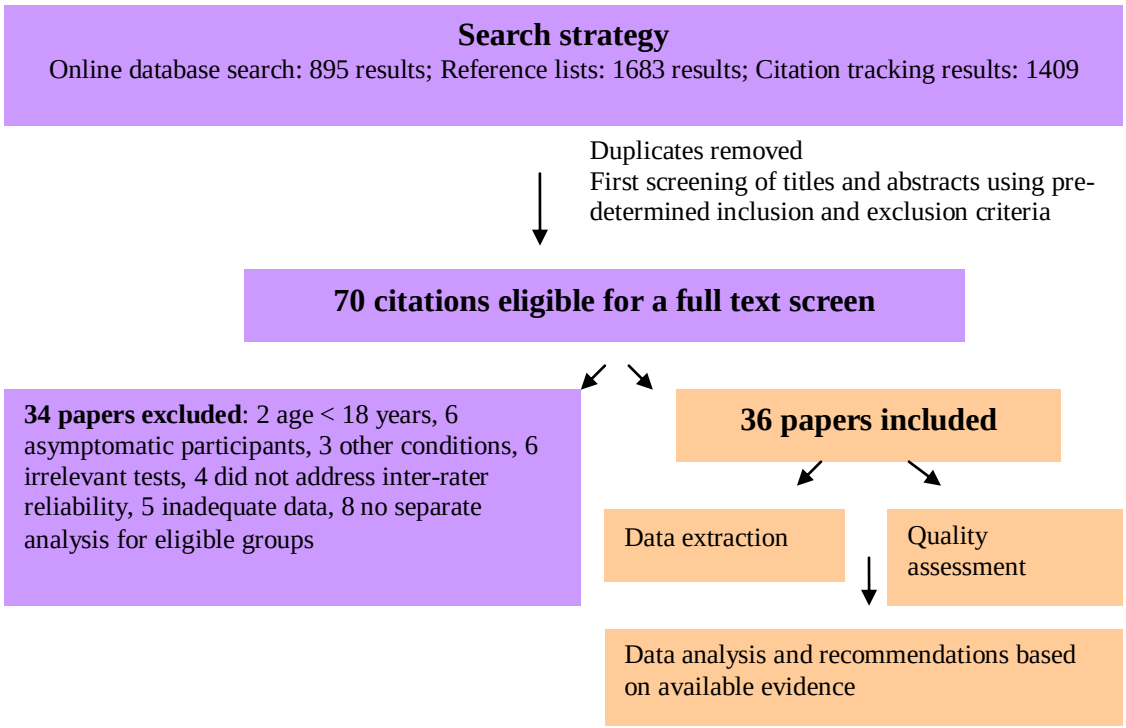
In view of the lack of consensus on thresholds for study quality and clinically acceptable reliability, it was decided to conduct a sensitivity analysis. Sensitivity analysis normally assesses the robustness of the review conclusions (de Vet et al., 2003b; Egger et al., 2001), but in this chapter, it was also used to explore the impact of different, previously used, quality and reliability thresholds on these conclusions. Clinically acceptable reliability and methodological quality were initially set at 0.7 and 60%, respectively. The pre-specified cut-off points for adequate methodological quality (60%) and minimally acceptable levels of reliability (0.7) were increased to a margin of 10% for methodological quality and decreased to 0.4 for reliability following commonly used thresholds in systematic reviews of reliability of physical tests.

2.4 RESULTS

2.4.1 Description of studies

A flowchart of the results of the search strategy and the review procedures is presented in Figure 2.4. The overall search strategy yielded 3987 results. After removing duplicates and screening abstracts, 70 citations were eligible for a full-text screen.

Figure 2.4 Flowchart of review procedures and results.



From these, 34 citations were excluded (Appendix 2.2), leaving 36 articles in the review. Six studies involved patients with NP and thirty studies recruited patients with LBP, with sample sizes of 12 to 127 individuals. No studies were found including patients with pain in other spinal areas. The mean age of samples of participating subjects ranged from 36 to 48 years. Investigated responses mainly involved symptom reproduction or changes in the intensity and / or location of symptoms in response to gross and segmental movement testing, palpation, non-

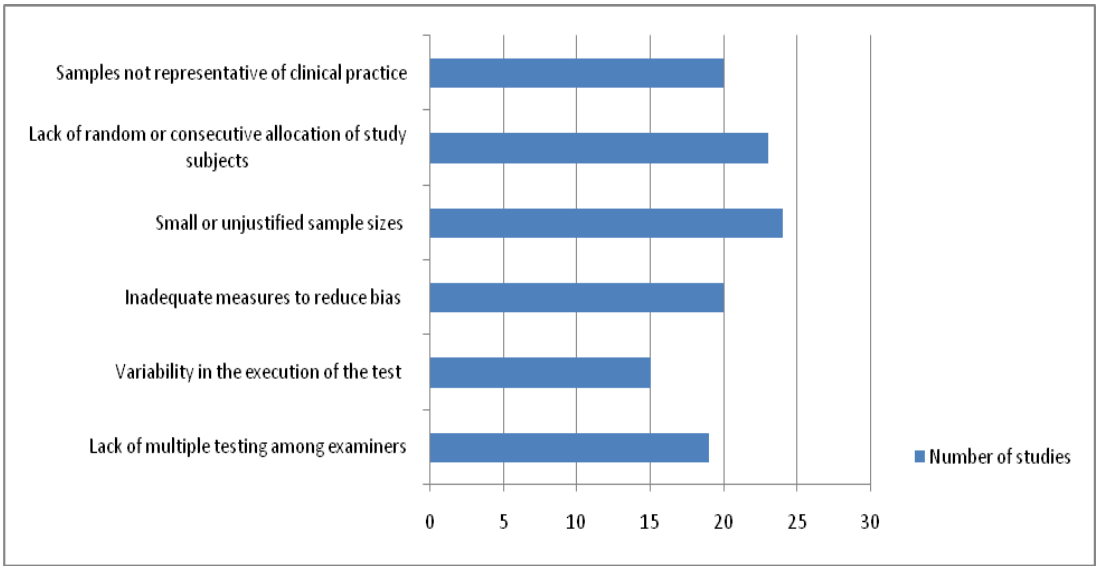
organic signs, and neural testing. Reliability studies were conducted using patients with one rater observing and one assessing (e.g. Razmjou et al., 2000), two adjacent examinations (e.g. Kilpikoski et al., 2002) or using videotaped examinations (e.g. Fritz et al., 2000a; Dionne et al., 2006). Most studies investigated a binary or ordinal scaled outcome and used kappa (Cohen, 1960) or weighted kappa (Cohen, 1968). For continuous variables, the ICC (Shrout and Fleiss, 1979) was most commonly selected for analysis. Five studies used percentage agreement or other statistics to measure reliability. Measures of precision, usually in the form of standard error or confidence intervals (95%CI) were used in some studies.

The characteristics of included studies are summarised in Table 2.9. The overall mean quality score of the studies was 60 / 100 (range 19-95 / 100), with approximately 56% of the studies scoring 60 or more. There was initially disagreement on 21 out of the 504 rated items (Appendix 2.3). All disagreements were resolved by discussion, and there was no need for the opinion of the third reviewer. The quality score (high or low quality) was associated with the year (before or after 2000) of publication [χ^2 (1, N=36) = 7.2, p <0.05]. Figure 2.5 presents the most commonly identified methodological weaknesses of included studies.

Table 2.9 Summary characteristics of included inter-rater reliability studies.

Characteristic	Number	Percentage (%)
Article publication date		
1980 - 1989	5	14
1990 - 1999	13	36
2000 - 2007	18	50
Study design		
Repeated measures	30	83
Parallel measures	5	14
Combination	1	2
Spinal symptoms		
Neck pain	6	17
Low back pain	30	83
Sample size of patients studied		
< 25	9	25
25 - 50	16	44
51 - 75	6	17
> 75	5	14
Sample size of examiners		
< 3	10	28
3 - 5	15	42
> 5	9	25
Not reported	2	5
Examiner background		
Physical therapist (PT), practitioner and / or student	16	44
Chiropractor, practitioner and / or student	4	11
Medical Doctor (MD)	6	17
Combination	4	11
Not reported	6	17
No of studies using different types of symptom responses		
Symptom response to gross movement testing	17	47
Symptom response to segmental testing	6	17
Symptom response to static testing	1	3
Symptom response to palpation / tenderness	13	36
Trigger point assessment	4	11
Symptom response to neural testing	7	19
Sacro-iliac joint pain provocation tests	4	11
Instability tests	2	6
Symptom response to a combination of strategies	2	6
Other	2	6
Non-organic signs	6	17
Symptom response to gross movement testing	5	14
Symptom response to palpation / tenderness	6	17
Other	3	8
Measure of statistics predominantly used		
Kappa	29	81
Intraclass correlation coefficient	2	6
Percent agreement	4	11
Other correlation statistics	1	3
Methodological quality		
≥ 0.60	20	56
< 0.60	16	45
Mean quality scores / category		
Study population		
Study population adequately described	2.6	65
Representative of clinical practice	1.8	45
Random or consecutive selection	2.5	36
Number of subjects	4.1	41
Study conditions		
Clearly described and reproducible procedure	3.3	66
Uniform execution	2.9	58
Adequate measures to reduce bias	4.4	44
Level of examiners	6.7	67
Consensus/ Training prior to testing	3.1	61
Study results		
More than one pair of examiners	6.1	61
Multiple testing between examiners	2.3	46
Standardised measure of outcome	4.2	85
Frequencies and agreement reported	7.9	79
Appropriate statistics	7.6	76

Figure 2.5 Most commonly identified methodological weaknesses of included studies.



2.4.2 Final approach to synthesis

The strong methodological and clinical heterogeneity across studies did not allow a meta-analysis to be undertaken. Instead, a qualitative synthesis of the evidence was performed, taking into account the number, consistency and validity of the study results. The kappa statistic (unweighted / weighted) was considered appropriate for the analysis of nominal / ordinal data (Thompson and Walter, 1988; Altman, 1991; Bartfay and Donner, 2001; Ludbrook, 2002; Sim and Wright, 2005; Viera and Garrett, 2005). The Bland-Altman method and the ICC (Shrout and Fleiss, 1979) were acceptable statistical techniques for continuous data. Other approaches, discussed previously, were considered inadequate when used exclusively in an included paper (Armstrong, 1981; Altman and Bland, 1983; Ludbrook, 2002). Studies using such approaches have been listed in this review (Appendix 2.5), but excluded from the analysis.

Levels of evidence (LOE) for included studies, adapted from van Tulder and associates (2003) and May and associates (2006), were assigned to the study results (Table 2.10). Like other reviews in the field (May et al., 2006; Stochkendahl et al., 2006; Myburgh et al., 2008), a moderate LOE strategy was followed (Ferreira et al., 2002).

Table 2.10 Levels of evidence for reliability studies (van Tulder et al., 2003; May et al., 2006).

Strong evidence	Consistent findings from multiple high quality studies
Moderate evidence	Consistent findings among low quality studies and/or one high quality study
Limited evidence	One low quality study
Conflicting evidence	Inconsistent findings among multiple studies
No evidence	No studies

2.4.3 Clinically induced symptom responses and inter-rater reliability

Results are presented below by spinal area and reliability threshold.

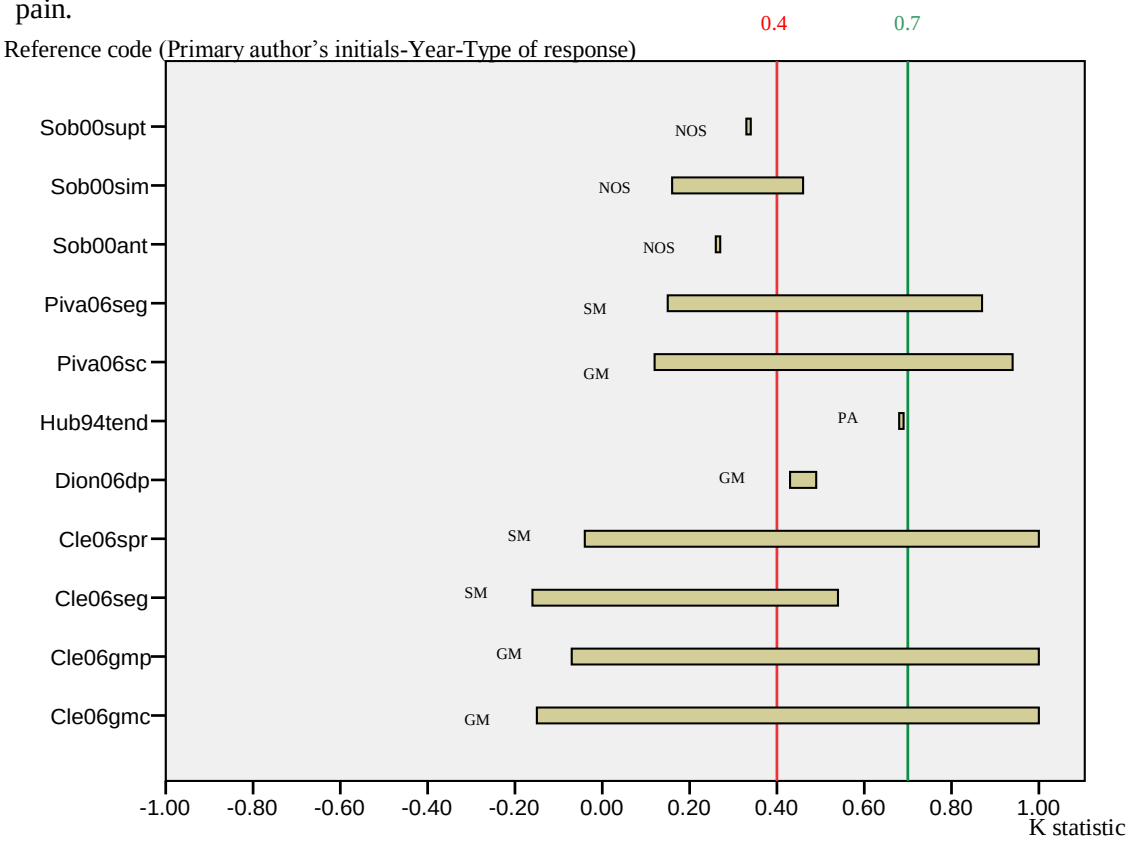
2.4.3.1 Reliability of clinically induced symptom responses in neck pain

A table of results for clinically induced symptom response procedures in NP is presented in Appendix 2.4. The study results across different thresholds are graphically displayed in Figure 2.6.

Four papers (Pool et al., 2004; Cleland et al., 2006a; Dionne et al., 2006; Piva et al., 2006) reported on symptom response to gross movement (GM) testing, three papers (Pool et al., 2004; Cleland et al., 2006a; Piva et al., 2006) on segmental movement (SM), one paper (Hubka and Phelan, 1994) on palpation (PA) and one paper (Sobel et al., 2000) on non-organic signs (NOS). When the ≥ 0.7 threshold was applied,

there was evidence that: changes in symptom location and / or intensity in response to single movements are reliable (strong evidence); spring testing is reliable (moderate evidence); judgements on directional preference, and non-organics signs are unreliable (moderate evidence); neck tenderness is unreliable (limited evidence). Evidence on pain response to segmental mobility testing was conflicting. When the ≥ 0.4 threshold was applied, judgements on directional preference (moderate evidence) and neck tenderness (limited evidence) were also reliable; and evidence on pain response to segmental mobility testing changed from conflicting to strong evidence of acceptable reliability. Changes in the quality assessment threshold did not affect the evidence on clinically induced symptom responses in neck pain.

Figure 2.6 Inter-rater reliability results of clinically induced symptom responses in neck pain.



Note: Ant, Non-anatomic tenderness; Dp, Directional preference; GM, Gross movement; GMC, Centralisation / Peripheralisation with single range of motion; GMP, Pain with single range of motion; PA, Palpation; SM, Segmental movement; NOS, Non-organic signs; Sc, Status change with single cervical range of motion; Seg, Pain provocation on segmental testing; Sim, Simulation; Spr, Spring testing; Supt, Superficial tenderness; Tend, Tenderness. Figure 2.9 illustrates studies with κ statistics.

2.4.3.2 Reliability of clinically induced symptom responses in back pain

The results for clinically induced symptom response procedures in LBP are presented in a table in Appendix 2.5. The study results across different thresholds are graphically displayed in Figure 2.7.

Eleven papers (McCombe et al., 1989; van Dillen et al., 1998; Kilby et al., 1990; Strender et al., 1997b; Razmjou et al., 2000; Kilpikoski et al., 2002; Seymour et al., 2002; White and Thomas, 2002; Hicks et al., 2003; Haswell et al., 2004; Fritz et al., 2006) reported adequate statistics on symptom responses to gross movement (GM) testing, and eleven papers (McCombe et al., 1989; Keating et al., 1990; Nice et al., 1992; Waddell et al., 1992; Boline et al., 1993; Njoo et al., 1994; Maher and Adams, 1994; Strender et al., 1997b; Hsieh et al., 2000; Fritz and Piva, 2003; Petersen et al., 2004) investigated symptom responses to palpation (PA), six papers (McCombe et al., 1989; Waddell et al., 1992; van den Hoogen et al., 1996; Strender et al., 1997b; Vroomen et al., 2000; Petersen et al., 2004) reported results on neural testing (NT), three papers (Boline et al., 1988; Strender et al., 1997b; Hicks et al., 2003;) on segmental movement (SM), seven papers on special testing (SP) (McCombe et al., 1989; Waddell et al., 1992; Laslett et al., 1994; Strender et al., 1997b; Vroomen et al., 2000; Hicks et al., 2003; Fritz et al., 2006) one paper (Fritz et al., 2006) on static testing (ST), one on non-organic signs (McCombe et al., 1989), and two papers on a combination of spinal strategies (COM) (Fritz et al., 2000a; Petersen et al., 2004).

When the $\geq$ 0.7 threshold was applied, there was evidence that the following symptom responses to physical examination procedures were reliable: changes in symptoms in response to single movements, or to a combination of strategies

(moderate evidence); pain aggravation / reproduction or classification based on repeated lumbar movements (moderate evidence); localised tenderness (moderate evidence); pain reproduction on neural tension testing (moderate evidence); straight leg raise (SLR) crossed sign (moderate evidence); posterior shear and pelvic torsion tests (limited to moderate evidence); prone instability test (strong evidence); pain on resisted hip flexion (limited evidence).

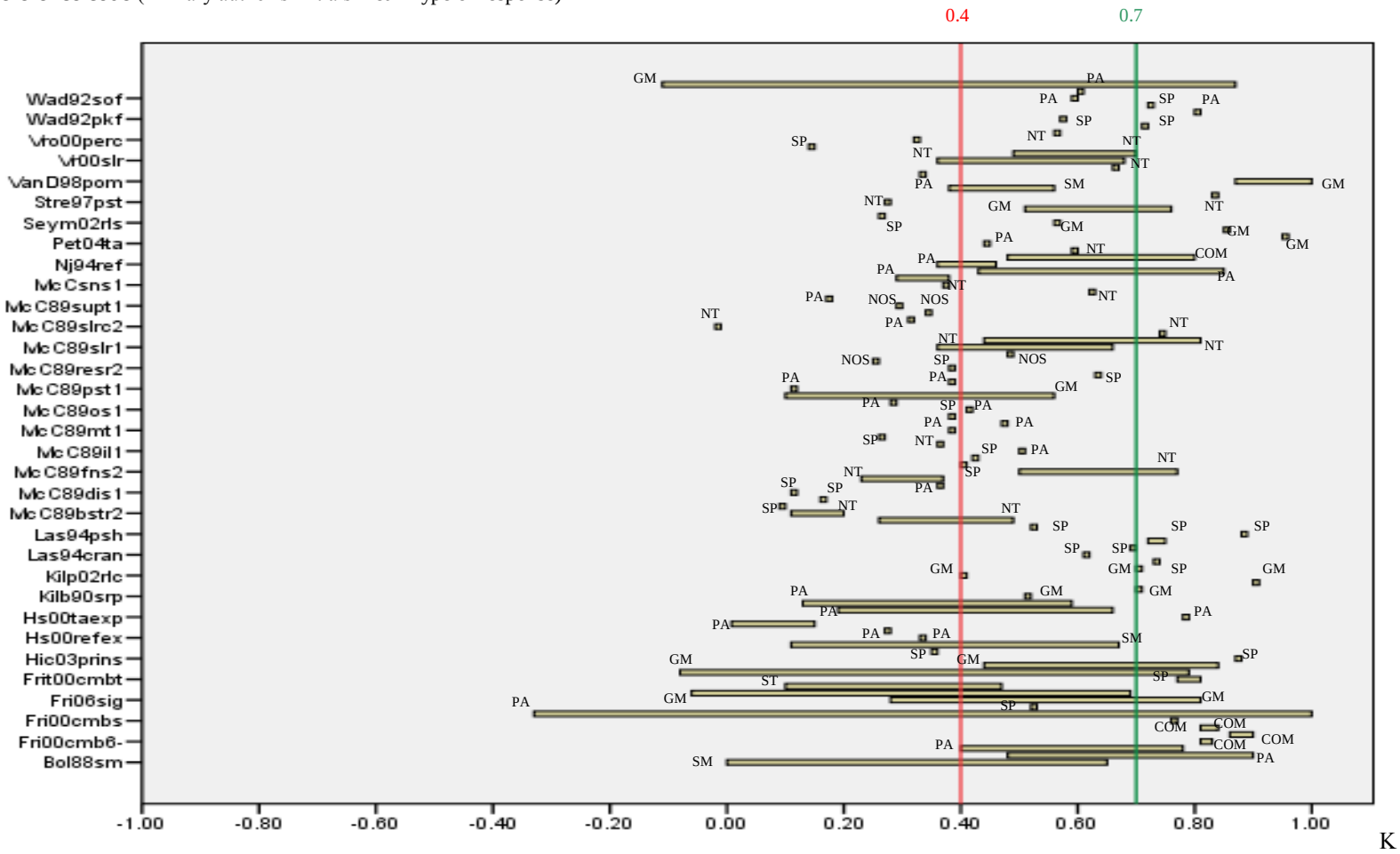
The following did not reach the 0.7 threshold of reliability: pain response to segmental mobility testing (strong evidence); changes in symptoms in response to sustained extension (moderate evidence); trigger point assessment (moderate evidence); SLR sciatic stretch test (limited evidence); Bragard sign (moderate evidence); Valleix pressure points (moderate evidence); sacro-iliac joint distraction test (moderate evidence); sacral thrust and cranial shear tests (limited to moderate evidence); Maitland sacro-iliac joint test (limited evidence); pain on resisted external hip rotation (limited evidence); posterior shear test (moderate evidence); non-organic signs (limited evidence); pain on hip abduction (limited evidence); pain on vertebral percussion (moderate evidence).

Conflicting evidence was found for the following responses: pain during lumbar movement; changes in symptoms in response to repeated movements; relevance of lateral shift; relevance of lateral component; spinal and paraspinal tenderness; soft tissue pain; taut band; referred pain pattern; pain on straight leg raising, knee flexion (femoral nerve stretch), hip flexion; sacro-iliac joint compression.

When the ≥ 0.4 threshold was applied, the following clinically induced symptom responses were also reliable: pain on movement (strong evidence); status change with repeated movement testing (strong evidence), or sustained extension (moderate evidence); relevance of lateral shift and lateral component (strong evidence); pain provocation on segmental movement testing (strong evidence); spinal tenderness (moderate evidence); soft tissue and osseous pain (moderate evidence); SLR sciatic stretch test (limited evidence); Bragard sign (moderate evidence); pain on Bowstring testing (limited evidence); femoral nerve stretch (moderate evidence); sacral thrust (limited to moderate evidence); cranial shear test (limited to moderate evidence); pain on hip flexion (moderate evidence); pain on hip abduction (limited evidence); pain on resisted external hip rotation (limited evidence); simulation (limited evidence); posterior shear test (moderate evidence). The evidence on sacro-iliac joint distraction changed from moderate to conflicting.

Changes in the quality assessment threshold (0.6 to 0.5) changed the evidence from limited to moderate on the following assessment procedures: sacro-iliac joint posterior shear (or thigh thrust test); pelvic torsion; sacral thrust; cranial shear test; status change with lumbar movement or sustained postures.

Figure 2.7 Inter-rater reliability results of clinically induced symptom responses in back pain.



Note: CMBT, Status change on combination of startegies; COM, Combination of strategies; CRAN, Cranial shear test; Dis, Distraction; FNS, Femoral nerve stretch; GM, Gross movement; IL, Iliac crest tenderness; MT, Maitland testing; NOS, Non-organic signs; NT, Neural testing.Os, Osseous pain; PA, Palpation; PERC, Pain on vertebral percussion; PKF, Pain on knee flexion; POM, Pain on lumbar movement; PRINS, Prone instability test; PSH, Posterior shear test; PST, Pain on segmental testing; Ref, Referred pain pattern; REFEX, Referred pain pattern (experts); RLS, Relevance of lateral shift; RESR, Pain on resisted rotation; RLC, Relevance of lateral component; SLR, Pain on straight leg raise; SLRC, straight leg raise crossed; SM, Segmental movement; SOF, Soft tissue tenderness; SP, Special testing; SRP, Symprom response on repeated movement; ST, Static testing; SUPT, Superficial tenderness; TA, Taut band; TAEXP,Taut band (experts) . Figure 2.7 illustrates studies with κ statistics.

2.4.3.3 Reliability of clinically induced symptom responses and training / experience of raters

Studies with results between groups with different types / levels of training or experience are presented in Appendix 2.6. Evidence was found in three LBP studies (McCombe et al., 1989; Fritz et al., 2000a; Hsieh et al., 2000).

2.4.3.3.1 Type of training

When applying the high threshold (≥ 0.7) of acceptable reliability, there were no differences in reliability for type of training (two orthopaedic surgeons versus orthopaedic surgeon with physiotherapist) for: pain on lumbar movement (limited evidence); midline tenderness (limited evidence); paraspinal tenderness (limited evidence); buttock tenderness (limited evidence); sacroiliac tenderness (limited evidence); iliac crest tenderness (limited evidence); SLR sciatic stretch test (limited evidence); pain on Bowstring testing (limited evidence); sacro-iliac joint compression, distraction, or the Maitland sacroiliac test (limited evidence); pain on hip flexion, or resisted external rotation (limited evidence); non-organic signs (superficial tenderness or simulation) (limited evidence). Differences in reliability for type of training (two orthopaedic surgeons versus orthopaedic surgeon with physiotherapist) were found for: pain on SLR (limited evidence); crossed SLR (limited evidence); femoral nerve stretch (limited evidence). When the lower threshold (≥0.4) was applied, differences in reliability for type of training were also observed for: midline, sacroiliac and iliac crest tenderness (limited evidence); SLR sciatic stretch test (limited evidence); pain on Bowstring testing (limited evidence); pain on resisted external rotation (limited evidence); simulation (limited evidence), but not for pain on SLR (limited evidence) or crossed SLR (limited evidence).

2.4.3.3.2 Level of training

When applying the high threshold (≥ 0.7) of acceptable reliability, there were differences in reliability for level of training (expert with examiners versus examiners only) for taut band (limited to moderate evidence) but not for referred pain pattern (limited to moderate evidence). There were no differences in reliability (≥ 0.7 or ≥ 0.4) for level of training or experience in judgements of symptom changes in response to repeated and / or sustained lumbar movements (limited evidence).

2.4.3.3.3 Experience

There were no differences in reliability (≥ 0.7or ≥ 0.4) for experience when investigating judgements of symptom changes in response to repeated and/or sustained lumbar movements (limited evidence).

2.4.4 Results of the sensitivity analysis

Changes in the threshold of acceptable reliability affected conclusions regarding clinically induced symptom responses to the physical examination of both the lumbar spine, sacroiliac joints (21 / 99) and the cervical spine (3 / 11). In particular, evidence changed from: conflicting to strong in 4 lumbar procedures; conflicting to moderate in 5 lumbar and 1 SIJ procedures; moderate to conflicting in 1 SIJ procedure; evidence of unacceptable to acceptable reliability in 7 lumbar and 4 SIJ procedures. Lowering the threshold for adequate methodological quality from 60% to 50% shifted levels of evidence from limited to moderate in 4 SIJ procedures and 1 lumbar procedure.

2.5 DISCUSSION

This is the most comprehensive systematic review on the inter-rater reliability of clinically induced symptom responses in the physical examination of the spine to date. Information originating from different professions was sought and found. Thus, this review may partly contribute to resolving uncertainty over inter-clinician consistency on such procedures across a range of professions. A systematic and reasoned approach was followed in the methodology, from the point of identifying relevant studies to appraising and analysing findings. In contrast to previous reliability reviews, the impact of commonly applied thresholds of acceptable reliability and methodological quality was explored. We found that the selection of reliability thresholds and adequate study quality can influence the review conclusions, demonstrated in the shift of observed trends when different thresholds were applied in the synthesis of the study findings. This finding may facilitate comparison of conclusions with previous reviews and highlight the need for uniform and well-informed decisions when summarising and systematically appraising the literature.

The primary findings of this chapter indicate that research on the inter-rater reliability of clinically induced symptom responses needs to be improved. Samples not representative of clinical practice, lack of random or consecutive allocation of study subjects, small or unjustified sample sizes, inadequate measures to reduce bias, variability in the execution of the test and lack of multiple testing among examiners were the most commonly identified study weaknesses. These methodological weaknesses should be addressed in future research. Current standards and expectations of scientific rigour might not have been expected at the time some older

studies were conducted and published (Seffinger et al., 2004). Nevertheless, steps towards improving methodological quality have already been made in the more recent reliability studies; this is evident in the three times lower number of low quality studies in more recent publications.

Few clinically induced symptom responses passed the high threshold of reliability frequently advocated for use of such procedures in clinical practice. However, for many tests, the evidence was preliminary and based on a single study. Fewer studies were conducted, and significantly less clinically induced symptom procedures were investigated in areas other than the lumbar spine. The comparability of findings among studies was often difficult due to study variability. With recorded measurements being the product of several factors, this variability was not attributable only to the clinically induced responses, but also the raters, the setting, the training and the protocol (de Vet et al., 2003a; Van Genderen et al., 2003). The variability in methods used by the source studies contributed to conflicting evidence. Several systematic reviews on the reliability of physical examination procedures have highlighted this problem (Seffinger et al., 2004; May et al., 2006; van Trijffel et al., 2005). Every effort was made in this investigation to present and analyse results considering the potential impact of various study characteristics and thresholds.

Differences in reliability between different professionals were observed in neural and SIJ tests. Laslett (1997) argues that this may be partly explained by differences in technique of examination among different professions. Thus, training in examination procedures is advocated by some for consistent findings (Aina et al., 2004).

Some informative trends also emerged from this review. In studies that used the kappa statistics, the proportion of studies showing reliability over 0.7 was higher for pain with gross movement testing compared to pain on segmental mobility testing. Regional range of motion has been found to be more reliable than segmental motion assessments (Seffinger et al., 2004), possibly because of the differences in the magnitude of applied pressure but also the difficulty in accurately locating and naming the spinal level (Hicks et al., 2003). Pain on palpation was not satisfactory for most spinal levels and had the lowest percentage of studies with reliable findings. This is in agreement with other systematic reviews questioning the reliability (Seffinger et al., 2004; May et al., 2006) and validity (Najm et al., 2003) of palpatory assessments.

The online database search was supplemented by manual searches and citation tracking to locate eligible articles. Similar to studies of diagnostic accuracy (Devillé et al., 2002), reliability studies are poorly indexed in databases possibly because of inconsistent terminology in reliability research (van Trijffel et al., 2005). Every effort was made to find all relevant studies, but some eligible studies may have been missed. Selection or language bias may have occurred because only English language articles were included. However, none of the screened foreign language studies seemed to be relevant to this systematic review.

Variability among examiners' ratings affects diagnostic test accuracy (Whiting et al., 2004). Clear and consistent measurements of potential prognostic and treatment indicators are required in research and clinical practice (Simon and Altman, 1994; Fritz and Wainner, 2001; Beattie and Nelson, 2007). However, the margin of error

that is acceptable within the context of intended use is currently unknown (Andersson and Granberg, 1997; Wainner, 2003) making numerical values often misleading. Most included studies provided summary statistics within the context of independent preliminary investigations, but none explicitly explored reasons for poor reliability values or proposed strategies for improving reliability. This is particularly important because it may lead to the premature exclusion of useful or the promotion of highly reliable, but clinically meaningless tests (Fritz and Wainner, 2001).

2.6 SUMMARY

This systematic review identified 36 studies that evaluated the reliability of clinically induced symptom responses in the physical examination of spinal pain. The findings of this review have implications for research and clinical practice. Clinicians need to be cognisant that many examination procedures commonly used in spine assessment either lack or demonstrate inconsistent reliability. However, more research is warranted before these symptom responses are abandoned, especially in areas other than the lumbar spine. Research on the inter-rater reliability of symptom responses should be improved. Attempts to determine and deal with the source of error, or making judgements by taking account of the current uncertainty, are essential in order to improve future use of these procedures. Finally, findings should be combined with data on the estimation of or contribution to the prediction of the future course of spinal pain in order for these tests to have further value or utility.

.

Chapter 3

The prognostic value of clinically induced symptom responses in the conservative management of spinal pain: a systematic review.

3.1 AIMS OF CHAPTER

This chapter presents a systematic review of prognostic studies on clinically induced symptom responses in spinal pain. The definition of clinically induced symptom responses is presented in Chapter 1. The aim was to provide a comprehensive review of the quality of research on and investigate the prognostic value of these procedures in the conservative management of spinal pain.

3.2 INTRODUCTION

Reviews on the reliability of physical examination suggest that symptom-related tests show more consistent results than other assessment procedures (Hestbaek and Leboeuf-Yde, 2000; Seffinger et al., 2004; May et al., 2006). In Chapter 2, a systematic review of reliability studies on clinically induced symptom responses concluded that some symptom responses demonstrate acceptable reliability levels in the spinal physical examination. However, the value of a test can not rely exclusively on reliability (Fritz and Wainner, 2001). The estimation of or contribution to the prediction of the future course of spinal pain adds value to the utility of a test.

The prognostic value of clinically induced symptom responses has been summarised for a few procedures, mostly in low back pan (LBP). Most reviews have focused on symptom responses to movement testing, or treatment (Wetzel and Donelson, 2003; Aina et al., 2004; Wessels et al., 2006; Berthelot et al., 2007). Changes in symptoms in response to repeated end-range movement (Wetzel and Donelson, 2003; Aina et al., 2004; Berthelot et al., 2007), or as a result of treatment strategies (Wessels et al., 2006) show promise in the prognosis of LBP outcomes. However, these reviews are

narrative or critical reviews of the literature using a descriptive approach to summarising findings. There have been no attempts to formally synthesise the data and no reasons given for this decision. This decreases confidence in results and conclusions made from these investigations (CRD, 2008).

The literature has expanded considerably since the last published search strategy in 2004, with articles reporting new information on some clinically induced symptom responses. The type of responses described in the literature is also much wider than what has been addressed previously. Some responses with prognostic utility may be common across different sites of spinal pain (Carnes and Underwood, 2007; Mallen et al., 2007). However, it is currently unknown which clinically induced responses are prognostically important. A comprehensive and systematic investigation of the literature on all clinically induced symptom responses across all spinal areas is still missing.

3.3 MATERIALS AND METHODS

Methods in this chapter have been based on recommendations on the conduct of reviews of prognosis (Laupacis et al., 1994; Stroup et al., 2000; Altman, 2001) as well as suggestions on ways of summarising and appraising the scientific literature (Egger et al., 2001; Deeks et al., 2003; Hayden et al., 2006). Reporting was made in accordance to current guidelines and suggestions (Moher et al., 1999; Stroup et al., 2000; McShane et al., 2005; Sampson et al., 2008).

3.3.1 Selection criteria

The selection of studies was made according to the criteria listed in Table 3.1. The eligibility criteria were defined in terms of the population, type of prognostic variables, interventions, and appropriate study designs (Counsell, 1997). The focus of this chapter was on the whole range of clinically induced symptom responses rather than particular types of responses investigated previously. Observational and experimental study designs offering the least biased answer were selected (Counsell, 1997; Altman, 2001, CRD, 2008; Hayden et al., 2008). No restriction was placed on the range of patient reported outcomes or the timing of the data collection. This decision was made because one of the chapters' objectives was to assess the quality of research, a component of which included the outcomes selected by studies.

Table 3.1 Eligibility criteria for the selection of prognostic studies.

Inclusion criteria
1. Primary research 2. Prospective longitudinal design (randomised controlled trial or cohort study) involving a cohort of patients 3. Current episode of spinal pain with or without radiating symptoms 4. Adults (≥ 18 years) 5. Investigation of at least one clinically induced symptom response 6. Conservative spine care 7. Patient reported outcomes 8. English language
Exclusion criteria
1. Inclusion criteria are not met[6] 2. Recruitment of participants on the basis of specific patho-anatomical and / or other confirmed serious pathologic conditions 3. Investigation of pain behaviours[7] 4. Full text unavailable

3.3.2 Search strategy and selection of studies

Different databases were searched to increase the coverage of journals (Counsell, 1997). Ovid-MEDLINE, Ovid-EMBASE, Ovid-CINAHL and Ovid-AMED were searched from inception up to March 2007. These databases were selected because of their relevance to the nature of the review question (CRD, 2008). Appendix 3.1 presents the search strategy for online databases. The search strategy was developed

[6] This also refers to studies where it was impossible to determine the nature of the investigated prognostic variables or no separate analysis was reported.
[7] For a definition of pain behaviours, please refer to Chapter 2.

by the author of this thesis (Angeliki Chorti) who is familiar with the topic area and systematic reviews methodology. This strategy involved combinations of keywords and Medical Subject Headings (MeSH) terms (Appendix 3.1). These combinations were the product of the research question being broken down into separate components (CRD, 2008). The combination of words and / or terms instead of the use of single terms was chosen to optimise the results of searches for prognostic studies (Haynes et al., 1994; Wilczynski et al., 2004; Wilczynski and Haynes, 2005). All databases were searched using a basic search template. Adjustments were made to individual database searches to maximise the relevant citation yield.

Online database searches are the most commonly used strategy for identifying articles in systematic reviews (Counsell, 1997). However, efforts to identify all relevant articles through the online databases search are usually inadequate because of indexing problems. Indexing in online databases has evolved and improved over time, but this change is quite slow (Wilczynski et al., 2002). Supplementing searches with manual screening of reference lists and citation tracking was necessary. The references of relevant systematic reviews identified by the searches and reference lists of included articles were searched. Forward citation tracking was performed for all included articles using the Web of Science (Bakkalbasi et al., 2006).

The titles and if available, abstracts of the retrieved citations were initially scanned by the author of this thesis to determine whether they were potentially relevant. The full text of eligible citations or citations where the eligibility could not be determined was then independently assessed by two reviewers (Angeliki Chorti and Anastasios

Chortis) who decided whether a study should be included or not. Any disagreements between reviewers were resolved through discussion. The reviewers were not blind to study details such as authorship, institution affiliations and journal name because of their familiarity with the literature in the field (Berlin et al., 1997).

3.3.3 Quality assessment

The quality of included studies influences the validity of results and conclusions made in a review (CRD, 2008). Therefore, an assessment tool, adapted from Hudak and associates (1996) and presented in Table 3.2, was used. This tool derived from theoretical methodological principles (Kernan et al., 1991; Sackett et al., 1991; Fletcher et al., 1996) and was originally used for the evaluation of studies on the prognosis of work-related neck and upper extremity disorders (Cole and Hudak, 1996). Because there was scarce or no empirical evidence to guide the selection of most methodological aspects likely to be important in the quality assessment of prognostic studies, theoretical considerations and reviewing quality assessment recommendations were used to meet minimal requirements of face and content validity[8] (Hayden et al., 2006; CRD, 2008).

There are no widely agreed quality criteria for assessing prognostic studies (Altman, 2001; Hayden et al., 2006; CRD, 2008). However, issues around quality assessment of prognostic studies involve the same domains: sample definition and selection, measurement of prognostic factors and outcomes, intervention details, follow-up of

[8] For a definition of these terms, please refer to Chapter 1.

patients, and statistical analysis (Laupacis et al., 1994; Hudak et al., 1996; Altman, 2001; Hayden et al., 2006; CRD, 2008). Most of the above criteria have been adopted in subsequent reviews on symptom responses (Aina et al., 2004) as well as systematic reviews of prognosis in musculoskeletal pain (Mallen et al., 2007; Kent and Keating, 2008). Sample size is also important when evaluating prognostic studies but has received little attention in lists of proposed quality criteria (Hayden et al., 2006; CRD, 2008). Therefore, instead of excluding sample size in the study quality assessment criteria, this issue was considered through sensitivity analyses (CRD, 2008).

Two raters (Anastasios Chortis and Nikolaos Strimpakos) independently assessed the quality of each included study. The raters were not blind to the details of the studies. Disagreements were resolved through discussion / consensus and if disagreement persisted, the opinion of the third rater (Angeliki Chorti) was sought. The maximum score that a study could achieve was 20. The cut-off point distinguishing a high from a low quality study was set to 50% of the maximum score or more (10 points). This cut-off point reflected adequate quality in reviews of whiplash patients (Williams et al., 2007; Williamson et al., 2007).

Table 3.2 Quality assessment tool for prognostic studies (adapted from Hudak et al., 1996, with permission).

Case definition	
Operational definition of cases including exclusion criteria	2
Operational definition of cases but no exclusion criteria	1
No explicit definition of cases	0
Source population	
Clear description of the source population	1
Unclear or no description of the source population	0
Representativeness	
Patients representative of clinical practice	2
Patients unlikely to be representative of clinical practice	1
Unable to determine	0
Patient Selection	
Inception cohort (defined in relationship to onset of symptoms)	2
Survival cohort, including a subset of the sample with an acute episode (which is analysed separately)	1
Survival cohort; unable to define subsets within the cohort or unclear	0
Participants	
Clinical and demographic characteristics described	2
Insufficient description of participants characteristics	1
No explicit description of participants characteristics	0
Treatment	
Description and standardisation and / or randomisation of provided treatment	2
Description of treatment but no standardisation or randomisation	1
No information on the treatment provided	0
Follow-up (extent and length)	
Follow-up of ≥ 80% of total sample to at least 1 year	3
Follow-up of ≥ 80% of total sample for less than 1 year or patients followed for varying lengths of time, including 1 year	2
Follow-up < 80% of total sample	1
Unclear	0
Outcome	
Blinded outcome criteria appropriate to the research question with reports of standardised or valid measurements	2
Outcome criteria appropriate to the research question	1
No explicit outcome criteria (e.g. patient significantly improved)	0
Prognostic factors	
Adequate description of potential prognostic factor(s) including information on standardised or validated measurements instruments	2
Adequate description of potential prognostic factors but insufficient detail on standardisation or validation	1
Inadequate description of potential prognostic factors	0
Analysis	
Adjusted proportions provided or appropriate multivariate techniques used to adjust for other prognostic factors	2
Crude proportions but data stratified or presented in a manner which would allow for analysis of subsets	1
Crude proportions for at least one outcome	0

3.3.4 Data extraction

There is no established form for data extraction of prognostic studies but general recommendations do exist (CRD, 2008). A search of common data extraction domains of interest to prognostic studies ensured that relevant information was not omitted from the final synthesis. Table 3.3 presents the types of information that was extracted from included prognostic studies. Data extraction was performed by the author of this thesis.

Table 3.3 Types of extracted information from included prognostic studies.

Data extraction on:

- **General information** (authors, title, citation, type of publication, country of origin, and source of funding)
- **Study characteristics** (aims / objectives of study, study design, eligibility criteria, recruitment procedures)
- **Participant information** (sampling strategy, numbers and characteristics)
- **Intervention details** (setting, description)
- **Predictor variables** (type, definition, measurement method)
- **Outcome data / results** (outcome measures, follow-up details, cut-points and rationale, statistical methods used, missing data, study findings).

3.3.5 Investigation of heterogeneity

In the context of this prognostic review, heterogeneity refers to the variability among studies and is divided into clinical, methodological and statistical diversity (Deeks et al., 2008). Clinical diversity may arise from differences in the patient population, prognostic variables, interventions and outcomes between studies, whereas methodological heterogeneity relates to study design and methodological variation (Deeks et al., 2008). Statistical heterogeneity, a consequence of clinical, methodological heterogeneity or both (Deeks et al., 2008), manifests itself in results from studies which may vary in the magnitude of effects or, the direction of effects (Dickersin and Berlin, 1992). Heterogeneity presents a problem for synthesis, as individual studies are likely to have different clinical and methodological characteristics (CRD, 2008). This may influence the decision to follow a meta-analytic (quantitative) approach and the use of a single summary measure (Dickersin and Berlin, 1992). Some have found ways to address such problems in meta-analyses of epidemiologic studies (Chene and Thompson, 1996; Shi and Copas, 2004; Hartemink et al., 2006; Key et al., 2006), but these approaches are rarely applicable in prognostic studies of spinal outcomes. Others have suggested the use of individual patient data to overcome such issues (Riley et al., 2003), but this approach is often characterised by failure to collect all relevant data (CRD, 2008).

Clinical and methodological sources of variation in included studies were considered first. A visual inspection of the study characteristics and findings and a comparison of these in studies investigating the same type of clinically induced symptom response were performed. Formal statistical testing to explore sources of

heterogeneity within the same type of clinically induced symptom response could not be used due to the small number of studies on different types of symptom responses (maximum number was four) and the scarcity of reported data on the size of the association with outcomes (Dickersin and Berlin, 1992; Deeks et al., 2008).

3.3.6 Data synthesis considerations

Publications using the same cohort were considered as one study (CRD, 2008) except in the case where different papers referred to different cohorts of patients. The latter were analysed separately, following suggestions that the influence of some prognostic factors may vary across patients with different types of pain (van der Windt et al., 2007). Symptom responses were clustered into groups based on the characteristics reported in included papers; outcomes were also classified into categories depending on the domain they referred to (i.e. symptoms, range of motion, strength, disability, health status, healthcare use, work status and work loss, perceived global change, satisfaction) (Bombardier, 2000). This approach has been used before to identify responses that appear to be comparable, despite variation in labelling (Kent and Keating, 2008). If crude associations with outcome were presented, significant associations were defined as an unadjusted p value < 0.05. If prognostic indications had been included in a multivariate model, those with an adjusted p value < 0.05 were judged to be statistically significant. Results of the quality assessment were presented separately for each included study. The extracted data were presented separately in two tables, the first containing information on the

characteristics of included studies and the second presenting the results from each individual study.

Synthesis of prognostic studies is a relatively new and evolving area with less well developed methods than for reviews of therapeutic interventions or of diagnostic accuracy (CRD, 2008). However, meta-analysis is often recommended in prognostic studies (Dickersin and Berlin, 1992). It was decided *a priori* that a meta-analysis would be performed only if the comparability between study characteristics and findings allowed data to be combined, there was an adequate amount of data (≥ 2 studies) and study quality and reporting was satisfactory (Tierney et al., 2007; CRD, 2008).

3.3.7 Sensitivity analysis

Considerations of investigating the robustness of review findings apply equally to reviews of prognostic studies (CRD, 2008). The influence of study quality, and in the specific context of prognosis the effect of smaller studies on the review conclusions is frequently a consideration (Egger and Smith, 1998; CRD, 2008). Sensitivity analyses taking account of these factors were performed to assess the robustness of the chapter conclusions.

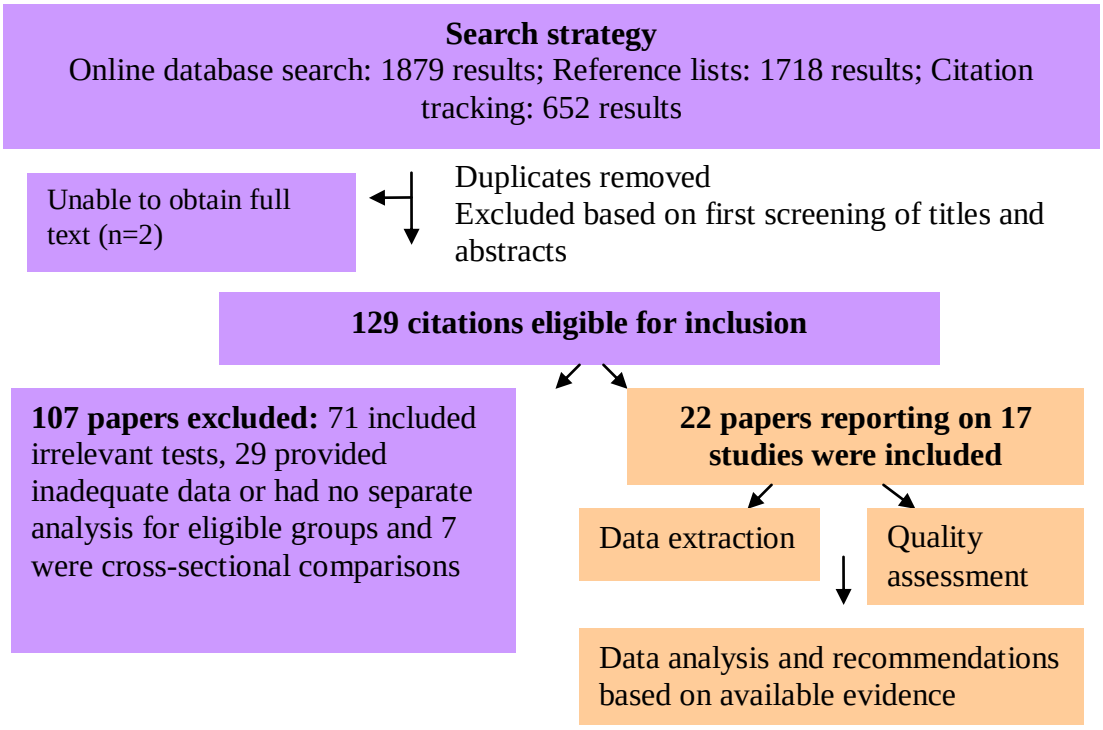
3.4 RESULTS

3.4.1 Description of studies

The search strategy yielded 4249 results. After the first screening, 131 citations appeared relevant and were eligible for a full-text review. Further information was required to determine eligibility in four studies reported in six citations (Burton and Tillotson, 1991; Burton et al., 1995; Seferlis et al., 2000; Hicks, 2002; Schultz et al., 2002; Hicks et al., 2005). Thus, contact with the authors of the studies was sought. The full text of two citations (Jordan, 1996; Hopwood et al., 1993) could not be found. One citation was a PhD thesis from the United States that was not available in the UK and the second citation was presented in a paper that could not be retrieved by the British Library Document Supply Centre. After unsuccessful efforts to contact the authors, these studies were excluded from the review. Appraisal of the full-text of the remaining articles resulted in another 107 articles being excluded (Appendix 3.2), leaving 22 articles for inclusion in the review (Figure 3.1).

The 22 citations reported data from 17 studies and 18 different cohorts of spinal patients. A summary of the characteristics of the included studies and resulting cohorts are presented in Appendix 3.3. Table 3.4 presents the statistical methods used in each study and the statistically significant results for all investigated prognostic variables in included studies.

Figure 3.1 Flowchart of review procedures and results.



The majority of studies were observational studies in LBP. Only one randomized controlled trial assessed prognostic factors using cohorts from the allocated treatment groups. Some studies focused on the association between clinically induced symptom responses and outcomes (i.e. explanatory studies, Table 3.5), others were outcome prediction studies i.e. aimed at identifying a combination of factors most strongly associated with outcome (Hayden et al., 2008).

A wide range of clinically induced symptom responses were candidate prognostic factors, including changes in symptoms with physical examination e.g. in response to gross or segmental movement testing, compression and distraction, neurodynamic, instability tests or symptom response to treatment sessions. A wide range of outcomes were also studied related to 10 domains (i.e. symptoms, range of motion,

strength, disability, health status, healthcare use, work status and loss, perceived global change and satisfaction). The most frequently used outcomes related to symptoms (11 studies), disability (10 studies) and work loss (4 studies). Duration of follow up ranged from 2 days to 4 years. Bivariate relationships between factors and outcomes were reported in included studies in a variety of methods e.g. differences between reported mean outcome scores between recovered and non-recovered groups, associations between the presence of a factor and outcome, and odds ratios. Multivariate relationships were also determined using a variety of methods e.g. linear or logistic regression analysis, Cox analysis, discriminant analysis, K-means cluster analysis, multinomial logit models, Kaplan – Meier survival curves.

3.4.2 Methodological quality of studies

The quality assessment results are presented for each study in Appendix 3.4. General comments related to limitations of the studies are also included (Appendix 3.4). Most studies were rated as high quality (overall mean score 14 / 20, range 6 - 18) and only one study was considered low quality. There was initially disagreement on 34 / 180 items among raters (Appendix 3.4). However, disagreement on individual study quality scores resulted only in one discrepancy in the judgement of high / low quality. The lack of an inception cohort was the most common methodological shortcoming. Weaknesses in terms of the representativeness of the population studied, outcome and follow-up assessments as well as inadequacies in the statistical analysis were also observed in included cohorts.

Table 3.4 Summary of the results for the included cohorts.

Reference	Type and measure of outcome	Follow-up	Statistical methods	Results (univariate)	Results (multivariate)
Neck pain					
Cleland (2007)	(1) Perceived global change (GROC)	After treatment, mean time 2.3 days to 6.3 days.	Independent samples t-tests (Co), cut-off determined by accuracy analyses ROC x^2 tests (Ca) ($p < 0.10$) Logistic regression ($p < 0.10$).	Symptom duration < 30 days [LR+ = 6.4 (1.60 - 26.3)], no symptoms distal to shoulder [LR+ = 1.4 (0.94 - 2.2)], FABQPA < 12 [LR+ = 3.4 (1.05 - 11.20)], FABQW,< 10 [LR+ = 1.8 (1.02 - 3.15), prior episodes of neck pain ≥ 3 [LR+ = 1.9(1.3 - 2.7)], subjects report that looking up does not aggravate symptoms [LR+ = 4.8 (2.07 - 11.03)], subject report of physical exercise > 3 times / week [LR+ = 1.9 (1.1 - 3.4)], Cx extension ROM < 30° [LR+ = 2.5 (1.34 - 4.57)], decreased upper Tx kyphosis (%), [LR+ = 1.1 (0.77 -1.60)], shoulder protracted (%) [LR+ = 2.7 (1.6 - 3.0)] were significantly associated with treatment response.	Significant association of symptoms < 30days, no symptoms distal to the shoulder, looking up not aggravating symptoms, FAB score < 12, decreased upper Tx kyphosis, Cx extension ROM < 30° with treatment response.
Tseng (2006)	(1) Pain intensity (11-NRS) (2) Perceived global change (GROC) (3) Satisfaction (5-point Likert scale)	After treatment (one session).	Univariate analyses for significant differences between groups ($p < 0.10$) followed by a stepwise multiple logistic regression analysis ($p < 0.05$) Accuracy analyses.	Not performing sedentary work > 5h / day, diagnosis of cervical spondylosis without radiculopathy, lower score of NDI, no sudden onset, bilateral involvement pattern, not worse in Cx flexion, not worse in Cx extension, better while turning head, better while moving neck, no worse while moving neck negative compression tests in extended position significantly associated with successful outcome.	Significant predictors to successful outcome: NDI < 11.50, bilateral involvement pattern, not performing sedentary work > 5h / day, better while moving neck, no worse in Cx extension, diagnosis of spondylosis without radiculopathy.
Tuttle (2005, 2006)	(1) Pain intensity (11-VAS) (2) Pain location (body chart) (3) Total and limited AROM (°) (4) Disability (NDI, PSFS) (5) Perceived global change (GPES)	At discharge, mean time 6.1 days (2 - 14 days).	1-tailed paired sample t-tests, Spearman's rank order r_s (report of coefficient of determination r^2_s with corrected $p < 0.01$) Stepwise multiple regression analysis.	Change in outcome measures in first 2 treatments able to predict change in same outcome by end of treatment. Between - treatment changes in limited ROM predicted changes in limited ROM (r^2_s = 0.53 and 0.57) and total ROM (r^2_s = 0.26 and 0.26) by the end of treatment. Within and between-session changes in pain location predictive of changes in pain location (r^2_s = 0.24, 0.27, 0.28 and 0.57) at discharge. Perceived global change within 1st treatment predictive of perceived change at discharge (r^2_s = 0.32).	No improved ability of combinations of impairments to predict change in disability.
Back pain					
George (2005)	(1) Pain intensity (11-NRS) (2) Disability (ODQ)	6 months.	Hierarchical multiple regression ($p < 0.05$)		(1) Stepwise addition of CP significantly increased variance explained to 49% (F for R^2 change = 10.55, p = 0.004). Adding FABQW significantly increased variance to 61% (F for R^2 change = 5.73, p = 0.027). CP (β = -0.51, p = 0.013), FABQW (β = 0.36, p = 0.027) and initial disability (β = 0.44, p = 0.011) were retained in the parsimonious model. The model explained

					49% of variance and was a significant predictor of 6-month disability. (2) Stepwise addition of CP significantly increased variance explained to 40% (F for R^2 change = 4.62, p = 0.044). Only CP significantly contributed to the model (β = -0.47, p = 0.044). The parsimonious model included initial pain intensity (β = 0.33, p = 0.108) and CP explaining 29% of the variance and being a significant predictor of 6-month pain intensity.
Hicks (2005)	(1) Disability (ODQ)	8 weeks.	Univariate analysis using independent sample t- tests (Co) and x^2 tests (Ca) (p < 0.10). Accuracy analyses. ROC (Co) Forward stepwise logistic regression (p < 0.15).	Success: age < 40years [LR+ = 3.7 (1.6 - 8.3), average SLR >91° [LR+ = 3.3 (0.90 - 12.4), aberrant movement during Lx ROM [LR+ = 1.6(1.0 - 2.3), prone instability test [LR+ = 1.7 (1.1- 2.8)] Failure: FABQPA< 9 [LR- = 0.26 (0.08 - 0.78)], pain rating < 3 [LR- = 0.58 (0.25 - 1.3), Lx flexion ROM < 37% [LR - = 0.51 (0.21 -1.2)], discrepancy in SLR >10° [LR- = 0.32 (0.11 - 0.90)], 3 or less previous LBP episodes (LR = 0.76 (0.30 -1.0)], no increasing frequency of LBP episodes [LR- = 0.64 (0.43 - 0.95)], aberrant movement absent during Lx ROM [LR- = 0.39 (0.21 - 0.69)], no hyper mobility during Lx spring testing [LR- = 0.74 (0.59 - 0.96)], prone instability test [LR- = 0.39 (0.24 - 0.63)].	Success: 3 or more variables identified by the univariate analysis [LR+ = 4.0 (1.6 - 10). Failure: 2 or ore variables (FABQPA > 8, aberrant movement, prone instability test, hyper mobility during Lx spring testing) [LR- = 0.18 (0.08 - 0.38)].
Skytte (2005)	(1) Disability (LBPRS) (2) Back pain (LBPRS) (3) Leg pain (LBPRS) (4) Health status (NHP) (5) Medication use (6) Work loss (days off work)	1, 2, 3, 6 and 12 months.	Repeated measurement of analysis of variance (p < 0.05).	Significant differences between CP and non-CP groups for NHP, disability (1, 2, 3 and 12 months), back (3 months) and leg pain (2, 3 months).	
Niemisto (2004)	Symptoms and disability (ODQ)	12 months.	Univariate analysis (p ≤ 0.10). Multivariate analysis: K-means cluster analysis (unsupervised pattern recognition) (p < 0.05). Stepwise logistic regression. Discriminant analysis.	Whole group: Civil status: single or divorced [OR = 2.28 (1.2 - 4.4)], university education [OR = 2.65 (1.3 - 5.3)], mild to moderate pain intensity [OR = 5.22 (2.8 - 9.8), work absence > 25 days [OR = 3.45 (1.3 - 9.0)], poor work ability [OR = 1.70 (1.0 - 3.0)], poor self rated prognosis of work ability after 2 years [OR = 1.75 (0.9 - 3.4)], poor life control [OR = 1.96 (0.9 - 4.2)], weak social support [OR = 1.85 (1.0 -3.3)], finger floor distance in forward flexion < 20 cm [OR = 1.75 (0.9 - 3.3)] SLUMP test [OR = 1.89 (1.1 - 3.3)]. Manipulative treatment: Civil status: single or divorced [OR = 2.19 (0.9 - 5.3)], university education [OR = 2.65 (1.0 - 6.7)], severe affective distress [OR = 2.65 (1.1 - 6.2)], non-CP [OR = 2.60 (0.9 -7.2)], SLUMP test [OR = 2.02 (0.9 - 4.6)]. Consultation only: Civil status: single or divorced [OR = 2.77 (1.0 -7.6)], university education [OR = 3.0 (1.0 - 8.6)], mild to	Whole group: university education [OR = 2.80 (1.1 - 6.9)], mild to moderate level of pain intensity [OR = 6.33 (2.8 - 14.3], work absence > 25 days during previous year [OR = 4.19 (1.5 - 11.3)], poor self-evaluated prognosis for work ability after 2 years [OR = 2.11 (0.9 - 5.0)], poor life control [OR = 2.77], normal SLUMP test [OR =1.96 (0.9 - 4.1)]. Overall model (74%). Manipulative treatment: mild to moderate level of pain intensity [OR = 5.99 (2.0 - 18.3)], severe affective distress [3.81 (1.3 -10.8)], non-CP [OR = 2.71 (0.8 - 9.6)]. Overall model (69%). Consultation only: university education [OR = 7.93 (1.6 - 39.7)], mild to moderate pain intensity [OR = 5.38 (1.5 -19.3)], work absence > 25 days

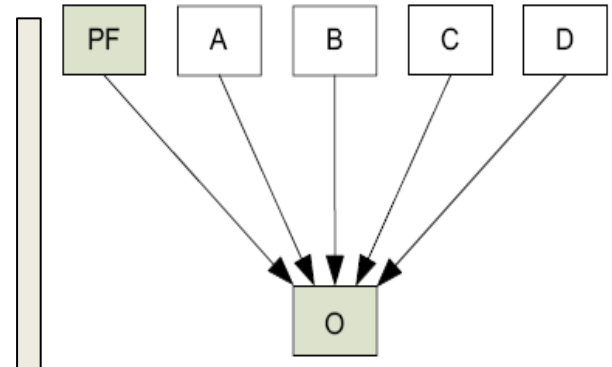
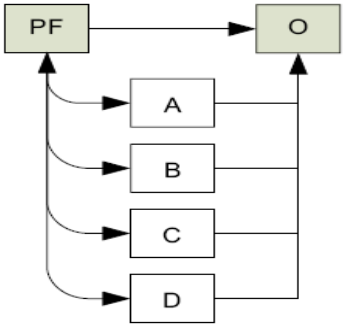
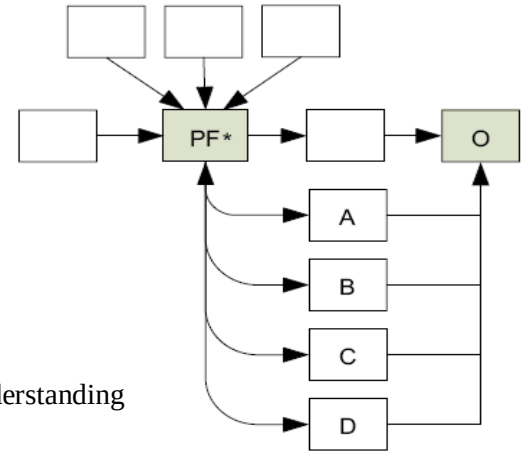
				moderate pain intensity [OR = 4.73 (2.0 - 11.4)], work absence > 25 days [OR = 7.33 (1.7 - 32.2), poor work ability [OR = 2.55 (1.1 - 5.8)], poor self-rated prognosis of work ability after 2 years [OR = 2.18 (0.9 - 5.4), weak social support [OR = 2.07 (0.9 - 4.6)], MSPQ ≥9 [OR = 2.53 (1.0 - 6.7)], finger floor distance in forward flexion [OR = 2.52 (1.0 - 6.1)].	[OR = 19.64 (3.8 - 102.5), MSPQ [OR = 3.18 (0.9 - 11.6), poor life control [OR = 9.40 (1.9 - 47.0). Overall model (76%).
Hahne (2004)	(1) Pain intensity (11-NRS) (2) AROM (°)	Mean time 4.8 days (2-11 days).	Linear regression analysis Pearson product moment r, coefficient of determination (r^2), RMS Subgroup analyses.	Within session changes in ROM LR+ ranging from 2.6 to 19.0 (0.9 -138.8), LR- = 0.3 to 0.8 (0.2 - 0.5), OR = 3.5 to 37.0 (0.8 - 330.8). Within session changes in pain LR+ = 2.3 to 4.4 (1.0 - 10.3), LR- = 0.3 to 0.5 (0.1 - 0.9), OR = 4.5 to 15.6 (1.3 - 68.4).	
Werneke (1999, 2001, 2004)	(1) Pain intensity (11-NRS) (2) Activity interference, downtime at home (3) RTW (4) Work loss (5) Healthcare use	12 months.	Univariate analysis: Two sample t-tests, x^2 test ($p \leq 0.05$). Multiple logistic regression.	(1): Non-organic physical signs, perceived disability at discharge, pain pattern classification. (2) Pain at intake, overt pain behaviours, perceived disability at discharge, pain pattern classification. (3) Multiple sites of pain, leg pain at intake, pain at intake, payer, overt pain behaviours, fear of work activities, perceived disability at discharge, pain pattern classification. (4) Perceived disability at discharge, pain pattern classification.	(1), (2), (3), 5) Pain pattern classification. (4) Leg pain at intake. (1) OR = 3.0 (1.4 - 6.4); (2) OR = 5.2 (2.4 - 11.3); (3) OR = 9.4 (3.4 - 26.0); (4) OR = 4.0 (1.5 - 10.5); (6) OR = 4.4 (2.0 -10.1).
Enthoven (2003)	(1) Pain intensity (VAS 0-100) (2) Disability (ODQ)	4 weeks, 12 months.	Mann-Whitney U test, x^2 test or Fisher's exact test. Spearman Rank sum correlation coefficient ($p<0.05$). Linear regression (12 months).	Low to moderate associations of changes in ROM, endurance and pain after examination with changes in disability (-0.22 to 0.64), and pain (-0.22 to 0.49).	Physical measures at baseline did not predict disability or pain, except for isometric endurance of back flexors for pain intensity at 12 months. At the 4-week examination, thoracolumbar rotation, isometric endurance back extensors and fingertip to floor distance were significant predictors of disability (r = -0.35, -0.41 and 0.40) and pain at 12 months (r = -0.32, -0.44 and 0.30).
Flynn (2002)	(1) Disability (OSW)	After treatment (2 -3 sessions).	Univariate analysis using independent sample t- tests (Co) and x^2 tests (Ca) ($p < 0.15$) Co: accuracy analysis, ROC curves Stepwise logistic regression ($p < 0.05$).	Success: FABQW, back symptoms only, symptoms distal to knee, duration of symptoms, increasing episode frequency, standing ranked as worse position, left and right hip internal rotation, hypomobility and pain with Lx spring testing, peripheralization with single lumbar movement testing.	Duration of symptoms < 16 days, at least one hip with > 35° internal rotation, hypomobility with Lx spring testing, FABQW < 19, no symptoms distal to knee. Overall model $x^2 = 48.5$, df = 5, $p< 0.001$, $R^2 = 0.67$).
Viikari-Juntura (1998)	(1) Work loss (days off work)	60 days.	Preliminary analysis followed by log linear modelling: multinomial logit models fitted by GLIM ($p < 0.05$).	Male gender, older age, blue collar occupation, sick leave 60 days before examination, most items on physical examination.	Gender [OR = 0.3 (0.1 - 1.4)], age [OR = 2.8 to 5.7 (1.1 - 19.7)], duration of symptoms [OR = 0.1 to 0.2 (0.0 - 0.6)], onset of symptoms [OR = 0.0 to 0.1 (0.0 - 0.4)], pain location [OR = 0.6 to 4.5 (0.2 - 15.0)], mode of pain [OR = 1.4 to 5.8 (0.6 - 19.4)], relief when lying [OR = 3.7 to 11.6 (1.3 - 53.9), trouble at work [OR = 3.5 to 7.0 (1.3 - 23.9)], pain in low back or buttock in lateral

					flexion [OR = 3.3 to 7.0 (0.9 - 24.7), side difference in SLR [OR = 1.0 to 5.9 (0.3 - 24.0)].
Karas (1997)	(1) Return to work (yes / no)	6 months.	x^2 tests ($p < 0.05$), Fisher's exact test. Logistic regression. Analysis of variance and Bonferroni post hoc analysis.	Patients with centralised symptoms returned to work more frequently than patients whose symptoms did not ($x^2 = 4.31$, $p < 0.05$). Patients with low scores returned to work more often than patients with high scores ($x^2 = 7.53$, $p = 0.006$). Among centralisers, more patients with low Waddell scores returned to work than patients with high Waddell scores ($p = 0.0003$). For patients who did not centralise, the Waddell score did not have a significant effect on RTW. Among patients with low Waddell scores, those who centralised had a higher RTW than non-centralisers. For patients with high Waddell scores, there was no difference between centralisers and non-centralisers in RTW.	For patients with centralised symptoms, the probability of returning to work increased with a low Waddell score ($p = 0.0005$). RTW depended on the interaction between CP and Waddell scores ($p = 0.0037$).
Van den Hoogen (1997)	(1) Time to recovery: number of weeks from initial visit to end of episode (2) Occurrence of relapse: LBP reports in 1 or more of the follow-up weeks after the end of episode present at initial visit.	12 months.	Bivariate correlation analysis ($p < 0.05$) followed by multivariate analysis. Kaplan-Meier survival curve, Cox regression with forward stepwise selection (time to recovery). Logistic regression (occurrence of relapse).	(1) Duration of the LBP episode preceding the initial visit, sciatica, maximal lumbar flexion, three aspects of perceived health, and receiving physical therapy. (2) Daily functioning.	(1) Duration preceding the initial visit (number of weeks) [HR = 0.98 (0.97 - 0.99)], receiving PT during first 5 weeks after initial visit [HR = 0.62 (0.49 - 0.78), pain as an aspect of perceived health [HR = 0.99 (0.99 - 1.0)], history of surgery [HR = 0.58 (0.36 - 0.94)]. Significance of overall model $p < 0.001$. (2) Daily functioning (2% variance explained).
Burton (1995, 2004)	(1) Disability (RMDQ) (2) Recurrence	(a) 12 months (b) 4 years.	Univariate analysis, t-tests and x^2 tests ($p < 0.05$). Stepwise multiple regression (all patients, acute patients, sub-chronic patients) ($p < 0.05$). Discriminant analysis.	(1a): No information available.	(1a) All patients: Coping strategies CSQ, praying / hoping, PPI, Somatic perceptions (MSPQ), SLR, root tension signs (39% of variance explained). (1a) Acute patients: Coping strategies CSQ catastrophising, somatic perceptions (MSPQ), SRL, coping strategies CSQ, praying / hoping, Leg pain (69% of variance explained). (1a) Sub-chronic patients: disability (initial RMDQ), PPI (18% of variance explained). (2a) Baseline Modified Zung Depression Index score, Baseline PPI (26% of variance explained). (2b) Longer duration of presenting symptoms, presence of leg pain, higher FAB, heightened somatic concern ($p < 0.05$).
Long (1995)	(1) Pain intensity (NRS-101) (2) Lifting capacity (3) Disability (ODQ) (4) Return to work	(a) 9 months (b) 24 months.	Multivariate analysis of variance.	CP demonstrated a greater decrease in maximum pain intensity compared to the non-CP groups at discharge. Significant difference in RTW status between CP and non-CP groups ($x^2 = 4.49$, $p = 0.034$) at 9 but not 24 months.	

Neck and back pain				
Werneke (1999, 2003)	(1) Pain intensity (11-NRS) (2) Disability (ODQ/ NDI) (3) Number of treatment visits	Discharge, 12 months.	One or two-way ANOVA with Sheffe post hoc analyses ($p < 0.05$). ANCOVA for discriminant validity of one versus multiple visit definition.	(1), (2) Statistically significant difference in pain intensity and disability between CP and non-CP groups. (3) CP had fewer visits than the partial and non-CP groups.
Hellsing (1994)	(1) Work loss	12 months.	Pearson x^2 tests, independent samples t-tests.	Previous sick leave $x^2 = 6.79$, $df = 1$, $p = 0.009$[OR = 6.24 (1.35 - 28.82)]. More than three positive pain tests $x^2 = 5.32$, $df = 1$, $p = 0.02$ [OR = 4.78 (1.16 - 19.72)]. Peripheral symptoms $x^2 = 7.44$ $df = 1$, $p = 0.0067$ [OR = 7.8 (1.4 -42.6)].

Abbreviations:
AROM, Active Range of Motion; Ca, Categorical; CP, Centralisation Phenomenon; Co, Continuous; CSQ, Coping Strategies Questionnaire; Cx, Cervical; FAB, Fear Avoidance Beliefs; FABQPA, Fear Avoidance Beliefs Questionnaire Physical Activity Subscale; FABQW, Fear Avoidance Beliefs Questionnaire Work Subscale; GROC, Global Rating of Change; GPES, Global Perceived Effect Scale; HR, Hazards Ratio; LBP, Low Back Pain; LBPRS, Low Back Pain Rating Scale; LR, Likelihood Ratios; Lx, Lumbar; MSPQ, Modified Somatic Perception Questionnaire; NDI, Neck Disability Index; NHP, Nottingham Health Profile; MSPQ, Modified Somatic Perception Questionnaire; NRS, Numerical Rating Scale; ODQ, Oswestry Disability Questionnaire; OR, Odds Ratio; OSW, Modified Oswestry Disability Questionnaire; PPI, Present Pain Intensity; PSFS, Patient Specific Functional Scale; RMDQ, Ronald Morris Disability Questionnaire; RMS, Root Mean Square; ROC, Receiver Operating Characteristics; ROM, Range of Motion; RTW, Return to Work; SLR, Straight Leg Raise; Tx, Thoracic; VAS, Visual Analogue Scale.

Table 3.5 Types of explanatory studies (adapted from: Altman and Lyman, 1998; Hayden et al., 2008, with permission). A, B, C, D, confounders; PF, prognostic factor; O, outcome.

Explanatory study type	*Description*
Exploration **Phase 1: identifying associations** 	Phase 1 studies provide hypothesis generating evidence indicative of a potential association of a prognostic factor with an outcome. These studies do not focus on one specific factor but instead, the prognostic factor of interest is investigated as one of many factors assessed for their association with outcome.
Phase 2: testing independent associations  Confirmation	Phase 2 studies measure the independent effect of a prognostic factor while controlling for confounders. Here, the importance shifts to assessing whether the factor independently adds to currently known prognostic factors.
Phase 3: understanding prognostic pathways  Understanding	Phase 3 studies attempt to describe the complexity of the prognostic pathways or processes. These studies usually apply knowledge from the previous phases of study on independent associations and incorporate other knowledge from the field of study.

3.4.3 Final approach to synthesis

The identified variability of symptom responses, methodological characteristics of the studies and the limited available data in other cases prevented the use of any meaningful meta-analytic techniques. Consequently, the synthesis of the data from included studies followed an approach taking into account only the consistency and validity of results (Slavin, 1995) (Tables 3.7 and 3.8). Levels of evidence for studies of prognosis (Scholten – Peeters et al., 2003) were applied, presented in Table 3.6.

Table 3.6 Levels of evidence for studies of prognosis (Scholten – Peeters et al., 2003)*.

- **Strong evidence:** Consistent findings from at least two high quality studies
- **Moderate evidence:** One high quality study and consistent findings among low quality studies
- **Limited evidence:** Findings in one cohort or consistent findings among low quality studies
- **Conflicting evidence:** Inconsistent findings among multiple studies
- **No evidence:** No studies

*Reproduced with permission of the International Association for the Study of Pain® (IASP®)

Some recent systematic reviews (Cote et al., 2001; Carroll et al., 2008a; Carroll et al., 2008b) propose making conclusions on the strength of recommendations based on a 3-level hierarchy of study type (Table 3.5). This system does not consider the quality or amount of identified evidence. However, because the study type was not included in the study quality assessment, it was decided *post hoc* to produce a table of statistically significant results where the strength of evidence regarding clinically induced symptom responses was compared to whether results derived from a Phase I, II, or III study (Table 3.8).

3.4.4 Clinically induced symptom responses as a prognostic factor

There was limited evidence for a positive or negative association for the majority of findings (Table 3.7). In patients with neck pain (NP), only changes in pain location within the same treatment session demonstrated a significant association with changes in pain between sessions. In LBP, strong evidence was found for a positive association between the lack of changes in pain location (and / or intensity) with lumbar motion testing and worse LBP symptoms and work status. However, there was evidence that lack of changes in pain location was not associated with more days of sick leave and the lifting capacity of those with chronic pain. There was also limited evidence for an association between changes in pain location with lumbar motion testing and health status. The evidence regarding symptom response to movement testing and disability or health care use was inconclusive.

Limited evidence was found that symptom changes during side-bend (regardless of the direction of movement) were associated with work loss, as opposed to other movements; however, patients recruited in this study represented a more severe spectrum of patients with suspected nerve root involvement (Viikari – Juntura, 1998). The evidence on neurodynamic testing was not consistent. Although there was a significant association with symptoms and disability in one high quality study (Niemisto et al., 2004), neural tension tests were not associated with symptoms and work loss in two studies (Viikari-Juntura, 1998; van den Hoogen et al., 1996).

Pain on segmental testing was only reported in LBP. Regarding instability tests for stabilisation programmes, there was limited evidence that only the prone instability

test was significantly associated with disability at 8 weeks (Hicks et al., 2005). There was also limited evidence that the sacroiliac pain provocation tests were not associated with disability in LBP patients (Flynn et al., 2002). Reporting more pain after physical examination was not associated with symptoms and disability (Enthoven et al., 2003). However, changes in pain intensity within the same session were predictive of changes in the same impairment between sessions (Hahne et al., 2004).

3.4.5 Results of the sensitivity analysis

Removing lower quality and smaller studies did not affect the overall conclusions. The only variations in review findings were in the evidence regarding changes in pain location within the same treatment session in NP. When excluding studies with smaller sample sizes (< 30) (Riley et al., 2003), evidence changed from limited to no evidence.

Table 3.7 Overall strength of evidence of statistical associations between clinically induced symptom responses and outcomes in spinal pain.

Prognostic factor	Outcome	Cohorts assessed	+ Findings	High quality	Low quality	- Findings	High quality	Low quality	Level of evidence
Neck pain									
Changes in pain location with Cx motion testing	Perceived global change	1				Cleland 2007	1		Limited
Compression testing	Perceived global change	1				Cleland 2007	1		Limited
	Symptoms, perceived global change or satisfaction	1				Tseng 2006	1		Limited
Cx distraction testing	Perceived global change	1				Cleland 2007	1		Limited
	Symptoms, perceived global change or satisfaction	1				Tseng 2006	1		Limited
Neurodynamic testing	Perceived global change	1				Cleland 2007	1		Limited
Changes in pain location with treatment	Symptoms (location)	1	Tuttle 2005, 2006	1					Limited
	Symptoms (intensity)	1				Tuttle 2005, 2006	1		Limited
	ROM	1				Tuttle 2005, 2006	1		Limited
	Disability	1				Tuttle 2005, 2006	1		Limited
	Perceived global change	1				Tuttle 2005, 2006	1		Limited
Changes in pain intensity with treatment	Symptoms (location or intensity)	1				Tuttle 2005, 2006	1		Limited
	ROM	1				Tuttle 2005, 2006	1		Limited
	Disability	1				Tuttle 2005, 2006	1		Limited
	Perceived global change	1				Tuttle 2005, 2006	1		Limited
Back pain									
Changes in pain location with Lx motion testing	Symptoms	4	Long 1995 Werneke 2001, 2004 George 2005 Skytte 2005	4					Strong
	Disability	5	Werneke 2001, 2004 George 2005 Skytte 2005	3		Long 1995 Flynn 2002	2		Inconclusive
	Symptoms and disability	1				Niemisto 2004	1		Limited
	Strength	1				Long 1995	1		Limited
	Health status	1	Skytte 2005	1					Limited
	Healthcare use	2	Werneke 2001, 2004	1		Skytte 2005	1		Inconclusive
	Work loss	2				Werneke 2001, 2004 Skytte 2005	2		Strong
Changes in pain location or intensity with Lx motion testing	Work status	3	Long 1995 Karas 1997 Werneke 2001, 2004	3					Strong
Pain during Lx motion testing (SB)	Work loss	1	Viikari-Juntura 1998						Limited
Pain during Lx motion					1	Viikari-Juntura 1998		1	Limited

testing (other)							
Pain on spring testing (Lx)	Disability	1			Flynn 2002	1	Limited
Lumbar instability tests							
Prone instability test (Lx)	Disability	1	Hicks 2005	1			Limited
Posterior shear test (Lx)	Disability	1			Hicks 2005	1	Limited
SI dysfunction tests							
Posterior shear test (SIJ)	Disability	1			Flynn 2002	1	Limited
Gaenslen test	Disability	1			Flynn 2002	1	Limited
Sacral thrust test	Disability	1			Flynn 2002	1	Limited
Resisted hip abduction	Disability	1			Flynn 2002	1	Limited
Compression-distraction test	Disability	1			Flynn 2002	1	Limited
Sacral sulcus test	Disability	1			Flynn 2002	1	Limited
Patrick test	Disability	1			Flynn 2002	1	Limited
Neurodynamic testing	Symptoms	2	Burton 1995, 2004	1	Van den Hoogen 97	1	Inconclusive
	Disability	1	Burton 1995	1	Burton 2004	1	Limited
	Symptoms and disability	1	Niemisto 2004	1			Limited
	Work loss	1			Viikari-Juntura 1998	1	Limited
Changes in pain intensity	Symptoms	1			Enthoven 2003	1	Limited
with physical examination	Disability	1			Enthoven 2003	1	Limited
Changes in pain intensity with treatment	Symptoms	1	Hahne 2004	1			Limited
Neck and back pain							
Number of positive pain tests on movement	Work loss	1	Hellsing 1994	1			Limited
Changes in pain location	Symptoms	1	Werneke 1999, 2003	1			Limited
with Cx or Lx motion testing	Disability	1	Werneke 1999, 2003	1			Limited
	Healthcare use	1	Werneke 1999, 2003	1			Limited

Note: + / - Findings indicate the presence / absence of a statistically significant association respectively.

Abbreviations:

Cx, Cervical; Lx, Lumbar; ROM, Range of Motion; SB, Side-Bend; SIJ, Sacro-iliac Joint.

Table 3.8 Statistically significant associations between clinically induced symptom responses and reported outcomes.

OUTCOME	CLINICALLY INDUCED SYMPTOM RESPONSES (STRENGTH OF EVIDENCE)	SPINAL AREA	PRIMARY AUTHOR	STUDY TYPE
Symptoms	Changes in pain location with treatment r^2s = 0.24, 0.27, 0.28 and 0.57 at discharge (Limited)	NP	Tuttle 2005, 2006	Phase 1
	Changes in pain intensity with treatment LR+ = 2.3 to 4.4 (1.0 - 10.3), LR- = 0.3 to 0.5 (0.1 - 0.9), OR = 4.5 to 15.6 (1.3 - 68.4) (Limited)	LBP	Hahne 2004	Phase 1
	Changes in pain location with lumbar motion testing (Limited)	NP & LBP	Werneke 1999, 2003	Phase 1
	Changes in pain location with lumbar motion testing β = -0.51, OR = 3.0 (1.4 - 6.4) (Strong)	LBP	Long 1995	Phase 1
			Werneke 2001, 2004	Phase 2
			George 2005	Phase 2
			Skytte 2005	Phase 1
Disability	Changes in pain location with lumbar motion testing (Limited)	NP & LBP	Werneke 1999, 2003	Phase 1
	Prone instability test LR+ = 1.7 (1.1- 2.8), LR- = 0.39 (0.24 - 0.63) (Limited)	LBP	Hicks, 2005	Phase 2
	Neurodynamic testing (Limited)	LBP	Burton 1995	Phase 2
Symptoms and disability	Neurodynamic testing OR = 2.71 (0.8 - 9.6) (Limited)	LBP	Niemisto 2004	Phase 2
Work status	Changes in pain location or intensity with lumbar motion testing x^2 = 4.31, OR = 9.4 (3.4 - 26.0) (Strong)	LBP	Long 1995	Phase 1
			Karas 1997	Phase 1
			Werneke 2001, 2004	Phase 2
Work loss	Pain during lumbar motion testing OR = 3.3 to 7.0 (0.9 - 24.7 (Limited)	LBP	Viikari-Juntura 1998	Phase 2
	Number of positive pain tests on movement x^2 = 5.32, OR = 4.78 (1.16 - 19.72) (Limited)		Hellsing 1994	Phase 2
Health status	Changes in pain location with lumbar motion testing (Limited)	LBP	Skytte 2005	Phase 2
Health care use	Changes in pain location with lumbar motion testing (Limited)	LBP	Werneke 1999, 2003	Phase 1

Abbreviations:

LBP, Low Back Pain; LR, Likelihood Ratio; NP, Neck Pain; OR, Odds Ratio.

3.5 DISCUSSION

Symptom changes in response to repeated movement testing and as a response to treatment offer promise as a predictor of LBP and NP outcomes, but this requires further investigation in future research, particularly for longer term prediction across a range of outcomes. Most of the included studies involved preliminary correlation designs. Despite the routine use of such designs in prognostic factor studies (Gamsa, 1994), their value in assessing etiological significance is limited. Conclusions drawn from such designs are often hypothesis generating rather than conclusive (Gamsa, 1994).

Several reviews have indicated that clinically induced changes in pain location and / or intensity may have a role in the prognosis and treatment of LBP patients (Wetzel and Donelson, 2003; Aina et al., 2004; Cook et al., 2005; Wessels et al., 2006). The findings in this chapter are in agreement with some of the above suggestions. In addition, several randomised controlled trials have provided preliminary evidence regarding the value of these responses in directing effective treatment in LBP (Fritz et al., 2003; Long et al., 2004; Brennan et al., 2006; Browder et al., 2007). However, more work is needed before clinically induced symptom responses can be recommended for routine clinical use and further research is required to investigate the relative importance of these clinical factors to other assessment domains in robust study designs (DIHTA, 1999).

Changes in the patient's condition during assessment and subsequent treatment have traditionally been used to guide prognosis and refine treatment interventions (APTA,

2001; Herbert et al., 2005). Clinically induced changes with often an immediate effect on clinical outcomes have recently attracted more attention (Axen et al., 2005b; Leboeuf-Yde et al., 2005; Tong et al., 2006) because these changes influence patients' perceived benefit of treatment and reasons for attending subsequent sessions (Grimmer et al., 1999). In a Swedish study on chiropractic treatment, improvements at the fourth visit could be predicted through patients' progress by the second visit (Axen et al., 2002; Axen et al., 2005b). Furthermore, patients who did not report definite improvement by the fourth visit were less likely to report definite improvement in 3-month and 12- month outcomes (Leboeuf-Yde et al., 2005). Clinically induced changes may be associated with a lasting effect, an assumption that has been consistent with findings of studies investigating clinically induced symptom responses in the United States and Australia (Werneke and Hart, 2003; Hahne, 2004; Tuttle, 2005; Tuttle et al., 2006).

With the exception of clinically induced symptom response with repeated movements and neurodynamic testing, the evidence on the association of most investigated clinically induced symptom responses with spinal outcomes was limited, with results mainly deriving from single prognostic studies. This may compromise the generalisability of most review findings since the validity of a prognostic factor can not be extrapolated directly to other populations unless it is reproduced across different independent samples (Kent & Keating, 2008). There was also a striking scarcity of prognostic studies in spinal areas other than the lumbar spine. Some studies on patients with NP were identified (Tuttle, 2005; Tuttle et al., 2006; Tseng et al., 2006; Cleland et al., 2007). However, none of these studies provided a

sufficiently long follow-up period (≥ 3 months, see Chapter 1). Therefore, these identified factors may not be important for the prognosis of the course of neck pain. Identifying factors that contribute to the prognosis of NP is highly desirable especially since prognosis and the relative importance of various factors in predicting future NP outcomes are largely unknown (Carroll et al., 2008d).

In LBP, the comparability of findings across studies was often difficult to achieve because of the variability of the study characteristics. Differences in the study population, operational definitions of prognostic variables, treatment, follow-up assessments and statistical analysis were often noticed between studies. Variations were also observed in the way prognostic questions were asked, or the purpose of the study. The above may explain the conflicting evidence found. For example, clinically induced symptom responses to movement testing (i.e. centralisation) had inconsistent results with regard to disability outcomes. Some of the studies failing to identify a statistically significant association used single movements (Flynn et al., 2002; Cleland et al., 2007), in contrast to studies using repeated movement testing (Long, 1995; Karas et al., 1997; Werneke et al., 1999; Niemisto et al., 2004; George et al., 2005; Skytte et al., 2005) to elicit this sign. Observed inconsistencies in reported definitions of centralisation have often been criticised and the standardisation of the definition of centralisation has been recommended (George and Fritz, 2005; Werneke and May, 2005; Berthelot et al., 2007).

Publication bias, i.e. the selective publication of studies based on the direction and strength of their results (Dickersin and Min, 1993), has been acknowledged as an

important problem in systematic reviews of prognosis (Altman, 2001). Studies showing significant findings are more likely to be published, translated into English and cited by other authors (Egger et al., 1997; Kyzas et al., 2005). This may result in a greater risk of missing papers that do not report significant findings; especially from small non-randomised studies (Easterbrook et al., 1991; Dickersin and Min, 1993). This review may be overstating the evidence for clinically induced symptom responses because of publication bias. A variety of strategies and sources were used to ensure that relevant studies were not missed. However, due to resource limitations, articles written in foreign languages and the two studies whose full text could not be found were excluded from this review.

Sensitivity analyses demonstrated that no conclusion of this chapter was affected by sample size and study quality. However, only one study was judged as low quality. Furthermore, conclusions in this chapter were primarily based on the overall level of evidence rather than the strength of the association between clinically induced symptom responses and spinal outcomes. This approach is common in systematic reviews of prognosis in NP (Scholten – Peeters et al., 2003; Williams et al., 2007; Williamson et al., 2007). However, the confidence in the utility of a prognostic factor is greater if the strength of the association with an outcome or effect sizes are available.

The issue of what constitutes a clinically important change or association between symptom response groups has not been clarified to date. The lack of reports regarding absolute changes in outcomes or the inability to use established outcome

measures have accounted for this confusion in some cases (Aina et al., 2004). In the light of this uncertainty, conclusions on the value of prognostic variables in many included studies were based on significance testing rather than judgments about clinical importance. It is therefore no surprise that conclusions in this chapter were also primarily based on the overall level of evidence rather than the strength of the association between clinically induced symptom responses and spinal outcomes.

Statistical significance indicates whether the hypothesis of no prognostic effect can be ruled out (Simon and Altman, 1994). Many studies arbitrarily dichotomised outcome scales and defined good or poor recovery at some point of the scale. Despite this being a quite common (Altman and Lyman, 1998; Altman and Royston, 2006; CRD, 2008) and legitimate approach in some clinical research (Hand, 1994), it can severely hamper the statistical power to detect a significant relationship (Altman and Royston, 2006; Royston et al., 2006).

The issue of distinct prognostic factors, their interrelations and relative importance remains unclear in spinal pain (Aina et al., 2004; Fejer and Hartvigsen, 2008). There is also a lack of theoretical models to explain prognostic associations (Kent and Keating, 2008; Kent et al., 2008). However, it is widely accepted that the course of spinal pain is likely to have multifactorial influences rather than depend on one factor (Carroll et al., 2008d). This review attempted to identify which clinically induced symptom responses are important in the prediction of outcomes in the conservative management of spinal pain. In view of the uncertainty in the strength of identified

associations and the extent of confounding between investigated prognostic factors, this goal has only been partially achieved.

3.6 SUMMARY

In this chapter, a comprehensive and systematic investigation of the literature on all clinically induced symptom responses across all spinal areas was undertaken. Symptom response to repeated movement testing and to treatment, often referred to as the centralisation phenomenon, may offer promise in the prognosis of LBP and NP outcomes, but this requires further investigation in future research, particularly for longer term prediction across a range of outcomes. Definitions and methods to identify centralisation need to standardised and further work is needed using robust study designs before symptom response can be used to inform management.

Chapter 4

A Delphi study on the centralisation phenomenon.

4.1 AIMS OF CHAPTER

This chapter presents a Delphi study on the centralisation phenomenon. The primary aim of this study was to achieve consensus on the operational definition of centralisation as a physical examination sign. The operational definition for physical measurements such as centralisation is 'a set of procedures that guides the process of obtaining a measurement, including descriptions of the attribute that is to be measured, in order to obtain the measurement' (Rothstein et al., 1991, p. 596). Emphasis was also placed on establishing criteria for the testing procedure and categories of classification for clinical practice and research. A secondary aim was to identify further issues for future research into centralisation.

4.2 INTRODUCTION

A clinical marker commonly used in the assessment of spinal patients is the centralisation phenomenon (CP). Definitions and methods to identify centralisation vary (Chapter 1) with studies not consistently favouring one definition over another (Werneke and May, 2005). This variation leads to differences in the prevalence, classification categories, prognostic indicators, management options and outcomes (George and Fritz, 2005; Werneke and May, 2005), and difficulties when comparing research findings between studies (Chapters 2 and 3). The standardisation of centralisation with the establishment of a core set of criteria has been frequently recommended (George and Fritz, 2005; Werneke and May, 2005; Berthelot et al., 2007; Werneke et al., 2008).

4.3. MATERIALS AND METHODS

4.3.1 Research background

Before the start of this project, relevant studies and discussion papers were searched and retrieved. Two systematic reviews on clinically induced symptom responses were produced as part of this effort, which are presented elsewhere (Chapters 2 and 3). Literature and reviews on existing classification systems for the conservative management of neck (Buchbinder et al., 1996; Childs et al., 2004b; Sterling, 2004) and back pain (Riddle, 1998; Petersen et al., 1999; Buchbinder et al., 1996; Cieza et al., 2004; McCarthy et al., 2004; O'Sullivan, 2005; Machado et al., 2006; Billis et al., 2007) were also explored before and throughout this project.

Contact with relevant individuals and organisations was sought[9]. From these communications, it was clear that there were no formal research collaborations between specialist interest groups who have developed classifications for spinal pain that include centralisation. However, informal unsuccessful attempts of collaboration between the McKenzie and Delitto groups had been made in the past. Communication appeared to be exclusively through the peer-reviewed literature thereafter. Reactions about the use of consensus to establish a uniform definition for centralisation were mixed. Some individuals expressed an immediate interest in being involved in a consensus effort whereas others expressed reservations to such an endeavour. The main arguments for participating in the study were that published literature so far has not resulted in resolving this issue and that a consensus

[9] This involved communications with the following: Stephen May, email communications 17th March and 31st May 2006; John Childs, email communication 24th July 2006; Anthony Delitto, email communication 25th July 2006; Steven George, email communication 2nd August 2006; Ron Bybee, email communication 9th August 2006; Audrey Long, email communication 18th September 2006; Thomas Dreisinger, email communication 20th September 2006; Julie Fritz, email communication 22nd September 2006; Gerard Brennan and Eric Parent, email communication 5th December 2006.

definition might be worth being tested further in future research. Arguments against participating included the notion that superiority of and preference for a particular definition should be demonstrated through research reports rather than anecdotal opinions, concern about resulting in a meaningless definition or difficulties in implementing change because of conflict of interest and lack of time of participants.

4.3.2 Consensus methods

Consensus methods provide a means of synthesising the insights of participants to create a product to be used with more confidence in the future (Cross, 2005). Consensus methods may be particularly important in the case of centralisation, since judgements on the definition and measurement of this sign have derived from a few clinicians rather than being the result of a collective decision (McCarthy et al., 2004). Formal consensus development methods are proposed as an initial step to resolving the issue of a lack of standard definition of centralisation in this chapter.

It is only since the 1950s that formal consensus methods have been used in healthcare (Murphy et al., 1998). Collective decisions were made before the introduction of these methods, but these decisions mainly resulted from informal discussions and agreement (Murphy et al., 1998). Some of the identified disadvantages were that the processes involved in informal group meetings did not promote decision – making. Domination of the discussion by particular individuals and limitation of ideas arising from the group due to time constraints were some examples of the problems often encountered (Murphy et al., 1998). Furthermore, the desire to reach agreement often overrode concerns about the accuracy of the result to

the extent that there was premature closure on a particular solution without consideration of alternatives (Janis, 1982).

When using formal consensus methods, the structure, process and output is explicit and established from the outset (Frances et al., 1998; Murphy et al., 1998). Because these methods are based on a systematic manner of reaching a group decision, they can be more advantageous over individual or informal group approaches (Murphy et al., 1998; Hicks, 2004). Safety in numbers (i.e. several people are more likely to reach a better decision than a single individual), authority (i.e. a selected group of individuals is more likely to lend some authority to the decision made), rationality (i.e. decisions are improved by reasoned argument in which assumptions are challenged and members are forced to justify their views) and scientific credibility (i.e. formal consensus methods meet the requirements of scientific methods) are some of the proposed benefits of formal consensus methods (Murry and Hammons,1995; Murphy et al., 1998).

Three main formal consensus approaches have been used in healthcare; the Delphi method, the nominal group technique (NGT) and the consensus development conference (Murphy et al., 1998; Pope and Mays 2000; Cross, 2005). All the above methods aim to reach consensus of opinion of a group of knowledgeable individuals often referred to as 'experts' (Murphy et al., 1998). The Delphi method, first introduced in the 1950s, involves a series of mailed questionnaires or consecutive 'rounds' interspersed by controlled feedback; the opinions of the expert panel (Couper, 1984) are then statistically aggregated in the light of the Delphi group feedback (Murphy et al., 1998). The NGT, which followed in the 1960s, involves the

discussion of proposed ideas 'face – to – face' in the presence of a group facilitator; this method also uses statistical aggregation techniques to analyse group responses (Fink et al., 1984; Murphy et al., 1998). Finally, in the consensus development conference, first used in the late 1970s, participants make decisions following an open meeting format (Murphy et al., 1998). Table 4.1 provides a comparative presentation of the characteristics of the above three formal consensus methods.

Table 4.1 Characteristics of the Delphi method, Nominal Group Technique (NGT), and consensus development conference (Murphy et al., 1998).

Consensus development method	Mailed question-naires	Private decisions elicited	Formal feedback of group choices	Face-to-face contact	Interaction structured	Aggregation method
Delphi method	Yes	Yes	Yes	No	Yes	Explicit
NGT	No	Yes	Yes	Yes	Yes	Explicit
RAND version	Yes	Yes	Yes	Yes	Yes	Explicit
Consensus development conference	No	No	No	Yes	No	Implicit

4.3.3 The Delphi method

The Delphi method is a research technique originally developed by Norman Dalkey and Olaf Helmer for technological forecasting for the RAND Corporation (Murry and Hammons, 1995; Hasson et al., 2000). Its conceptual roots originate from the legend of the Greek Delphi oracle (Baker et al., 2006). The Delphi technique is considered by many as quasi-experimental research, aiming to bridge the gap between quantitative and qualitative approaches (Procter and Hunt, 1994). It is also a

multi-stage process where each stage builds on the results of the previous one (Sumsion, 1998). Some authors have defined this method as '... a method for structuring a group communication process so that the process is effective in allowing a group of individuals, as a whole, to deal with a complex problem' (Linstone and Turoff, 1975, p.13). Other authors have used more detailed definitions: 'a method for the systematic solicitation and collection of judgements on a particular topic through a set of carefully designed sequential questionnaires interspersed with summarised information and feedback of opinions derived from earlier responses' (Delbecq et al., 1975, p.10).

Many have described and applied Delphi studies in various situations and ways depending on the requirements of individual Delphi projects (Erffmeyer et al., 1986; Powell, 2003; Cross, 2005). This has led to the development of modified, but not always widely acceptable or rigorous, versions of the original 'classic' technique (Couper, 1984; Keeney et al., 2001). In this chapter, the 'classic' Delphi technique was selected as an initial research step for the standardisation of centralisation.

The reasoning behind using the Delphi technique was (Murry and Hammons, 1995):

(a) The topic under investigation did not lend itself entirely to precise analytical techniques and could benefit from subjective judgements on a collective basis;

(b) Time and cost would make face-to-face group meetings infeasible. In Delphi studies, there are no geographical limitations on the selection of experts (Fink et al., 1984); this may result in a significant saving of time, money and inconvenience (Walker and Selfe, 1996);

(c) The past history of disagreement amongst some special interest groups meant that open discussion was unlikely to succeed; the communication process had to be refereed and / or anonymity or 'quasi-anonymity' i.e. when respondents are known to the researcher (and sometimes to one another) but their judgements remain strictly anonymous (McKenna, 1994) ensured to prevent influential individuals from dominating the group's decision (Dalkey, 1969; Couper, 1984; Murry and Hammons, 1995).

4.3.4. Study Design

A three-round postal / email questionnaire Delphi survey was used. This decision was made *a priori* because there is clear evidence that three rounds are usually adequate to reach consensus with very little change in opinion occurring thereafter (Erffmeyer et al., 1986; Rowe et al., 1991; Murphy et al., 1998). Furthermore, although reliability is suggested to increase with the number of rounds (Fink et al., 1984), the Delphi members can become fatigued after two or three rounds. This may lead to a substantial drop-out rate compromising the validity of the study (Webler et al., 1991).

The methods in this study were based on the conceptual model that has been proposed for components of consensus development in the healthcare sector (Murphy et al., 1998). This model is summarised in Table 4.2.

Table 4.2 Components of consensus development methods (Murphy et al., 1998). Shaded cells represent the areas on which Murphy and associates (1998) focused in their review.

	Planning	**Individual judgement**	**Group interaction**
Questions	Selection of topic Selection of cues Comprehensiveness	Influence of cues Question structure Level of detail	Modification of question(s)
Participants	Number Type Degree of heterogeneity Selection of individuals	Representation of others Representation of self	Combination of backgrounds
Information provided for participants	Amount Selection Presentation	Read Understand Interpret	Use of information New information Feedback of group view
Method of structuring interaction	Choice of method Particular brief	Perceptions of process Past experience	Setting Structure of interaction
Output: method of synthesising individual judgements	Type Target audience Aggregation rules	Perceptions of output Acceptance	Production of output

4.3.5. Procedures

This study was carried out in compliance with the principles of the Helsinki Declaration (WMA, 2008) and was approved by the Biomedical Research Ethics Sub-Committee of the University of Warwick (Appendix 4.1).

4.3.6. Identification and selection of participants

An important element of the consensus decision making process is the choice of participants (Murphy et al., 1998). In qualitative research, the choice of the sampling method is determined by the methodology selected and the topic under investigation (Higginbottom, 2005). Because the Delphi method aims to reach consensus of opinion from a group of knowledgeable individuals (Murphy et al., 1998),

participants were purposefully selected to apply their knowledge and experience based on criteria developed from the research question and the aims and objectives of the study (Akins et al., 2005; Tuckett, 2005).

Purposive or purposeful sampling is a common sampling technique when the selection of research participants is made to represent the range of beliefs and experiences that the researcher thinks as relevant to the nature of the problem under investigation (Kuper et al., 2008b). Purposive sampling is defined as judgemental sampling involving the conscious selection by the researcher of certain study participants (Crooks and Davis, 1998). Murry and Hammons (1995) suggest that in order to ensure the participants' motivation and commitment to the study, purposive sampling is necessary.

Two strategies were used for the identification and selection of the Delphi panel. These strategies were selected to capture a wide range of perspectives from the panellists but at the same time to ensure 'expertise' and relevance to the research question. Diversity of panellists is hypothesised to improve the validity of findings (Mead and Moseley, 2001). Murphy and associates (1998) also concluded that participants should reflect the full range of key characteristics of the population who will use the results, in order to enhance the credibility and widespread acceptance of the results. However, care had to be taken so that the heterogeneity of the members of the Delphi group did not compromise the quality of the group or the study results by including individuals without knowledge or clinical experience of using centralisation in the panel.

The first strategy involved recruiting representatives of two specialist interest groups who have developed classifications for spinal pain that include centralisation, the McKenzie and the Delitto groups. These organised groups acted as 'gatekeepers' helping to identify potential participants. The Delitto and McKenzie groups were asked to nominate equal numbers of eligible individuals to participate in the study. Permission to obtain the contact details of nominated individuals was requested if their details were not publicly available. The second strategy involved contacting relevant individuals identified through the international peer-reviewed literature. For this strategy, a systematic search of the literature on centralisation as well as other clinically induced symptom responses (Chapters 2 and 3) was performed.

Potential participants were sent the study pack and were invited to join the Delphi group. In the study pack, an invitation letter (Appendix 4.2), a study information sheet (Appendix 4.3), the participant information and eligibility questionnaire (Appendix 4.4), two copies of a formal consent form (Appendix 4.5) and a pre-paid envelope were enclosed. Each individual had 3 weeks to decide whether or not to participate in the study and 3 weeks were also allowed for the return of questionnaires in the Delphi rounds. This quick turnaround time was selected to reduce respondent attrition by maintaining attention and motivation (Wilson et al., 2003; Hicks, 2004).

To be included, eligible individuals had to:

- Have used centralisation and the system involving it in clinical practice;
- Provide a written consent to participate in the study.

Plus one of the following:

- Have been nominated by the McKenzie or Delitto group;
- Have published or co-authored research in the peer-reviewed literature on the centralisation phenomenon and the system involving it;
- Currently participate in ongoing research on the centralisation phenomenon and the system involving it;
- Have teaching responsibilities including the centralisation phenomenon and the system involving it.

Individuals not fulfilling the aforementioned inclusion criteria were excluded.

The constitution of the expert panel is one of the most fundamental but also controversial components of the Delphi technique (Hasson et al., 2000; Cross, 2005). Definitions of the type and level of expertise needed for consensus development methods have been variable and subject to a great deal of criticism (Rothstein, 2004; Baker et al., 2006). Some authors (Parenté and Anderson-Parenté, 1987, cited in Baker et al., 2006) have even argued against the use of expert panels on the grounds that there are no guidelines on the definition of an expert or evidence that using experts increase the accuracy of Delphi studies.

The dictionary definition of an expert is a person with extensive knowledge about, or skill in a particular area (Soanes and Stevenson, 2003). However, some authors argue

that this definition may not be entirely applicable in Delphi studies in healthcare. Pill (1971) suggested that an expert should be defined as anyone with a relevant input. Mead and Moseley (2001) argue that experts can be defined in more ways, such as their position in a hierarchy, public acknowledgement or as recommended by other participants in a study. Acknowledging expertise or influence may validate the choice of participants; however, such attributes may not always be adequate for the inclusion of participants (Cross, 2005). For example, commitment to the investigation, motivation to comply with the demands of procedure and the acceptance of the consensus may be fundamental (Cross, 2005).

In Delphi studies, three key themes are reported to emerge from the definition of expertise: knowledge, experience and ability to influence policy (Baker et al., 2006). Knowledge in this study was demonstrated by the possession of a professional qualification, although having a higher degree in a specific area may increase the credibility of an expert (Baker et al., 2006). The authoring of materials such as books or peer-reviewed articles and participation in ongoing research and professional education was also used as an eligibility criterion. This was done in the light of suggestions that this increases the likelihood of including participants with specific and cutting-edge knowledge in an area (Duncan et al., 2004). Content validity increases with the use of participants who have knowledge and interest in the topic (Goodman, 1987). In addition, the ability of such individuals to influence policy may have implications on the ability to implement findings or do research, an area in which Delphi studies have also been reported to be inadequate (Fink et al., 1984). Murphy and associates (1998) suggest that a selected group of individuals is more likely to lend some authority to the decision produced. Another consideration when

defining criteria for selection of an expert panel in Delphi studies is that those participating may possess knowledge but not clinical experience (Baker et al., 2006; Delitto, 1998). For this reason, participants were required to have clinical experience of using centralisation in spinal practice.

The inclusion of patients and potential research participants has also been recommended (Streiner and Norman, 2003). However, service users such as spinal pain groups were not invited to take part in this Delphi panel or provide feedback on the questionnaire development stage. This decision was made because the inclusion of service users to an expert panel dealing with technical information or expert opinion based on such knowledge prerequisites does not necessarily add additional validity to this study method (Baker et al., 2006). The use of alternative methods, such as ratifying the findings, triangulating the results or using different methodologies (Baker et al., 2006) seemed more appropriate when taking account of people with spinal problems.

4.3.7. Sample size

Qualitative studies do not usually follow a predetermined sample size (Kuper et al., 2008a). This is because most qualitative research does not aim to generalise findings in the same way as quantitative research (Greenhalgh, 1997). Instead, qualitative studies aim to provide information-rich data on a chosen topic (Higginbottom, 2005). In the case of the Delphi method, as Helmer (1977, p.18-19) points out "a Delphi inquiry is not an opinion poll, relying on drawing a random sample from 'the population of experts'; rather, once a set of experts has been selected (regardless of how), it provides a communication device for them, that uses the conductor of the

exercise as a filter in order to preserve anonymity of responses". Thus, the representativeness of the panel was judged on the qualities of the expert panel rather than its numbers (Powell, 2003).

There are no straightforward rules for the estimation of the sample size in the Delphi method or consensus development studies (Murphy et al., 1998; Akins et al., 2005). Some authors argue that sample size should be guided by the degree of homogeneity of the participating group and the nature of the investigation (Murphy et al., 1998). Others suggest that the number of participants should vary according to the requirements of the particular technique used, the scope of the problem and the available resources (Fink et al., 1984).

The impact of the number of participants on the reliability or validity of consensus studies lacks empirical evidence, with effects due to size being subtle and difficult to detect (Murphy et al., 1998). As a result, sample sizes in musculoskeletal research have ranged from 24 (Binkley et al., 1993), 30 (McCarthy et al., 2006), 45 (Miro et al., 2008) to several hundred individuals (Couper, 1984). Groups beyond 30 may not result in more information and such a large sample may only increase administration difficulties and cost (Delbecq et al., 1975; Fink et al., 1984; Cross, 2005). Because the Delphi approach also uses quantitative analysis methods to aggregate results, it was decided *a priori* that the number of participants should not be less than six, so that the reliability of the group decision making does not decrease dramatically (Murphy et al., 1998).

4.3.8. Pilot testing

The structure and organisation of the first round may influence subsequent responses (Procter and Hunt, 1994). Questionnaire design can be open to many researcher biases and errors in respondent judgements (Mead and Moseley, 2001). If a question is leading or ambiguous, the interpretation of responses is also difficult or 'unsafe' (Mead and Moseley, 2001).

Before sending the first-round questionnaire, pilot testing and advice was sought from five individuals who participated in the actual Delphi study. This strategy was followed to identify ambiguities and improve the feasibility of questionnaire administration (Jairath and Weinstein, 1994). Apart from receiving the first-round questionnaire, these five individuals also received the participant information leaflet and the eligibility questionnaire and were asked to make comments regarding format, clarity, content, and wording. Following their suggestions, the final version of the first-round questionnaire was sent to the consenting participants.

4.3.9. Round 1

Consenting participants in the first round were asked to make suggestions for categories and questionnaire items relevant to the purpose of the study. The following questions were asked in the first-round questionnaire (Appendix 4.6):

Question 1

'*Please list the criteria that should be used for your preferred operational definition of centralization as a physical sign and related symptom response groups.'*

Question 2

'Please write any further issues that do not fit into the context of the above question but are important when considering centralization in spinal clinical practice and related research.'

Question 3

Please list any important questions around centralization that you would like to be addressed in future research.'

Open-ended questions were used to give the respondents as much scope as possible in generating relevant items (Procter and Hunt, 1994). Asking participants to suggest items may promote their commitment in participation and help them justify their judgements (Murphy et al., 1998). Open-ended questions can generate a vast amount of unusable data; however, because the research question was unlikely to produce an unmanageable number of definitions, this strategy was not expected to cause significant delays and lack of focus on the study issues (Binkley et al., 1993; Murry and Hammons, 1995; Murphy et al., 1998; Mead and Moseley, 2001).

A literature review was not used to produce items for the first round because potentially useful information could have been lost by limiting the number of suggestions (Mead and Moseley, 2001). Participants were also not provided with literature reviews on the topic despite recommendations for a review of research-based information to all participants at an early stage (Murphy et al., 1998). This decision was based on the grounds that one of the prerequisites for inclusion as experts was knowledge in the field.

In each round, a reminder (email, letter or phone call) was sent to individuals who had not returned their completed questionnaires within 3 weeks. Responses from the

first-round questionnaire were subject to content analysis (Krippendorff, 1980) by two members of the research team (Angeliki Chorti and Chris McCarthy). One first-round questionnaire did not arrive in time and was not included in the analysis; however, no new issues arose from this questionnaire.

Content analysis has been described as a research technique that 'provides systematic and objective means in order to describe and quantify phenomena' (Downe-Wamboldt, 1992). A thematic frequency analysis was performed by the author of this thesis with accuracy of themes and statements checked by Chris McCarthy, who is familiar with the area under investigation. In order to utilise the Delphi method in this study it was also necessary to develop a theoretical framework for the operational definition / criteria and suggestions for future research (Procter and Hunt, 1994). The Standards for Tests and Measurements (Rothstein et al., 1991) were used to inform the questionnaire development and the collation of responses.

When analysing responses from the first round, not only general definitions were collated for the second round but also specific operational criteria and future research questions. The aim of this approach was to avoid ambiguities in the resulting guidance and recommendations. Specifically stated consensus findings are more amenable to action and less likely to be subjected to misinterpretation than are generally stated findings (Fink et al., 1984). This is particularly important since one of the criticisms of the Delphi technique is that it may result in statements of broad generalities that are not always helpful in making clinical decisions (Rennie, 1981; Frances et al., 1998).

All participants' responses were included in the analysis of the first round results including the views of 'outliers'. This decision was made to allow the Delphi members to express and judge their opinions without any possible investigator influence (Hasson et al., 2000). Minimal editing was applied to the statements presented for consideration in the second round for the same reason. Hitch and Mugatroyd (1983) have also described a similar attempt to avoid manipulation of the data given by the respondents in order to preserve its authenticity.

One reported disadvantage of including all options suggested by participants is that judgements of scenarios which never or rarely occur in practice may be less reliable (Murphy et al., 1998). The decision to include outliers may compromise the level of reported agreement (Black, 1994) by alienating participants from the task at hand (Murphy et al., 1998). For this reason, although all suggestions on general definitions were presented to participants in the first round, a threshold for including items had to be applied in subsequent rounds.

4.3.10. Round 2

In Delphi studies, the second round is commonly used by the researcher to ask the participants to consider, rate, edit and comment upon items on the developed questionnaire from the previous round (Procter and Hunt, 1994; Murry and Hammons, 1995). In this study, participants were asked to vote for their preferred definition by indicating their personal level of support on a 5-point Likert scale ('strongly agree' to 'strongly disagree') for all general definitions as well as

component criteria in the second-round questionnaire (Appendix 4.7); the same ranking also applied to the list of clinical and research-related issues.

Likert scales are the most common tools for rating Delphi questionnaire criteria and quantifying experts' views (Murry and Hammons, 1995; Mead and Moseley, 2001). The number of points in a Likert scale usually ranges from 3 to 11 (Mead and Moseley, 2001). The minimum recommended number of categories is in the region of 5 to 7 (Streiner and Norman, 2003). The five-point scale was selected here because the larger the number of options, the more time it takes to complete and the more difficult it is to hold items in short-term memory (Baddeley, 1994; Miller, 1994).

Participants were encouraged to give reasons for their answers; this was done in an attempt to promote interaction and discussion among the group members. Participants were also allowed to add items or make comments that they considered important and were not included in the questionnaire. This strategy helped confirm whether an adequate analysis of the original data had been undertaken (Procter and Hunt, 1994). Delphi studies are often criticised in terms of exchange of information over the nominal group technique (Murphy et al., 1998; Cross, 2005).

Following the results of the second round, a feedback report (Appendix 4.8) and the third-round questionnaire (Appendix 4.9) were designed. The feedback report included the summary statistics of the group response and the participant's responses for each questionnaire item, all comments made by respondents as well as graphical

displays of the group response for each statement. This would enable participants to see where their response stood in relation to that of the group (Couper, 1984).

Statistical analyses were performed using SPSS version 14 (Chicago, ILL). Total agreement within the group in the second round was investigated by using the Kendall coefficient of concordance (Hicks, 2004; Grzegorzewski, 2006). The best method of mathematical aggregation is not yet established (Murphy et al., 1998). However, the median and interquartile ranges (25th and 75th percentiles) were used to express the level of support for each statement because they are considered more robust than the mean for describing group agreement (Murphy et al., 1998), as response patterns tend to be bimodal (Binkley et al., 1993). Percent agreement was also calculated for each item.

The level or the type of agreement for establishing consensus has not been consistent in the Delphi methodology literature (Fink et al., 1984; Murry and Hammons, 1995; Walker and Selfe, 1996; Wilson et al., 2003). Some suggest that the level used is at the researchers' discretion and depends upon sample numbers, the aims of the research and responses (Wilson et al., 2003). Nevertheless, consensus parameters need to be determined in advance (Fink et al., 1984).

Many types of criteria have been used to describe when consensus is achieved (Murphy et al., 1998). Fink and associates (1984) have described some, presented below:

- On the final vote, any topic or issue supported by a pre-determined proportion of participants is adopted;

- After a predetermined number of rounds of voting, a number of topics receiving the most votes are approved;
- All topics are rated on a scale of 1 to 5. Only those topics receiving a mean rating of 3 or more are accepted;
- All topics are rated on a scale of 1 to 3. All topics receiving a rating of 1 from the majority of participants are adopted;
- Any topic is dropped if it is vigorously opposed by a pre-determined proportion of the participants.

Thresholds for consensus levels have been reported to range from 51% to 80% (Hasson et al., 2000). The stricter the criteria, the more difficult it is usually to establish consensus (Fink et al., 1984; Murphy et al., 1998). A small Delphi survey of practising orthopaedic clinicians on low back pain classifications indicated that agreement should reach at least 75% of opinions to have an impact on clinical practice (Binkley et al., 1993). This is in keeping with others (Murry and Hammons, 1995) who propose the 75% cut-off as a minimum point of agreement for reaching consensus. Based on the above considerations, it was determined *a priori* that items reaching 80% of agreement should be included, whereas items below 70% should be excluded from the list of statements of the third-round questionnaire. Items between these margins were sent for reconsideration in the next round.

4.3.11. Round 3

The goal of the third round is to achieve consensus or stability of the panel member responses (Murry and Hammons, 1995). In the final round of the Delphi study,

participants are asked to reconsider their views in the light of other Delphi members' opinions (Couper, 1984; Murry and Hammons, 1995). Thus, the Delphi panellists were presented with the feedback report from round 2 (Appendix 4.8) and the third-round questionnaire (Appendix 4.9) and were asked to answer whether they agreed or disagreed with each statement using a 'yes / no' answer format in the light of other panellists' opinions. This limited range of response options can decrease the scope for fine judgements; however, this format may also put more pressure on the participants to make a decision (Mead and Moseley, 2001).

The results of the third round were fed back to the participants (Appendix 4.10) who then had the opportunity to make their final comments. To promote the dissemination of findings, the results of the Delphi study were also presented to professional conferences and will be submitted for publication in international peer-reviewed journals.

4.4. RESULTS

4.4.1. The Delphi panel

The study pack was sent to 72 individuals. Thirty individuals did not return the forms to the research team. Twenty-six did not respond at all, whereas 4 individuals emailed to decline participation because of time commitments or lack of relevance to the topic under investigation. From the 42 forms that were received by the research team, 7 envelopes were returned as undeliverable, leaving 36 individuals returning completed consent forms to participate in the Delphi study. However, one participant did not agree to complete the eligibility questionnaire and was excluded from the

study. Therefore, thirty-five individuals, from which 24 were males and 11 females, aged 32 to 65 years (mean = 48, SD = 8.4), agreed to participate in the Delphi study and were sent the first-round questionnaire. Participants worked in various settings, including clinical practice, research and education. Most individuals were physical therapists (77%) who were currently in clinical practice (71%) and had a MSc or PhD (66%) qualification. The years of experience in using centralisation ranged from 3 to 31 years (mean = 15, SD = 7.6).

The response rate for the first, second and the third round was 89% (31 / 35), 89% (30 / 35) and 89% (31 / 35) respectively. These rates are high (well above the minimum recommended response rate of 70%) and acceptable for maintaining rigour in this technique (Procter and Hunt, 1994; Walker and Selfe, 1996; Hasson et al., 2000). Two participants did not return any of the questionnaires. Reasons for not returning questionnaires are presented in Figure 4.1. The characteristics of the Delphi group over the 3 rounds are presented in Table 4.3.

4.4.2. The Delphi list of items

Fourteen general definitions (Table 4.4) were identified from the analysis of the first-round questionnaire. Issues arising from the criteria for the operational definition referred to the population for whom the test is intended, potential test users, tools used for documentation, type of the loading strategy (including planes / directions and characteristics of movement testing), criteria for a positive test, timeframe of response and safety issues.

Other issues related to centralisation reflected opinions about the prognostic and diagnostic implications of the test. Future research suggestions included the establishment of the operational definition of centralisation, a conceptual model for centralisation, the diagnostic accuracy of centralisation, the prevalence, reproducibility, course and prescriptive validity of centralisation, centralisation and the cervical spine, case studies, and finally education and training. After the second round, nineteen new items were added to the list of third-round statements. These items mainly reflected clarifications or refinements of pre-existing items from the second round questionnaire.

The Delphi survey was successful in establishing a definition and operational criteria, as well as future research questions for centralisation after the third round. Tables 4.5 and 4.6 present included and excluded items after the final round of the Delphi study.

Figure 4.1 Delphi study flowchart.

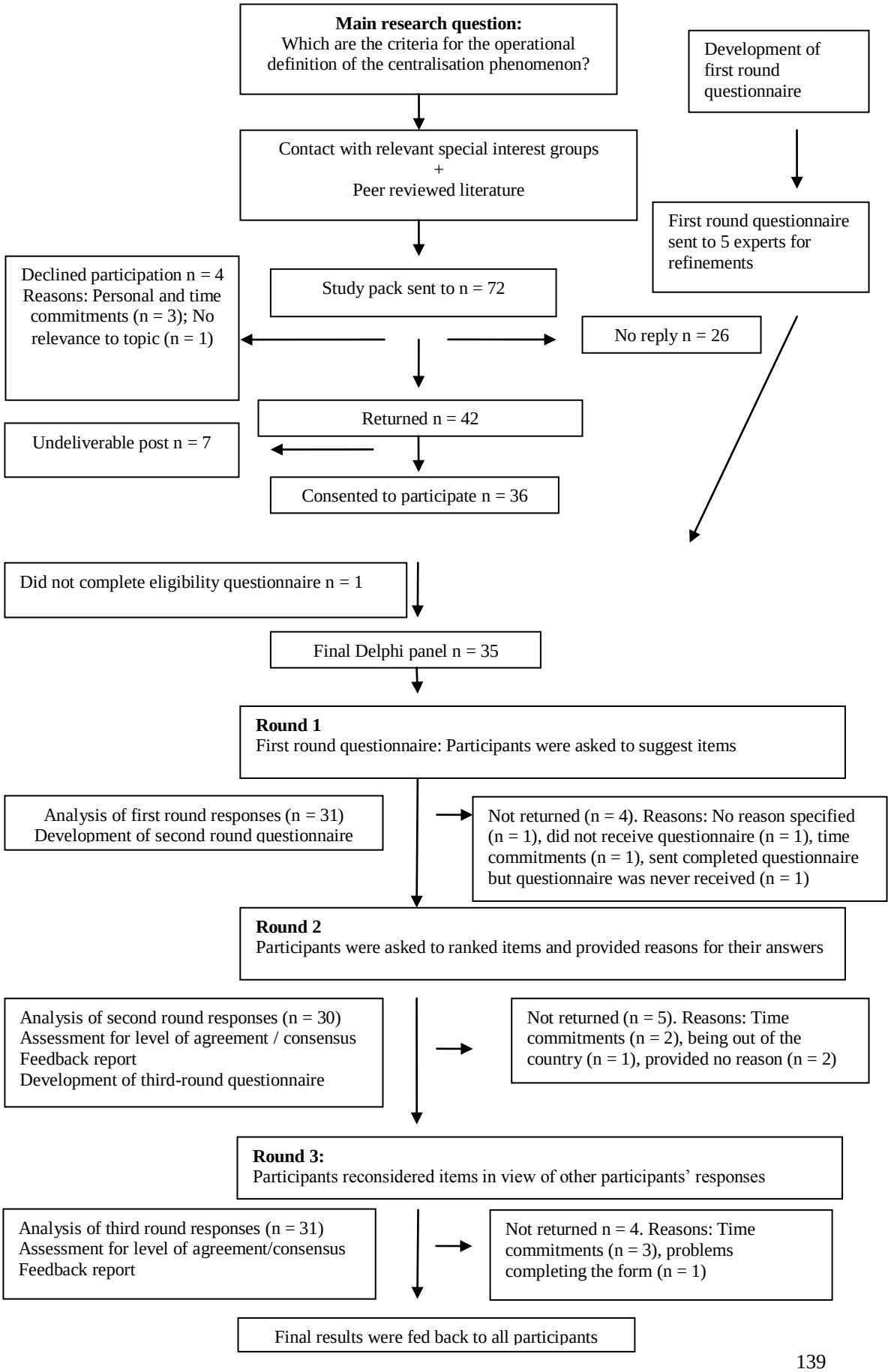


Table 4.3 Characteristics of the Delphi group in the three rounds*.

Characteristic		**Consenting participants (n=35)**	**Participants round 1 (n=31)**	**Participants round 2 (n=30)**	**Participants round 3 (n=31)**
Gender	Male	24 (69)	23 (74)	21 (70)	23 (74)
	Female	11 (31)	8 (26)	9 (30)	8 (26)
Continent	North America	23 (66)	21 (68)	21 (70)	22 (71)
	Europe	9 (26)	7 (23)	6 (20)	6 (19)
	Australasia	3 (9)	3 (10)	3 (10)	3 (10)
Basic Qualification	PT	27 (77)	24 (77)	23 (77)	24 (77)
	MD	3 (9)	3 (10)	3 (10)	2 (7)
	Other	3 (9)	3 (10)	3 (10)	3 (10)
Highest qualification	MS	13 (37)	13 (42)	12 (40)	13 (42)
	PhD	10 (29)	9 (29)	8 (27)	8 (26)
	PG Dip	5 (14)	4 (13)	4 (13)	4 (13)
	Professional membership	3 (9)	2 (7)	3 (10)	2 (7)
System predominantly using	McKenzie	25 (71)	22 (71)	21 (70)	22 (71)
	Delitto	8 (23)	8 (26)	8 (27)	7 (23)
	Other	1 (3)	1 (3)	1 (3)	1 (3)
Previous relevant publications	Yes	33 (94)	29 (94)	28 (93)	29 (94)
Participation in ongoing relevant research	Yes	17 (49)	15 (48)	16 (53)	16 (52)
Previous or current relevant educational responsibilities	Yes	21 (60)	18 (58)	17 (57)	18 (58)

*Expressed as number of participants (percentage of participants in parentheses). Percentages have been rounded.

Abbreviations:

MD, medical doctor; MS, Masters degree; PG Dip, Postgraduate Diploma PhD, Doctor of Philosophy; PT, physiotherapist.

Table 4.4 Reported general definitions presented in the second Delphi round questionnaire.

Section 1: General definition

*Please **mark one box** for each statement that is closest to how you feel.*

(1)= Strongly Agree (2) = Agree (3) = Neutral (4) = Disagree (5) = Strongly Disagree

Centralization should generally be defined as:

	(1)	(2)	(3)	(4)	(5)
1. A lasting improvement in patient status (abolishment of distal symptoms or improvement of signs) in response to a defined movement which can vary in terms of direction, degree, duration, repetition and position used	☐	☐	☐	☐	☐
2. A lasting abolishment or decrease in intensity of the most distal radicular symptoms and signs in response to repeated movements or static positioning, traction or a combination (traction and repeated movements).	☐	☐	☐	☐	☐
3. An abolishment of the most distal radiating symptoms in response to repeated movement testing.	☐	☐	☐	☐	☐
4. An improvement in location, intensity or frequency of symptoms in response to single or repeated movement testing or sustained postures	☐	☐	☐	☐	☐
5. An abolishment of peripheral symptoms in response to repeated movement testing and overpressure	☐	☐	☐	☐	☐
6. A reduction or abolishment of peripheral symptoms in response to repeated movement testing and patient or therapist overpressure	☐	☐	☐	☐	☐
7. The movement of symptoms in a proximal direction in response to repeated end-range movement testing only	☐	☐	☐	☐	☐
8. Distal symptoms moving proximally in response to repeated movement testing and/or sustained positions or therapist mobilization.	☐	☐	☐	☐	☐
9. Distal symptoms moving and remaining proximally in response to repeated movement testing and/or sustained positions or therapist mobilization. In patients with axial symptoms only, the decrease in such symptoms is defined as 'centralization' if such decrease is (a) substantial (>50% reduction) and (b) sustained	☐	☐	☐	☐	☐
10. Distal symptoms (pain, numbness or tingling) traveling proximally towards the central spine in response to therapeutic loading strategies	☐	☐	☐	☐	☐
11. An improvement of the most distal symptom regardless of the testing methods used (i.e. during movement testing, treatment, over time). In order of priority, the type (i.e. pain, paresthesia or anesthesia), location and intensity of symptom are considered in the hierarchy of improvement in the definition.	☐	☐	☐	☐	☐
12. Movement of pain only to a proximal location in response to movement testing	☐	☐	☐	☐	☐
13. The most distal pain disappearing and moving proximally in response to repeated end-range movements or static loading	☐	☐	☐	☐	☐
14. The progressive and stable reduction of the most distal pain towards the spinal midline in response to standardized repeated end-range movement or sustained loading testing procedures. Testing may involve multiple directions and various starting positions, progression of forces or alternative forces (e.g. clinician assistance) as well as more than one visit	☐	☐	☐	☐	☐
Please feel free to provide reasons for your choices and/ or add any important comments/ questions here:					

Table 4.5 Included items after the third round of the Delphi study.

1. General definition

Centralisation should generally be defined as:

- The progressive and stable reduction of the most distal pain towards the spinal midline in response to standardised repeated end-range movement or sustained loading testing procedures. Testing may involve multiple directions and various starting positions, progression of forces or alternative forces (e.g. clinician assistance) as well as more than one visit.

2. Population for whom the test is intended

- Centralisation can be best appreciated only in patients who demonstrate referred or radiating symptoms originating from the spine
- Before testing, patients should be screened for the presence of red flags (i.e. serious pathology indicators).

3. Potential test users

- Centralisation should be recognized by the average clinician rather than requiring years of training
- Training of examiners is essential in order to have consistent results among clinicians

4. Tools used for documentation

No items were included in the consensus

5. Loading strategy / Testing

- Centralisation can be elicited by repeated and/or sustained end-range movement testing, overpressure by patient or therapist or manual techniques executed by therapist if appropriate
- Testing for centralisation should involve the standard planes of movement available to the spinal region and/or a combination of movements if appropriate
- Movement testing for centralisation should not be based solely on single movements. Test movement must be performed repeatedly, to the fullest of the patient's available end-range
- Movement testing for centralisation should include repetitions that are continued long enough to ensure that the status of the symptoms has changed and/or is clear.

6. Criteria for a positive test

- Nonspinal conditions or underlying disease states potentially causing or influencing symptoms should be excluded or taken into consideration
- Observed changes should be retained over time

7. Timeframe

- The response to testing may be obvious during the first examination or may require evaluation over a period of time to confirm the phenomenon

8. Safety issues

- Symptoms getting worse proximally but improving distally should be considered a positive sign
- Education of patients is essential following the use of these techniques so that movement is normalized as symptoms settle and patients do not develop fear of movement

9. Attributes related to centralisation

- People with centralisation have a good prognosis

10. Future research

- Cross validation of diagnostic accuracy in different patient populations and examiners
- Expected rates of centralisation in clinical settings for acute and chronic patients with low back pain
- Effect of training and procedures in the prevalence and outcomes of centralizers
- Do centralizers have a favourable course when left untreated (natural history)?
- Clinical response versus natural course of centralizers
- Do centralizers benefit from interventions other than directional movement exercise and which intervention provides the greatest benefit?
- Comparison of outcomes of non-responders to centralisation for various interventions
- Centralisation and outcomes
- Clinical predictors (CPR) of patients responding (or not) to directional preference exercises (including: Effect of patient compliance and attitudes of centralizers on outcomes; Effect of severe disability and psychosocial distress on the diagnostic, prognostic and treatment characteristics of centralizers; Centralisation versus other prognostic factors)
- Are centralizers and other clinical subgroups (e.g. patients satisfying CPR for manipulation or trunk stabilization treatment or candidates for disc surgery) made up of the same patients? How often do centralisation findings co-exist with other findings? How soon will positive electromyographic findings become normal after centralisation has been achieved and maintained?
- Reliability of detecting, prevalence and outcomes in neck pain
- Predictive validity of centralisation in the management of neck pain
- Effect of standardization of the test on the reliability of centralisation
- Role of history in predicting the presence of centralisation
- Role of examiner's training in predicting the presence of centralisation

Table 4.6 Excluded items after the third round of the Delphi study.

1. General definition

Centralisation should generally be defined as:

- A lasting improvement in patient status (abolition of distal symptoms or improvement of signs) in response to a defined movement which can vary in terms of direction, degree, duration, repetition and position used
- A lasting abolishment or decrease in intensity of the most distal radicular symptoms and signs in response to repeated movements or static positioning, traction or a combination (traction and repeated movements).
- An abolishment of the most distal radiating symptoms in response to repeated movement testing.
- An improvement in location, intensity or frequency of symptoms in response to single or repeated movement testing or sustained postures
- An abolishment of peripheral symptoms in response to repeated movement testing and overpressure
- A reduction or abolishment of peripheral symptoms in response to repeated movement testing and patient or therapist overpressure
- The movement of symptoms in a proximal direction in response to repeated end-range movement testing only
- Distal symptoms moving proximally in response to repeated movement testing and/or sustained positions or therapist mobilization.
- Distal symptoms moving and remaining proximally in response to repeated movement testing and/or sustained positions or therapist mobilization. In patients with axial symptoms only, the decrease in such symptoms is defined as ‘centralisation’ if such decrease is (a) substantial (>50% reduction) and (b) sustained
- Distal symptoms (pain, numbness or tingling) traveling proximally towards the central spine in response to therapeutic loading strategies
- An improvement of the most distal symptom regardless of the testing methods used (i.e. during movement testing, treatment, over time). In order of priority, the type (i.e. pain, paresthesia or anesthesia), location and intensity of symptom are considered in the hierarchy of improvement in the definition.
- Movement of pain only to a proximal location in response to movement testing
- The most distal pain disappearing and moving proximally in response to repeated end-range movements or static loading

2. Population for whom the test is intended

- Centralisation can be best appreciated only in patients with low back pain i.e. and not patients with symptoms originating from other spinal areas
- Before testing, patients should be screened for the presence of yellow flags (e.g. Waddell signs)
- Patients with distal symptoms above the knee that meet other parts of a clinical prediction rule for success with spinal manipulation are most likely to benefit from manipulation regardless of their response to repeated movement testing
- The centralisation phenomenon becomes complex in chronic cases with an adherent nerve root involved. In such cases, the close monitoring of the symptoms and potential postural deformities is required and patients are treated on an individual basis

3. Potential test users

- The experience of examiners is essential in order to have consistent results among clinicians

4. Tools used for documentation

- The pattern of questioning and documenting patient responses i.e. pain intensity and location should be highly standardised in clinical practice and research
- The pattern of questioning and documenting patient responses i.e. pain intensity and location should be highly standardised in clinical practice only
- The pattern of questioning and documenting patient responses i.e. pain intensity and location should be highly standardised in research only
- For changes in pain location, a clear overlay numeric template should be used
- The area of the most distal symptoms should be marked and reassessed using a measurement from a bony landmark. The determination of the most distal pain should also be confirmed through palpation

5. Loading strategy/ Testing

Centralisation can be elicited by:

- Repeated and/ or sustained end-range movement testing only
- By single and repeated movements. Sustained movements are used with caution to prevent confusion with instability
- By repeated and/or sustained end-range movement testing, or manual techniques executed by therapist if appropriate
- Regardless of testing methods used. This can involve test movements or after immediate treatment application, or post treatment over time
- Testing for centralisation should involve only the sagittal plane
- Testing for centralisation should involve the standard planes of movement available to the spinal region involved
- Movement testing for centralisation should not include more than 10 repetitions for each test movement
- Movement testing for centralisation should not necessarily be performed in a specific order. The sequence of movement testing does not affect the outcome or the ability to detect centralisation

Cont'

6. Criteria for positive test

- When defining the presence of centralisation, changes in the intensity of symptoms should not be considered
- When defining the presence of centralisation, changes in the neurological status (symptoms or signs) should not be considered
- When defining the presence of centralisation, underlying disease states potentially causing or influencing symptoms should be excluded
- The reduction of peripheral symptoms towards the spinal midline should be progressive and stable. Distal pain which simply abolishes rather than progressively receding toward the spine should not be considered in the definition

7. Timeframe

- Symptoms may be sequentially and lastingly abolished at each session (pure response) or gradually abolished in a progressive manner but not sequentially (partial response)

8. Safety issues

- The patient's status (improving or worsening) should be considered beyond symptom relief i.e. neurological picture, range of motion. For example, if the patient's symptoms are improving but his neurological status is compromised, centralisation should be considered an undesirable outcome for this patient.

9. Attributes related to centralisation

- Centralisation may be a stronger prognostic factor than psychosocial variables
- Centralizers have an internal disc disruption

10. Future research

- Head to head comparisons of different operational definitions
- Clarification of the term centralisation i.e. number of repetitions, type of change, magnitude of change (e.g. belt line pain), duration and timeframe of required change, necessity of provocative testing for central symptoms
- Is centralisation an anatomical phenomenon? Centralisation in relation to tissue response
- Mechanism causing the centralisation phenomenon (including physiological mechanism)
- Criterion validity of centralisation (e.g. using provocation discography as a standard; centralisation as a tool for assessing the severity of a disc lesion; centralisation as a tool for ruling out pathologies other than discogenic pain)
- Potential irreducible derangement i.e. pain moving centrally but neurological symptoms moving distally: at what point should patients be referred for further investigation?
- Are centralizers and other clinical subgroups (e.g. patients satisfying CPR for manipulation or trunk stabilization treatment or candidates for disc surgery) made up of the same patients?
- How often do centralisation findings co-exist with other findings? (i.e. relationship of centralisation with other variables e.g. psychosocial, clinical findings e.g. segmental provocation signs, sciatica and electromyographic findings)
- Centralisation and contained cervical pathology
- Stability/ reversibility of the centralisation phenomenon
- Effect of the clinician characteristics on the reliability of centralisation
- Effect of the patient characteristics on the reliability of centralisation
- Effect of knowledge of the test on the reliability of centralisation
- Case study: The rare patient whose pain centralizes up to the lumbar spine but then remains unchanged and may worsen with exercise. Generally improves over time and is usually initiated by long sitting time
- How do we best educate healthcare professionals that abolishment of leg or arm symptoms in a patient with central or foraminal stenosis is not a "centralizer", but a separate subgroup of their own?

4.4.2.1 Operational definition

General definition: Centralisation was agreed by 80% of the participants to be defined as the progressive and stable reduction of the most distal pain towards the spinal midline in response to standardised assessment procedures (Figure 4.2).

Figure 4.2 Section on the general definition from the third round feedback report Note: N = number of valid responses; % agree = percentage of participants that agreed with statement, green ≥ 80% agreement.

Section 1: General definition

Centralization should generally be defined as:

	N	% agree
1. A lasting improvement in patient status (abolition of distal symptoms or improvement of signs) in response to a defined movement which can vary in terms of direction, degree, duration, repetition and position used	**30**	**63.3**
14. The progressive and stable reduction of the most distal pain towards the spinal midline in response to standardized repeated end-range movement or sustained loading testing procedures. Testing may involve multiple directions and various starting positions, progression of forces or alternative forces (e.g. clinician assistance) as well as more than one visit	**30**	**80.0**

COMMENTS
Statement 1: Does not include other loading strategies; The general definition should include the movement of symptoms toward the midline; Signs should not be considered in the definition
Statement 14: Too focused; More comprehensive than other definitions; Too flexible; Does not reflect standardized examination; Problematic when referring to central pain only; Potential confusion with natural course of symptoms or non-specific improvement due to time element; Stability and lasting change should be defined better.

Population to whom the test is intended (Figure 4.3): Screening for serious pathology was considered essential before testing for this type of symptom response. It was also acknowledged that some conditions can potentially have an impact on symptoms and for this reason they should be taken into account in the examination. However, screening for psychosocial factors i.e. yellow flags was not deemed necessary by most participants. Centralisation was considered to be best appreciated in patients who demonstrate referred or radiating symptoms originating from the spine. The location of symptoms below or above the knee did not seem to make a difference to

participants as to the type of intervention that low back sufferers who centralise should receive. Some participants were concerned about the definition of centralisation in relation to central pain only. Their concerns mainly involved difficulties in determining changes in location in small surface areas or other spinal areas such as the thoracic spine.

Figure 4.3 Section on the patient population from the third round feedback report Note: N = number of valid responses; % agree = percentage of participants that agreed with statement, green ≥ 80% agreement.

A. Population for whom the test is intended

	N	% agree
15. Centralization can be best appreciated only in patients who demonstrate referred or radiating symptoms originating from the spine	30	86.7
16. Centralization can be best appreciated only in patients with low back pain i.e. and not patients with symptoms originating from other spinal areas	29	0.00
17. Before testing, patients should be screened for the presence of red flags (i.e. serious pathology indicators)	30	96.7
18. Before testing, patients should be screened for the presence of yellow flags (e.g. Waddell signs)	31	32.3
19. Patients with distal symptoms above the knee that meet other parts of a clinical prediction rule for success with spinal manipulation are most likely to benefit from manipulation regardless of their response to repeated movement testing	31	22.6
20. The centralization phenomenon becomes complex in chronic cases with an adherent nerve root involved. In such cases, the close monitoring of the symptoms and potential postural deformities is required and patients are treated on an individual basis	30	36.7

COMMENTS
Statement 15: Best and only contradict each other
Statement 16: Leaves out other spinal areas
Statement 17: The dangers associated with the testing procedure are minimal; Patients with red flags should not be excluded from the testing unless they can not perform the movements; If a patient is not appropriate for mechanical therapy, he/she will not centralise.
Statement 18: Psychosocial factors can have an impact on musculoskeletal pain and may confound the clinical interpretation of physical findings; Prefer FABQ to Waddell signs; Research has shown that physical therapists do not fully understand how to integrate and evaluate the psychosocial domains with physical domain during the clinical examination of patients with low back pain
Statement 19: Childs et al. Ann Int. Med, 2004 have provided evidence for this statement and also good evidence that patients with symptoms distal to the knee are not as likely to respond to manipulation; In a study by Browder et al. PT 2007, the treatment effect was dramatic in comparison to an extension-oriented treatment approach (EOTA); The manipulation CPR has not been thoroughly tested and we know of one published example that shows manipulation to not be the optimal treatment for a patient who fits the manipulation CPR characteristics; Any benefit from manipulation would be overshadowed by the benefit of teaching self-care to centralize and abolish symptoms as well as empowerment for prevention of recurrences using directional exercises and posture modifications; This statement is more relevant to the McKenzie method
Statement 20: Close monitoring and individualized treatment would apply to all patients anyway; No reliable/valid method for 'diagnosing' an adherent nerve root; If the patient has an adherent nerve root, there is no centralization

Potential test users (Figure 4.4): Participating experts agreed that centralisation should be a recognisable sign by the average clinician. Some participants raised the issue of what 'average' means and whether sufficient education is provided in entry level programmes in order to be able to use this sign. Training of examiners was considered essential in order to achieve standardisation and to have consistent results among clinicians. However, the experience of clinicians did not result in the same certainty as training.

Figure 4.4 Section on the clinician population (potential test users) from the third round feedback report Note: N = number of valid responses; % agree = percentage of participants that agreed with statement, green ≥ 80% agreement.

B. Potential test users (examiners)

	N	% agree
21. Centralization should be recognized by the average clinician rather than requiring years of training	30	80.0
22. Training and experience of examiners is essential in order to have consistent results among clinicians	30	56.7
22.1. Training of examiners is essential in order to have consistent results among clinicians	30	86.7
22.2. The experience of examiners is essential in order to have consistent results among clinicians	29	44.8

COMMENTS
Statement 21: To be useful and widely used it should be recognized by average and entry-level clinicians; We do not have a definition of what "average" is. More research is needed to clarify the issues around training and experience; This statement depends on whether clinicians have sufficient education in entry level physiotherapy programs; Post-graduate training may be required
Statements 22, 22.1, 22.2: Standardization (e.g. specificity and clarity of definitions) is required to achieve the same shared baseline knowledge of procedures; Training is important for consistency, experience is not; Good education and learning experiences are more important than years of experience; Experience is desirable and improves matters, but we all start out as novices; Prior research has demonstrated that appropriate training is needed in order to have consistency and reliability; Consistency of results might depend on the quality, level and amount of training and experience; Current research does support training but there is no research on the level of experience; Training is required, and should be at undergraduate level. All physiotherapists should know how to do the test procedures and interpret the patient's responses.

Tools used for documentation (Figure 4.5): Many participants (76.7%) agreed that the pattern of questioning and documenting patient responses should be highly standardised in clinical practice and research. However, this statement did not reach the acceptable cut-off point of 80%. The use of palpation, the overlay template for

changes in pain location or a bony landmark as a reference point for the area of the most distal symptoms were not supported by the majority of participants.

Figure 4.5 Section on the tools used for documentation from the third round feedback report Note: N = number of valid responses; % agree = percentage of participants that agreed with statement, 70% ≤yellow ≤ 79% agreement.

C. Tools used for documentation

	N	% agree
23.The pattern of questioning and documenting patient responses i.e. pain intensity and location should be highly standardised	30	73.3
23.1. The pattern of questioning and documenting patient responses i.e. pain intensity and location should be highly standardised in clinical practice and research	30	76.7
23.2. The pattern of questioning and documenting patient responses i.e. pain intensity and location should be highly standardised in clinical practice only	30	3.3
23.3. The pattern of questioning and documenting patient responses i.e. pain intensity and location should be highly standardised in research only	30	20.0
24. For changes in pain location, a clear overlay numeric template should be used	31	16.1
25. The area of the most distal symptoms should be marked and reassessed using a measurement from a bony landmark. The determination of the most distal pain should also be confirmed through palpation	31	0.00

COMMENTS
Statements 23, 23.1, 23.2, 23.3: The word "highly" should be defined or excluded; Standardisation may help, but we do not know. It has not been studied; Standardization does not necessarily have to be prescriptive. It may just imply consistency and clarity of reporting method; Failure to recognise the pain distribution and intensity prior to test manoeuvres is common among clinicians, therefore standardisation of assessment is important; Research may require standardisation but with different standards depending on the research question; Standardisation is essential for some research studies in order to be replicated; Routine clinical work may require less standardisation; In research, a researcher may choose highly standardized procedure different from McKenzie's recommendations for the purposes of determining if another method is superior.
Statements 24: Too prescriptive for clinical practice, and unnecessary; For research purposes only; The evidence and a recent systematic review on centralization support a measurement tool to document centralization. Perhaps one of several reasons for the large variance in the prevalence rates of centralization across studies is the lack of a standardized documentation process; This is an objective measurement tool and physical therapy guidelines encourage therapists to use objective measurement procedures; Pain overlay template is one possibility however, there are other ways especially if using computer-based assessment.
Statement 25: Too prescriptive for clinical practice, and unnecessary; Palpation has no validity

Loading strategy (Figures 4.6, 4.7, 4.8): Participating experts agreed that testing should mainly involve repeated and / or sustained end-range movement techniques in the standard planes of movement of the spine. Progressive or alternative forces introduced by the patient or the therapist could also be used if appropriate as well as combined movements. Relying exclusively on single movement testing was deemed insufficient by the vast majority of participating experts to elicit this sign, with the

number of repetitions relying on the clarity of the elicited symptom response. Finally, participating experts did not agree on whether the order of testing affects the outcome of or the ability to detect centralisation.

Figure 4.6 Section on the type of loading strategy from the third round feedback report Note: N = number of valid responses; % agree = percentage of participants that agreed with statement, green ≥ 80% agreement; 70% ≤yellow ≤ 79% agreement.

D. Loading strategy – Type

Centralization can be elicited:

	N	% agree
26. By repeated end-range movement testing only	29	3.4
27. By repeated and/ or sustained end-range movement testing	28	60.7
27a. By repeated and/ or sustained end-range movement testing **only**	29	10.3
28. By single and repeated movements. Sustained movements are used with caution to prevent confusion with instability	28	14.3
29. By repeated and/or sustained end-range movement testing, or manual techniques executed by therapist if appropriate	29	79.3
30. By repeated and/or sustained end-range movement testing, overpressure by patient or therapist or manual techniques executed by therapist if appropriate	30	90.0
31. Regardless of testing methods used. This can involve test movements or after immediate treatment application, or post treatment over time	28	17.9

COMMENTS
Statement 28: I have no idea what the statement means as instability has not been defined; Centralization should be distinct from instability and manual techniques, therefore should avoid sustained movements and therapist overpressure.
Statements 29, 30: Overpressure is a manual technique so there is no distinction between 29 and 30; It is essential to allow for manual overpressure by the therapist because some patients cannot achieve the appropriate end range for many different reasons. In the case of correction of a lateral shift, self correction by the patient often fails, but manual shift correction causes the centralization phenomenon easily; I think a good compromise would be to put a period in #30 after the phrase "over pressure by patient or therapist." And the delete the manual techniques wording
Statement 31: I have no idea what this statement means or what it refers to

Figure 4.7 Section on the planes and directions of the loading strategy from the third round feedback report Note: N = number of valid responses; % agree = percentage of participants that agreed with statement, green ≥ 80% agreement.

E. Loading strategy - Planes and directions

Testing for centralization should:

	N	% agree
32. Involve only the sagittal plane	28	3.6
33. Involve the standard planes of movement available to the spinal region involved	27	33.3
34. Involve the standard planes of movement available to the spinal region and/or a combination of movements if appropriate	29	93.1

COMMENTS
Statement 33: What are the standard planes of movement?
Statement 34: Side gliding is not considered a standard plane of movement by many, but an essential inclusion in the lumbar spine. Retraction in the cervical spine suffers the same divergence of opinion and is very important in eliciting centralization; This statement does not include sustained positions

Figure 4.8 Section on the movement testing of the loading strategy from the third round feedback report Note: N = number of valid responses; % agree = percentage of participants that agreed with statement, green ≥ 80% agreement.

F. Loading strategy - Movement testing

Movement testing for centralization should:

	N	% agree
35. Not be based solely on single movements. Test movements must be performed repeatedly, to the fullest of the patient's available end-range	30	86.7
36. Not include more than 10 repetitions for each test movement	30	20.0
37. Include repetitions that are continued long enough to ensure that the status of the symptoms has changed and/or is clear.	30	93.3
38. Not necessarily be performed in a specific order. The sequence of movement testing does not affect the outcome or the ability to detect centralization	30	40.0

COMMENTS
Statement 35: Does not include sustained positons.
Statement 37: The exact number of repetitions is not important, the pattern recognition is
Statement 38: This is supported by published data (Donelson et al. in Spine vol 16 1991); We do not know, it has not been studied; In some instances eg patient with a lateral shift, the order of movements is important but not in all cases

Criteria for a positive test (Figure 4.9): Changes in the intensity of symptoms or neurological changes were considered important when testing for the centralisation phenomenon. However, some participants highlighted the need to differentiate between centralisation and directional preference (i.e. preference for specific postures or movements in one direction compared to another) (McKenzie and May, 2003; McKenzie and May, 2006), two overlapping but not synonymous terms. There was no agreement as to the way that distal pain should change location (i.e. simply abolishing rather than progressively receding toward the spine). However, there was substantial agreement that observed changes should be lasting.

Figure 4.9 Section on the criteria for a positive test from the third round feedback report Note: N = number of valid responses; % agree = percentage of participants that agreed with statement, green ≥ 80% agreement.

G. Criteria for positive test

When defining the presence of centralization:

	N	% agree
39. Changes in the intensity of symptoms should not be considered	30	20.0
40. Changes in the neurological status (symptoms or signs) should not be considered	31	12.9
41. Nonspinal conditions or underlying disease states potentially causing or influencing symptoms should be excluded or taken into consideration	31	87.1
41a. Underlying disease states potentially causing or influencing symptoms should be excluded	31	67.7
41b. Nonspinal conditions potentially causing or influencing symptoms should be taken into consideration	31	87.1
42. The reduction of peripheral symptoms towards the spinal midline should be progressive and stable. Distal pain which simply abolishes rather than progressively receding toward the spine should not be considered in the definition	31	22.6
43. Observed changes should be retained over time (lasting change after testing)	29	86.2

COMMENTS
Statements 39 & 40: The clinician should differentiate between a directional preference and centralization. Both terms overlap but these terms are not synonymous; Further work is required to operationally define these terms to decrease the confusion between clinical identification of directional preference and/ or centralization. Without this important clinical discussion and continued research confusion will continue regarding how best to define centralization.
Statement 41, 41a, 41b: I'm not sure nonspinal conditions/disease states are criteria for a positive test. They definitely need to be considered as part of complete evaluation
Statement 42: More testing is usually needed
Statement 43: Most appropriate for pure criteria for positive test; Statement needs to be rephrased: over how much time and under what conditions? How long after testing?; Does not apply to centralization, but if referring to the process of centralization or centralizing in the direction of full centralization, that is another topic; It is just the ability of changes to predict longer lasting changes that are of primary concern. Otherwise one is at risk of a circular argument simply suggesting that if someone improves over time that they have improved over time.

Timeframe (Figure 4.10): Participants agreed that the response to centralisation tests may not necessarily be obvious on the day of assessment. In some cases, assessment may be required over a period of time to confirm centralisation. There was uncertainty about the expected change in symptoms over time, or the time course of change. Some participants expressed some reservations for a potential confusion with the natural course of improvement and recovery.

Figure 4.10 Section on the timeframe for determining centralisation from the third round feedback report Note: N = number of valid responses; % agree = percentage of participants that agreed with statement, green ≥ 80% agreement; 70% ≤yellow ≤ 79% agreement

H. Testing – Timeframe

When testing for centralization:

	N	% agree
44. The response to testing may be obvious during the first examination or may require evaluation over a period of time to confirm the phenomenon	31	100.0
45. Symptoms may be sequentially and lastingly abolished at each session (pure response) or gradually abolished in a progressive manner but not sequentially (partial response)	27	74.1

COMMENTS
Statement 45: This statement is too complex and confusing; Sequentially should be defined; Can't say I've seen a non-sequential abolition moving gradually toward complete abolition; Centralization rarely if ever takes more than a week to be identified, unless the patient is slow to adopt the postural changes and exercise regime between assessments; I agree with 45, but it does not include all possibilities.
Statements 44 & 45: These questions are hard to answer because they both combine the timeframes. I can disagree with both of them and still answer yes.

Safety issues (Figure 4.11): Participants agreed on the importance of the change in the distal symptom regardless of the proximal presentation when defining whether a patient is getting better or worse as a result of a movement, position etc. There was no consensus on the importance of centralisation in the face of deteriorating changes in other signs e.g. neurological status. Some participants supported that centralisation is accompanied with improvements in other signs but others argued that this is not always the case. Patient education was considered an integral part of the clinical

interaction so that any fear of movement is minimised.The content of education based on incorrect assessment findings concerned one of the participants.

Figure 4.11 Section on safety issues for determining centralisation from the third round feedback report Note: N = number of valid responses; % agree = percentage of participants that agreed with statement, green ≥ 80% agreement.

I. Safety issues

	N	% agree
46. The patient's status (improving or worsening) should be considered beyond symptom relief i.e. neurological picture, range of motion. For example, if the patient's symptoms are improving but his neurological status is compromised, centralization should be considered an undesirable outcome for this patient.	27	59.3
47. Symptoms getting worse proximally but improving distally should be considered a positive sign	30	100.0
48. Education of patients is essential following the use of these techniques so that movement is normalized as symptoms settle and patients do not develop fear of movement	30	90.0

COMMENTS
Statement 46: Unclear question. Was neural status compromised at the outset and is unchanged, or is it deteriorating as a result of the assessment? If the latter, one must cease, although, again, that never happens in the face of centralization. That's one of the wonderful safe guards about the MDT assessment. Certainly, monitoring neural status is paramount and deterioration cannot be accepted; Neurolgical status should be monitored additionally e.g. if neurologic status is worsening and only pain is improving, this should not be considered a positive sign; Question 46 seems to me a purely "researcher" question. Never happened to me that centralization were linked to a worsening of the general status; True centralization with concurrent worsening of neurological status must be vanishingly rare. I cannot recall a case in 30 years of doing this; In this scenario centralization is not the undesirable outcome - the worsening neurological status is the undesirable outcome and centralization of symptoms is besides the point - this is an important distinction but not a good description of it. The point is that if an intervention is worsening the patient's neurological status then it may be inappropriate despite centralization of symptoms. This appears to happen occasionally in patients with spinal stenosis- the little used 'pheasant's sign' was reported to look for this phenomenon.
Statement 48: This seems to speak to intervention, not examination; I would agree with 48, but it would appear to be a general principle and have little to do with concepts of centralization; Education is certainly essential. However, some therapists make the error of discontinuing the movement causing centralization and the expected increase in proximal pain. The end result is that the wrong patient education is then provided

Note: MDT, Mechanical Diagnosis and Therapy

Attributes (Figure 4.12): There was unanimous consensus that the presence of centralisation translates into a good prognosis. However, the same certainty was not expressed on the importance of centralisation in relation to psychosocial factors in spinal pain. Finally, participating experts strongly disagreed that centralisation is present only in patients with internal disc disruption, not connecting this phenomenon to an anatomical event.

Figure 4.12 Section on other issues related to centralisation from the third round feedback report Note: N = number of valid responses; % agree = percentage of participants that agreed with statement, green ≥ 80% agreement.

Section 3: Other issues related to centralization

	N	% agree
49. Centralizers have a good prognosis	30	96.7
50. Centralization may be a stronger prognostic factor than psychosocial variables	29	65.5
51. Centralizers have an internal disc disruption	27	14.8

COMMENTS
Statement 50: Psychosocial variables are another issue
Statement 51: This is supported by Laslett M et al TSJ 2005; There is currently evidence in the lumbar spine but not the cervical spine; More evidence is needed in support of #51; I disagree strongly with 51 and think its inclusion potentially reduces the credibility of the concept of centralization; I don't know, I don't care
Statements 49, 50 & 51: All three statements are true to some extent. However, those with severe psychologcial distress may require psychological intervention before being able to adhere to the self-treatment program required to centralize and abolish pain. Those with severe IDD (i.e., grade IV on the Dallas Discogram Scale) or who have satellite fissures not connected to the nucleus of the disc may not centralize.

Note: IDD, Internal Disc Disruption.

4.4.2.2. Future research recommendations

Results after the third round are presented in Figure 4.13. The effect of training and procedures in the prevalence and outcomes of patients who centralise, the prevalence, reliability and outcomes of centralisation in neck pain and the comparison of outcomes of individuals who fail to centralise for various interventions were the most popular questions for future research. Participants were also interested in the natural and clinical course of individuals whose symptoms centralise in response to testing as well as the best intervention for this symptom response group. The extent to which centralisation co-exists with other clinical findings, the diagnostic accuracy of centralisation across different populations and examiners and expected rates of centralization for acute and chronic patients with LBP were some of the research questions that were also of interest to participants. Studies on clinical predictors for symptom response groups and the relative importance of centralization, the effect of standardisation on the reliability of centralisation and the role of the history and the examiner's training in predicting centralisation were also included in the consensus list.

Figure 4.13 Section on future research from the third round feedback report Note: N = number of valid responses; % agree = percentage of participants that agreed with statement, green ≥ 80% agreement; 70% ≤yellow ≤ 79% agreement

Section 4: Future research

Future research should look at:

	N	% agree
52. Operational definition for centralization:	**29**	**86.2**
57. Mechanism causing the centralization phenomenon (including physiological mechanism)	**29**	**79.3**
58. Diagnostic accuracy of centralization:	**28**	**92.9**
60. Cross validation of diagnostic accuracy in different patient populations and examiners	**30**	**86.7**
62. Centralization and prevalence:	**28**	**96.4**
63. Expected rates of centralization in clinical settings for acute and chronic patients with LBP	**29**	**86.2**
64. Effect of training and procedures in the prevalence and outcomes of centralizers	**28**	**100.0**
65. Centralization and course/ prognosis:	**28**	**92.9**
66. Do centralizers have a favourable course when left untreated (natural history)?	**28**	**92.9**
67. Clinical response versus natural course of centralizers	**30**	**93.3**
68. Prescriptive validity of centralization:	**28**	**89.3**
69. Do centralizers benefit from interventions other than directional movement exercise and which intervention provides the greatest benefit?	**29**	**89.7**
70. Comparison of outcomes of non-responders to centralization for various interventions	**30**	**96.7**
71. Centralization and outcomes (i.e. ability to return to work, psychosocial outcomes, economic outcomes, health care utilization, QoL, recurrences and ability to self-manage ; Cost effectiveness of the McKenzie assessment prior to disc surgery, injections, imaging or any other treatments)	**28**	**85.7**
72. Centralization and subgroups:	**28**	**85.7**
73. Clinical predictors (CPR) of patients responding (or not) to directional preference exercises (including: Effect of patient compliance and attitudes of centralizers on outcomes; Effect of severe disability and psychosocial distress on the diagnostic, prognostic and treatment characteristics of centralizers; Centralization versus other prognostic factors)	**28**	**82.1**
74. Are centralizers and other clinical subgroups (e.g. patients satisfying CPR for manipulation or trunk stabilization treatment or candidates for disc surgery) made up of the same patients? How often do centralization findings co-exist with other findings? (i.e. relationship of centralization with other variables e.g. psychosocial, clinical findings e.g. segmental provocation signs, sciatica and EMG findings). How soon will positive EMGs become normal after centralization has been achieved and maintained?	**28**	**89.3**
74a. Are centralizers and other clinical subgroups (e.g. patients satisfying CPR for manipulation or trunk stabilization treatment or candidates for disc surgery) made up of the same patients?	**28**	**71.4**
74b. How often do centralization findings co-exist with other findings? (i.e. relationship of centralization with other variables e.g. psychosocial, clinical findings e.g. segmental provocation signs, sciatica and EMG findings).	**27**	**77.8**
74c. How soon will positive EMGs become normal after centralization has been achieved and maintained?	**27**	**77.8**
75. Centralization and the cervical spine:	**27**	**92.6**
77. Reliability of detecting, prevalence and outcomes in neck pain	**29**	**96.6**
77a. Reliability of detecting centralization in neck pain	**25**	**92.0**
77b. Prevalence of centralization in neck pain	**27**	**92.6**
77c. Outcomes of centralization in neck pain	**27**	**92.6**
78. Predictive validity of centralization in the management of neck pain	**27**	**92.6**
79. Reproducibility:	**29**	**86.2**
80. Stability/ reversibility of the centralization phenomenon	**28**	**78.6**
81a. Effect of the clinician characteristics on the reliability of centralization	**27**	**74.1**
81b. Effect of the patient characteristics on the reliability of centralization	**28**	**78.6**
81c. Effect of standardization of the test on the reliability of centralization	**28**	**82.1**
81d. Effect of knowledge of the test on the reliability of centralization	**28**	**71.4**
82a. Role of history in predicting the presence of centralization	**28**	**82.1**
82b Role of examiner's training in predicting the presence of centralization	**28**	**85.7**
85. Education and training:	**29**	**86.2**

COMMENTS
Statement 70: "non-responders to centralization"? Makes no sense. Perhaps "non-centralizers"?
Statement 71: Change to 'physical therapy assessment' rather than 'McKenzie assessment' and I will agree.
Statement 74: While useful for prognosis, there has not been any indication in the line of research utilizing clinical prediction rules to determine the effectiveness of various treatment approaches that centralization of symptoms in the absence of neurological signs is a sign that should trump other factors (i.e. time in the manipulation CPR), particularly given different in effect sizes- symptoms distal to the knee seems to be important. These two in combination (symptoms distal to the knee and centralization of symptoms) seem to be likely candidates for patients that will respond best to a direction specific treatment approach more than other treatment approaches. Utilizing 'centralizers' as a stand alone subgroup does not seem to be the best approach; Question 74 contains 3 questions, at which am I supposed to answer? Questions 74 b and c: how is possible to answer with yes or no to questions starting with: "How"?
Statement 85: No answers for question #85?

Note: EMG, Electromyography; QoL, Quality of Life

4.4.3. Final comments

Responding participants' final comments are listed in Appendix 4.11.

4.5 DISCUSSION

4.5.1. Discussion of findings

This is the first study to achieve formal consensus on the centralisation phenomenon as a physical examination sign. Some of the proposed definitions in this chapter have been identified by previous authors (George and Fritz, 2005; Werneke and May, 2005; Berthelot et al., 2007) but there has never been such a comprehensive account of operational components for this test. Indeed, Delphi participants bring a wide range of direct knowledge and experience to the decision – making process (Murphy et al., 1998).

The Delphi survey was successful in establishing a general definition and operational criteria, as well as future research questions for centralisation. However, comments made by participants indicated that opinions and knowledge around centralisation are not necessarily uniform across the expert panel. Thus, the use of feedback and comments to inform panellists between rounds and the results of the final round may have risen or led to an increased collective awareness of the current knowledge base and its possible discrepancies, limitations or gaps (Stokes, 1997).

Participating experts supported a broader definition including not only testing approaches such as repeated and / or sustained end-range movement testing, but also progression of forces or alternative forces in multiple directions and in various

starting positions when appropriate. This may be expected considering that most members of the Delphi group were predominantly using the McKenzie (MDT) classification system (71%) over the Delitto (TBC) (23%) or other classification and management approaches. The MDT system has been in common use for more than 20 years (Hefford, 2008), therefore popularity of this method over other classification approaches may be understandable.

The philosophy underpinning the MDT approach is argued to be different from the TBC approach (May et al., 2008). In the MDT approach, movement testing, force progressions and a range of force alternatives are tested over 3-5 sessions in order to identify a specific exercise strategy (May et al., 2008). In contrast, the TBC system exhausts movement testing much earlier (May et al., 2008). Treatment prescription (e.g. manipulation, stabilisation exercises, and specific exercise) is primarily based on criteria from clinical prediction rules. This preference against extensive movement testing has also been implied in studies of the TBC group (George et al., 2005; Brennan et al., 2006; Browder et al., 2007).

Overlap within the definition 'centralisation' of both systems exists (May et al., 2008). What this Delphi study indicates is that there is mutual appreciation of the different classifications for centralisation. This concept is not new; various practitioners may have different but equally acceptable approaches to the management of a particular classification category (Binkley et al., 1993). In fact, some authors argue that in clinical practice, the same therapist may often combine approaches instead of exclusively following one system only (Battié et al., 1994; Jette et al., 1994; Pinto et al., 2007). This lack of a preference for a specific approach

may also be reasonable since no comparative studies have been reported between classification approaches.

Participating experts acknowledged that an average clinician should be able to identify centralisation, but agreed that training is essential to use the test. The type and amount of required training was not determined. Earlier studies have demonstrated that centralisation can be identified reliably when trained individuals are involved (Fritz et al., 2000a; Kilpikoski et al., 2002). However, recent research on the identification of symptom response groups failed to provide reliable results with trained clinicians (Dionne et al., 2006; Fritz et al., 2006). The effect of training and procedural variations in testing on prevalence and outcomes of individuals who centralize is still unknown and requires further investigation. This research question was the most popular for future investigation.

Participating experts showed a particular interest in investigating the utility of this sign in the cervical spine. In contrast to the lumbar spine, research and evidence on centralisation in the cervical spine (Werneke et al., 1999; Tuttle, 2005; Tuttle et al., 2006; Cleland et al., 2006; Cleland et al., 2007; Fritz and Brennan, 2007) is preliminary. Clinicians often base their theories on the assumption that all spine areas have similar characteristics (Mercer and Jull, 1996). Thus, classifications and treatment for the cervical and thoracic spine groups have been suggested to follow similar patterns to the lumbar spine (Hefford, 2008). Some authors have supported that changes in the shape or position of the intact nucleus pulposus as a result of movements or sustained positions of the spine are responsible for ‘centralisation’ and peripheralisation of symptoms in the cervical (McKenzie, 1990) as well as the

lumbar spine (McKenzie, 1981; Stevens and McKenzie, 1988; Magnusson et al., 1995). This was not supported by participating experts and is in keeping with the literature (Mercer and Jull, 1996; Parent et al., 2006). The lack of a conceptual model did not detract the Delphi panelists from the potential clinical utility of this sign. This may also reflect current views that clinical anatomy relates weakly to applications other than those directed at surgical or medical practice and decision making (Mercer and Rivett, 2004; Zusman, 2005). However, the need for more research on centralisation in the cervical spine was identified and prioritised by participating experts. Further work into the prevalence, reliability of identification and the outcomes of centralisation is therefore required in neck pain.

Most experts supported the notion that centralisation is most applicable to patients with referred or radicular symptoms. One reason identified for this was the difficulty in defining changes in location when only central symptoms are involved. Werneke and associates (1999) used the strictest criteria by defining centralisation as the abolition of symptoms when central pain only exists. In contrast, other authors (Karas et al., 1997) have argued that a decrease in symptom intensity is a sufficient criterion for the presence of centralisation, or have excluded patients with spinal symptoms above the knee or elbow area (Childs et al., 2004a). Recent research (Powers et al., 2008) found that the effects of spinal manipulation or extension press-ups may be similar in low back pain patients with central symptoms. This study did not however follow-up patients for long-term effects or extent its findings to other spinal areas.

Some subtle differences appeared to emerge in the participants' perception of the process of testing and the end result of identifying centralisation. Aina et al. (2004) have argued that the presence and not the stability of centralised symptoms are important for spinal outcomes. This was also reflected in the deliberation of the panel that included lasting changes after testing as a criterion for centralisation. Centralisation may not be elicited easily in all patients; experts agreed that for some spinal sufferers, it may require more than one testing session to confirm. For example, Werneke and Hart (2003) found that 60% of people who did not centralise on day one of assessment had their symptoms fully or partially centralised over the next few visits. However, caution was expressed by Delphi participants that centralisation should not merely be a product of the natural history of spinal pain.

Participants acknowledged that conditions influencing symptoms should be taken into account when assessing spinal patients for the presence of centralisation. The value of red flags in spinal screening was not disputed by the panel. Red flags are signs and symptoms that raise suspicion of serious spinal pathology (Greenhalgh and Selfe, 2006). Despite studies suggesting an association between psychosocial factors (i.e. yellow flags) and pain responses (Laslett et al., 2005; Werneke and Hart, 2005) and the acknowledgement by experts that other factors may play a role when using symptom response testing, there was uncertainty regarding the use of psychosocial screening. This observation may indicate that not all clinicians fully understand how to and when to evaluate psychosocial factors. Although the biopsychosocial model is widely accepted within the spine care community (Weiner, 2008), its implementation may still be problematic.

Although objective measurements and standardisation are advocated in patient assessment (Rothstein et al., 1991), using objective tools for day-to-day practice was not recommended. In a UK study, Turner et al (1996) suggested that the use of recognised quantified methods for pain assessment is not standard practice. This is also consistent with other studies (Kirkness et al., 2002; Abrams et al., 2006; Copeland et al., 2008). There are a range of ways in which centralisation could be documented, but the impact of variation in recording methods is unknown.

4.5.2. Strengths and Limitations

This study should be viewed in the light of its strengths and limitations. While formal consensus methods were developed to meet the requirements of scientific methods, they have been subjected to relatively little methodological research within the healthcare field (Black, 1994; Murphy et al., 1998). Neither the validity nor the reliability of the Delphi method is well investigated (Walker and Selfe, 1996). In addition, there is little scientific evidence regarding the effectiveness of such methods in improving the quality of healthcare or reducing costs (Black, 1994). However, some authors argue that methodologies such as those applied in Delphi studies can not lend themselves to traditional scientific approaches of evaluation (Mullen, 2003). Nevertheless, there are still mechanisms of critically appraising the study findings when using such methods (Powell, 2003).

Investigator bias. A potential limitation might exist if the team or individuals who monitor the study have bias that distorts the results, exploits the privacy of the respondents or imposes too restrictive a process on the participants, not allowing

consensus to occur (Wilson et al., 2003). A number of strategies were used to minimise any such risk: use of explicit rules for the procedures followed and the quantification of consensus before the beginning of the study, inclusion of all responses in the first round, minimal editing on presented statements, analysis checks by a second reviewer and an opportunity for panelists to suggest, add or refine items in the round questionnaires.

Consensus or compromise? An advantage of iteration in Delphi studies is that in successive rounds, participants can change their opinions, support them further, agree or disagree with other opinions and present their arguments (Couper, 1984). Feedback from other panel members might also convince some to consider items they might have missed or thought unimportant (Couper, 1984). This is in agreement with arguments supporting rationality in formal consensus methods (Murphy et al., 1998). Some authors, however, question the value of successive rounds, because conformity may be produced rather than consensus (Goodman, 1987; Binkley et al., 1993).

A reason behind this criticism may be that many Delphi studies do not report the stability of consensus or the convergence of agreement between rounds (Greatorex and Dexter, 2000). In other words, Delphi studies should not focus only on whether consensus was achieved and what the final opinion was but also whether the consensus agreement existed throughout each round or was reached in the later rounds as a result of the Delphi process (Greatorex and Dexter, 2000). Some authors even suggest that results from the second round may be more informative of participants' opinions and perspectives than findings from subsequent rounds

(Binkley et al., 1993). In this study, when looking at the second round responses, the consensus definition was still the preferred definition for the majority of the Delphi members (79% of participants). However, approximately half (38% out of the 79%) of the panellists who supported the consensus definition strongly (rather than simply) agreed with it. Relevant comments made by the participants at the third round indicate that lack of strong agreement by all supporting panellists may relate more to the level of detail used in rather than the core essence of the general definition per se. The latter fact also confirms the importance of investigating agreement for individual operational criteria for centralisation, rather than restricting exclusively to the establishment of a general definition.

Identifying the conformist panellists who abandoned the majority of their original opinions and exclude them from the analysis could have shed further light on whether consensus or conformity was achieved (Greatorex and Dexter, 2000). It has been suggested that such panellists are not really experts (Linstone and Turoff, 1975, cited in Greatorex and Dexter, 2000). If there was a clarification of opinion when these individuals are removed, then the Halo effect (individuals conforming to the group opinion whether they agree or disagree with an opinion) would be less likely to be the cause of any change of opinion (Greatorex and Dexter, 2000). However, removing members of an expert panel may also be problematic since experts who refuse to change their minds may be as damaging to the decision making process as experts who always conform (Greatorex and Dexter, 2000).

An alternative approach would be for experts who deviated in the third round from their original opinion to be asked to explain why they have changed their opinion

(Greatorex and Dexter, 2000). This may have offered some insight into whether the observed change was a result of a genuine change of opinion and may have had the additional benefit of discouraging panellists from simply conforming to group opinion (Greatorex and Dexter, 2000). However, whilst attempting to gain feedback on why experts are changing their opinions may have had many benefits, it would also have added to the complexity, cost and length of the study (Greatorex and Dexter, 2000). Furthermore, this approach would have been more useful if stability of responses was not achieved between the second and third round, which was not the case in this study.

Reliability. Reliability in Delphi studies reflects the extent to which the Delphi rounds produce similar results under constant conditions on all occasions (Hasson et al., 2000). The selection of items in the consensus list reflected the opinions and experience of participating experts (Pincus et al., 2008). In contrast to another recent Delphi survey (McCarthy et al., 2006) testing for centralisation was not limited to single movements. However, the primary aim of the UK study (McCarthy et al., 2006) was not to define centralisation but to produce a list of the most important discriminatory items for the sub-classification of LBP. The majority of our Delphi group consisted of individuals with postgraduate training and extensive experience in using centralisation in clinical practice. Our participants felt that single movement testing only is inadequate for eliciting centralisation in some people with spinal pain. However, it is likely that single movement testing may represent the approach used by some practitioners who may not be familiar with all test procedures required to elicit this sign (McCarthy et al., 2006). Although healthcare professionals favour what they are most familiar with (Murphy et al., 1998), literature to date does not

support the use of single active range of motion assessment for centralisation (Chapters 2 and 3).

Validity. Despite efforts to include equal numbers between the two main classification systems, the disproportionate numbers of professionals with McKenzie training over other groups using centralisation may have biased the study findings. This may have decreased the validity of this study because if the panel is skewed in some way, this will not truly reflect the range of opinions (Streiner and Norman, 2003). However, opinions on many consensus items were consistent with current evidence on centralisation (Appendix 4.12). In a similar way, uncertainty in some items was in keeping with missing or conflicting evidence on centralisation (Appendix 4.12). The findings in this chapter are still useful despite uncertainty in some areas; they can be used to raise issues of debate that can be explored in future research on centralisation. Recent emphasis on evidence-based practice (EBP) appears to favour opportunities for greater integration of qualitative research findings into the professional knowledge base of healthcare professions (Fritz, 2004). In view of the significant amount of interest in sub-grouping in the clinical literature (Borkan et al., 2002; Childs et al., 2008), this Delphi consensus may pave the way for the standardisation of centralisation as a physical sign and for further study of potential sub-groups and classification of spinal syndromes.

4.6 SUMMARY

This is the first study attempting to establish consensus on the centralisation phenomenon internationally among researchers and clinicians. Centralisation was defined as the progressive and stable reduction of the most distal pain towards the

spinal midline in response to standardised assessment procedures. This definition remains to be tested, first in an inter-reliability design. Nevertheless, the results of this study may contribute to the standardisation of centralisation as a physical sign and provide common directions for future research in the field.

SECTION 3

INVESTIGATING AND IMPROVING RELIABITY OF CLINICALLY INDUCED SYMPTOM RESPONSES IN NECK PAIN

CHAPTER 5

The inter-rater reliability of the identification of centralisation and related clinically induced symptom response groups in neck pain.

5.1 AIMS OF CHAPTER

In Chapters 2 and 3, two systematic reviews on the reliability and prognostic value of clinically induced symptom responses in spinal pain concluded that there was strong evidence supporting the use of distal symptom location and / or intensity changes in response to repeated spinal movement testing for the prediction of long-term LBP and work-related outcomes (Appendix 5.1). Limited evidence was also found supporting an association of treatment induced changes in pain location or intensity with neck pain (NP) and low back pain (LBP) (Appendix 5.1). These clinically induced symptom changes, cited by some as the 'centralisation phenomenon' (e.g. McKenzie, 1981; Karas et al., 1997; Werneke et al., 1999; Laslett et al., 2005), are consistent with the recommended operational criteria by the Delphi panel (Chapter 4).

Although the usefulness of centralisation has been investigated extensively in LBP, its reliability and prognostic value are still unclear in NP. This was acknowledged by the Delphi panel (Chapter 4) who identified the investigation of the reliability and prognostic utility of centralisation in NP as in the top priorities for future research. This chapter presents the reliability investigation on centralisation in NP.

The aim of the study presented in this chapter was to assess the inter-rater reliability of the identification of clinically induced symptom responses based on the centralisation phenomenon in patients with NP. Reasons for identified variability in clinicians' judgements (measurement error) were also explored.

5.2 INTRODUCTION

There are few data on the inter-rater reliability of centralisation in NP. Two studies focused on the inter-rater reliability of the McKenzie classification / sub-classifications and directional preference[10] (Clare et al., 2005; Dionne et al., 2006), but the detection of centralisation was not addressed in either of these studies. Werneke et al. (1999) investigated the agreement in coding the location of the most distal pain in an overlay body template and categorising patients into the centralisation and non-centralisation groups. Although agreement of therapists was excellent both in coding the distal location (κ = 0.92 to 1.0) as well as categorising 15 NP and 15 LBP patients (κ = 0.96), these results reflected judgements related to interpreting the pain diagrams and did not extend to differences in the clinical examination. Cleland et al. (2006a) found fair to substantial agreement (κ = 0.44 and 1.00) when assessing centralisation in response to cervical active range of motion in the sagittal plane (extension and flexion respectively), but poor reliability for side bending and rotation (κ = -0.05 to 0.2) in 22 participants with NP. These results were in contrast with Piva et al. (2006) who found moderate to substantial agreement for symptom response to these movements in 30 people with NP.

Variations in reported definitions and methods to identify clinically induced symptom response groups may account for inconsistencies in observed values. For example, Werneke et al. (1999) used clinician judgement of patient completed diagrams of primarily repeated movements acquired over multiple visits to define centralisation in contrast to others (Cleland et al., 2006a; Piva et al., 2006) who used

[10] Directional preference implies the presence of centralisation in most cases; however, preference for a specific spinal strategy may be determined even if centralisation is not observed (Werneke, 2005).

a single assessment of active range of motion. Reliability may have varied depending on the two methods requiring difference in clinical skills as well as the potential for symptom variations over time. In the light of the findings of the Delphi exercise, that single movements are inadequate for eliciting centralisation (Chapter 4) and reservations about the practicality of using a multiple-visit definition, further research was needed to investigate the reliability of the identification of centralisation in the cervical spine.

5.3 MATERIALS AND METHODS

5.3.1 Study design

This was an inter-rater reliability study using a single-group repeated measures design in a sample of participants with NP.

5.3.2 Participants

5.3.2.1 Recruitment

Participants with NP were recruited through a prospective pilot cohort study of patients with a primary complaint of NP, with / without symptoms in the head and / or upper extremity. The study was originally planned to be part of a trial of exercise in the management of neck / shoulder symptoms in the occupational health setting (SENSE study) (Luime, 2005), but this investigation was not undertaken because of lack of funding. In keeping with the SENSE study, recruited participants were referred from the safety and occupational health departments (SOHDs) of companies in the area (University of Warwick and Coventry University). SOHDs acted as 'gatekeepers' through whom potentially eligible individuals could be contacted,

informed about the study, and screened for referral to the research team,. The study's information leaflets are presented in Appendices 5.2 and 5.3. Promotion of the study was made through advertisement - posters and email alerts. A study website was also available at the following address:

www2.warwick.ac.uk/fac/med/research/ctu/trials/otherresearch/neckpainstudy/

5.3.2.2 Eligibility criteria

The eligibility criteria for the selection of participants are listed in Table 5.1. Participants were classified as Grade I to III in the severity category of the Task Force on Neck pain, encompassing all neck pain cases without an identified serious structural or psychological pathology (Guzman et al., 2008).

Table 5.1 Eligibility criteria for the selection of participants.

Inclusion criteria
Eligible participants were included in the study if:
(1) their age was between 18-65 years
(2) they presented to the Safety and Occupational Health Departments with mechanical symptoms in the neck and / or shoulder area (i.e. pain in the cervical area with or without referral to the upper extremity reproduced/influenced by neck movements, provocation tests or sustained postures)
(3) they were able / willing to give informed consent and complete self-report questionnaires
(4) they were able to understand, speak and write in the English language.
Exclusion criteria
Eligible participants were excluded if:
(1) they did not fulfill the aforementioned inclusion criteria
(2) they presented with signs and symptoms potentially indicating the presence of serious pathology, or a confirmed non-musculoskeletal problem e.g. people with severe psychiatric or personality disorders
(3) they suffered from a previous traumatic injury to the affected upper limb(s) or shoulder girdle(s) resulting in current or prolonged disability
(4) they were expected to receive major medical or surgical treatment within the next 3 - 4 months
(5) they did not consent or had a legal inability to participate in the study.

5.3.3 Examiners

Five physiotherapists (1 male and 4 females), with a mean age (± SD) of 39 (± 9.8) years (range 29 to 54 years) took part as the examiners for the reliability analysis. A practical strategy of forming physiotherapist pairs, previously applied in inter-rater reliability investigations of sacro-iliac joint physical examination tests was followed (Laslett and Williams, 1994; Robinson et al., 2007). This strategy involved one physiotherapist being the examiner in all assessment sessions and the second physiotherapist randomly selected from a pool of four based on their availability on a given participant.

The participating physiotherapists were all members of the UK Health Professions Council (HPC) and the Chartered Society of Physiotherapy (CSP) and had been in clinical practice for mean (± SD) duration of 20 (± 11.7) years (range 7 - 33 years). Some of the participating physiotherapists, but not all, had been working in full time musculoskeletal practice prior to the study. They also spent on average 11% of their working week managing patients with NP. Three out of five physiotherapists had previously used symptom response assessments based on the McKenzie classification system, however, experience in using centralisation or a relevant system was not a prerequisite for participating as an examiner. This decision was made on the basis of some evidence (Fritz et al., 2000a) and the Delphi consensus recommendations that training in the procedures rather than experience is important for consistent results (Chapter 4). It was also made to ensure external validity in the light of suggestions that not all clinicians are familiar with this physical sign (McCarthy et al., 2006).

All participating physiotherapists received eight hours of training to ensure that study procedures were performed and interpreted in the same manner. Approximately 3 hours were spent on eliciting and interpreting centralisation. Training was provided by the author of this thesis who has experience in using centralisation and symptom response tests and is an accredited clinician in the McKenzie method. Training involved studying a manual of the test procedures with the operational definitions as well as a practical session involving the assessment of a volunteer with neck problems (selected sections in Appendix 5.4). The manual also including information on the study background, procedures and treatment, was developed by the author of this thesis and refined in consultation with participating physiotherapists.

Each physiotherapist met with the investigator before data collection to ensure proper reporting and performance of the test procedures. A formative competency test (Appendix 5.5) comprising questions about the study procedures as well as clinical reasoning completed the training. No formal scoring was applied after the completion of the competency test (only discussion of options and reasons for selection of responses) since there is no widely agreed method of assessment of knowledge and skill.

5.3.4 Development of the operational definition for centralisation

When the operational definition was developed, there were no published guidelines or systematic reviews specific to NP assessment. In view of this, decisions had to be based on the available spinal literature and conclusions made in Chapters 2, 3 and 4. After the definition was finalised, guidelines for NP have been reported elsewhere (Childs et al., 2008; Nordin et al., 2008)[11]. A critical review of the literature on diagnostic procedures for NP and LBP (Rubinstein and van Tulder, 2008) has also been published. Critical appraisal information on the guidelines is found in Appendix 5.6.

The guideline produced by the Bone and Joint Decade 2000-2010 Task Force on Neck Pain and its associated disorders, was based on the results of a systematic review looking at various components of the assessment and management of people with NP (Carroll et al., 2008a; Nordin et al., 2008). From the clinical examination, some data were found on procedures such as inspection, range of motion, strength, palpation and neurologic tests covering reliability, diagnostic accuracy or utility but not prognostic value (Carroll et al., 2008a; Carroll et al., 2008b; Carroll et al., 2008c; Nordin et al., 2008). Others (de Koning et al., 2008b; Rubinstein et al., 2008) also found scarce evidence to support the use of most physical examination tests in clinical practice. Thus, an immediate and strong need was identified to test commonly used procedures for the aforementioned attributes[12] (Nordin et al., 2008).

[11] Many of the centralisation studies in the cervical area date later than the performed search strategies. Evidence is currently updated for the US Veterans Affairs (Nordin, personal communication 22nd June 2009).

[12] The only exception was provocation tests for cervical radiculopathy and manipulation. With regards to centralisation, no data were found in the cervical spine.

The guideline produced by the Orthopaedic section of the American Physical Therapy Association was also based on a review, but methods to identify and appraise the evidence were not explicitly stated. This guideline recommended measures of impaired function of muscle, connective and neural tissues associated with the identified pathology (Childs et al., 2008), i.e. NP classifications based on the *International Classification of Functioning, Disability and Health* (WHO, 2001). This recommendation was based on theoretical evidence and was surprising in view of the authors' suggestions that the tissues causing patients' neck complaints are frequently unknown (Childs et al., 2008). According to the same authors, centralisation procedures / exercises were not beneficial in reducing neck disability compared to other interventions (Childs et al., 2008). However, this recommendation was based on a single study by Kjellman and Oberg (2002) who treated patients irrespective of their symptom response classification. Furthermore, supporting evidence from more recent relevant studies (Wang et al., 2003; Klaber-Moffett et al., 2006) was ignored.

5.3.5 Procedures

5.3.5.1 Ethical approval and initial contact

The study was carried out in compliance with the principles of the Helsinki Declaration (WMA, 2008) and IASP guidelines for pain research on humans (Charlton, 1995). Approval was given by the Biomedical Research Ethics Subcommittee at the University of Warwick (Appendix 5.7).

After obtaining ethics approval, participating SOHDs received the relevant study forms. Once a potentially eligible individual was identified by the SOHDs using a screening form (Appendix 5.8), an appointment was scheduled with the research team to provide further information (if required), verify eligibility and obtain a written consent as well as the baseline measurements. A flowchart of the procedures at baseline is presented in Figure 5.1.

All participants provided informed consent (Appendix 5.9) prior to participating in the study. Within the context of the VideoNeck study (Chapter 1), consenting participants were also asked if they agreed to be videotaped. Following this, participants provided demographic and General Practitioner (GP) information and were given a standard proforma (baseline questionnaire) (Appendix 5.10) comprising various self-report questionnaires (Table 5.2).

A history and physical examination took place after the completion of the self-report questionnaires using a standardised procedure and assessment form (adapted from McKenzie and May, 2006). Two physiotherapists were used for the assessment of each patient. The order of testing by either physiotherapist 1 or 2 was random, determined by a blind draw to avoid the same physiotherapist consistently undertaking the first examination and introducing systematic order bias (McCarthy et al., 2007). Two cards with the number of the order in the assessment (1 or 2) were prepared, folded and placed in sealed opaque envelopes. The physiotherapists who were going to undertake the assessment opened their selected envelope on the day indicating who was first and who was second in the assessment order.

5.3.5.2 History

The history was undertaken by one physiotherapist, with the second in attendance. This allowed for both physiotherapists to receive the same background information before testing and reduce the time with each participant. The first physiotherapist led the history taking, and the second was permitted to ask questions once the first physiotherapist had finished. The mean (± SD) duration of history taking was 33 (± 8.9) (range from 7 to 50) minutes.

All participants were asked a sequence of pre-specified questions, including information about current and previous episodes of NP (e.g. the location, duration and mode of onset of symptoms, the course of symptoms, whether symptoms are constant or intermittent, aggravating and easing factors) as well as personal and work-related information (Table 5.2). A body diagram was used to record information about the area, location and type (e.g. pain, abnormal sensation) of symptoms the participant was experiencing during the current episode (Bryner, 1996; McKenzie and May, 2006).

The history took place with the participant sitting on a plinth in an unsupported position so that the participant's posture could be observed and naturally progress to the physical examination procedures. History taking is part of the NP assessment process and does not seem to affect the reliability of physical examination tests for the neck / shoulder (Bertilson et al., 2003).

5.3.5.3 Physical examination

Upon completion of the patient history, the physical examination started with an observation of the participant and his / her posture. This was followed by an alteration of the participant's posture and documentation of his / her response to this procedure (see correction of posture, Appendix 5.4). Then, the second physiotherapist left the examination room in order for the first to proceed with the remainder of the physical examination.

During the physical examination, each physiotherapist examined patients and recorded findings independently from the other physiotherapist. Independence of findings is necessary in reliability studies so that results are unbiased and reliability coefficients are not overestimated (Sim and Wright, 2005). Findings were not shared among physiotherapists and patients were instructed not to disclose information about their examination during testing[13]. The mean (± SD) duration of physical examination and symptom response assessment was 71 (± 18.2) (range from 45 to 139) and 45 (± 14.9) (range 24 to 102) minutes.

Before the symptom response spinal testing, active cervical range of motion (ROM) measurements were obtained by both physiotherapists using the Cervical Range-of-Motion (CROM) device (Performance Attainment Assoc, St Paul, Minn). The CROM device is an instrument specifically designed for measurements of movement and posture of the cervical spine (Jordan, 2000; de Koning et al., 2008a). Recent reviews comparing various instruments for the measurement of cervical spine range

[13] The only exception was if the safety of the participant would be compromised by not disclosing relevant information.

of motion have suggested its superiority over other tools (Jordan, 2000; de Koning et al., 2008a). The order of movements was protrusion, retraction, flexion, extension, lateral flexion and rotation (right side tested first). Physiotherapists also recorded the limiting factor for each movement. Assessing physiotherapists were allowed to use other tests e.g. neurological testing, shoulder tests if appropriate.

Valid estimation of reliability requires the characteristics under study to be stable (Piva et al., 2006). Pain properties may be susceptible to change as a result of natural variation over time (British Pain Society, 2008) or mobilising effects of the test procedure itself (Piva et al., 2006). The examinations of physiotherapist 1 and 2 were done within the same day / session with 2-5 minute breaks allowed between them. To minimise the possibility that symptom response testing by the first physiotherapist would cause a true change in the participant's mechanical presentation and influence the range of motion measurements, both physiotherapists performed the active range of motion tests before the symptom response testing.

Figure 5.1 Flowchart of study procedures as described in the research clinicians' manual at baseline. CROM, Cervical Range of Motion; GP, General Practitioner.

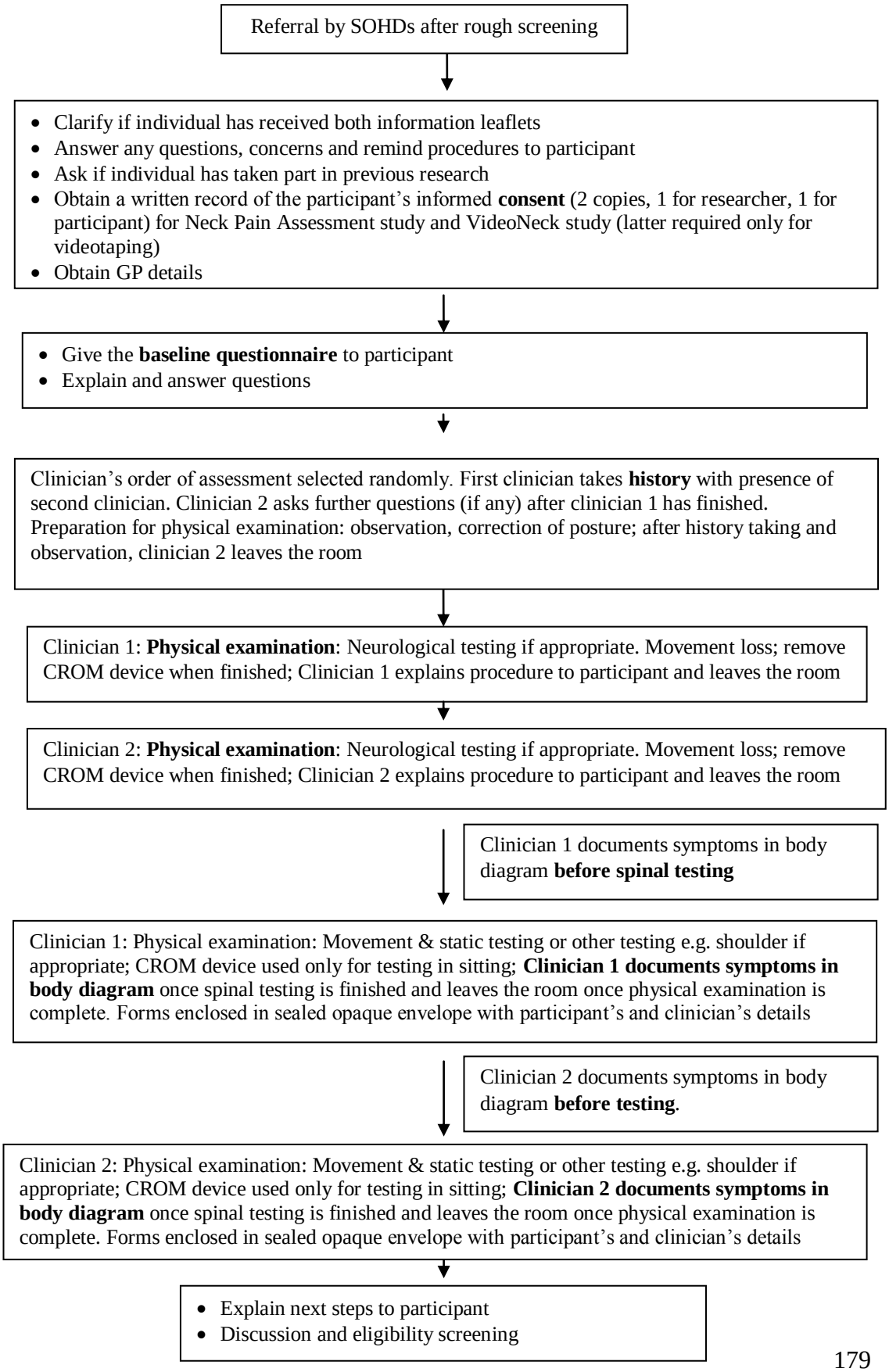


Table 5.2 Domains at the time of the baseline examination

Domains	Source	Measurement method
Pain intensity*	Baseline questionnaire	11-point Likert numeric rating scale (NRS-11)
Pain bothersomeness	Baseline questionnaire	5-point pain bothersomeness scale
Neck Disability	Baseline questionnaire	Neck Disability Index (NDI) (Vernon and Mior, 1991)
Upper extremity disability	Baseline questionnaire	Disability of the Arm, Shoulder and Hand (DASH) scale (Hudak et al., 1996)
Fear avoidance	Baseline questionnaire	Tampa Scale of Kinesiophobia (TSK) (Kori et al., 1990)
Self – efficacy	Baseline questionnaire	Pain Self-Efficacy questionnaire (PSEQ) (Asghari and Nicholas, 2001)
Duration of symptoms	History	Patient self-report
Course of symptoms	History	Patient self-report
Mode of onset	History	Patient self-report
Symptom location	History	Patient self-report / Body diagram
Constant/ Intermittent	History	Patient self-report
Aggravating factors	History	Patient self-report
Easing factors	History	Patient self-report
Disturbed sleep	History	Patient self-report
Previous episodes	History	Patient self-report
Comorbidity	History	Patient self-report
Red flags	History	Patient self-report
Work-related information	History	Patient self-report
Personal circumstances	History	Patient self-report
Posture	Physical examination	Observation
Relevance of posture correction	Physical examination	Patient self-report
Neurological signs	Physical examination	Neurological examination
Active range of motion/ Movement loss	Physical examination	Cervical Range of Motion (CROM)
Symptom response to spinal testing	Physical examination	Patient self-report
Mechanical response to spinal testing	Physical examination	Cervical Range of Motion (CROM)

*This refers to the current pain and the average pain during past week.

The assessment for symptom response involved single and repeated movement testing, sustained postures, or manual techniques and overpressures if appropriate. For each testing procedure, the physiotherapist asked the participant to describe the characteristics i.e. nature and location of symptoms and to rate the intensity of symptoms before testing. After the testing, the physiotherapist again asked the participant[14] to describe the status of same characteristics of his / her symptoms. All the above characteristics were required to ensure relevance to the symptoms of the reported current episode and to enable physiotherapists to make a classification judgement about the participant. Procedures causing a lasting peripheralisation of or increase in the distal symptom intensity were not tested further, following recommendations of previous studies (Fritz et al., 2000a), relevant textbooks (McKenzie and May, 2006) and suggestions on the importance of the change in the distal symptom regardless of the proximal presentation when defining whether a patient is getting better or worse (Chapter 4, Safety issues).

Pain drawings (in body diagrams) were used to record the results of the symptom response testing (Appendix 5.11). Each patient was instructed by the evaluating physiotherapist to shade all the relevant areas on the body diagram where he / she was experiencing spinal pain and referred symptoms twice: before and after the spinal examination. Body diagrams were completed in sitting. The recording position was standardised to sitting, as this was the starting point for all tests, and any change in position may have changed symptoms, as a result of the position *per se* rather than

[14] Pain is a subjective experience and as such, it can usually be ascertained by individuals' reports (IASP Task Force, 1994).

the testing procedure previously performed. After the completion of the body diagrams, examining physiotherapists provided their classification judgements.

Patients were classified into centralisation, non – centralisation or 'other' groups in two ways; first, based on changes in the location of the most distal symptoms, and second, based on changes in symptom location and / or intensity of the most distal symptoms. A 'better', 'worse' and 'no change' classification was also considered based on changes in symptom location and / or intensity of the most distal symptoms. Table 5.3 presents the list of categories that were evaluated in the reliability analysis.

Table 5.3 List of categories evaluated for reliability.

Classification based on location	
Centralisation (CP_L)	Distal symptoms migrate to a more proximal location or central symptoms are abolished
Non-centralisation (non-CP_L)	Symptoms peripheralise or distal symptom location remains the same after testing
Other ($Other_L$)	No symptoms prior to the mechanical assessment; unable to classify patient
Classification based on location and / or intensity	
Better (B)	Distal symptoms migrate to a more proximal location, are abolished or reduced
Worse (W)	Distal symptoms peripheralise, are produced or increase in intensity
No Change (NC)	Distal symptom location and / or intensity remain the same after testing

5.3.5.4 Intervention

Following the assessment, an appointment was scheduled where participants received advice about their condition (Appendix 5.12). This involved a one-to-one session. All participants, regardless of their symptom response classification, were given the Neck Book (Waddell et al., 2004) including evidence-based information on their problem, advice to remain active[15], exercises and effective self-care strategies. This pragmatic approach is in accordance with current practice (Chapter 1), population beliefs (Bostick et al., 2009) and policies (DoH, 2001), core self-management principles (Chou et al., 2007; Liddle et al., 2007; DoH, 2008) and recent guidelines on the management of neck pain at the individual level (Haldeman, 2008).

Advice was provided by the author of this thesis who reinforced the Neck Book key messages, demonstrated exercises and answered all participants' questions. A letter to the GP was also sent informing of the subject's participation in the study (Appendix 5.13).

5.3.6 Study endpoints

The endpoints were agreement in the overall classification and individual symptom response categories involving the centralisation phenomenon (Table 5.3).

[15] "Activity" refers to the following: mobility and activities of daily living; recreational and sports – related activities; occupational activities (Abenhaim et al., 2000).

5.3.7 Data handling

5.3.7.1 Data protection and missing data

All data were treated with confidentiality and were not accessed by anyone other than the research team or authorised individuals at the University of Warwick. None of the participants were referred to with their names or any identifiable data. Participants' information was stored in locked filing cabinets and on password protected computer files. Participants were free to withdraw from the study at any time and without having to give any reason. However, if possible reasons for withdrawal could be ascertained, these were recorded accordingly in notification forms (Appendix 5.14).

5.3.7.2 Data entering and cleaning

Anonymised questionnaires and examination form data were transferred into an electronic format. All data were entered by the author of this thesis. Analyses were performed using SAS, the Statistical Package for Social Sciences (SPSS), version 15.0 (Chicago, ILL), Microsoft Excel, and hand calculations where appropriate. For the purposes of reducing error and inconsistencies, the followings steps were used: (a) 10% of data entered was independently checked by a second person (SPl), (b) the SPSS descriptive statistics and data validation module was used, and (c) reliability data from statistical packages were double checked against hand calculations (Angeliki Chorti and Tim Friede).

5.3.8 Statistical analysis

5.3.8.1 Quantitative data

Descriptive statistics were used as appropriate to characterise the sample and describe patterns of missing data. Raw agreement was reported as the percentage of agreement, calculated as the ratio of the number of agreements between observations of the first and second physiotherapist to the total number of comparisons made (Haas, 1991). For the reliability of overall symptom response classification and individual categories, point estimates and measures of variance were calculated whenever possible for each pair of physiotherapists as well as for the whole sample. This approach was followed because the investigation of both individual and overall agreement is more informative than calculating overall agreement only (Maclure and Willett, 1987; Haas, 1991; Armitage and Berry, 2002). The rationale behind, and the specifics for the chosen statistical analyses are discussed in detail below. An overview of approaches for measuring agreement and reliability statistics is provided in Chapter 2.

5.3.8.1.1 Reliability of pairs of physiotherapists

In order to identify the components of the overall reliability, separate measures of the inter-rater reliability of the judgements of physiotherapists were calculated for each pair of physiotherapists (Armitage and Berry, 2002). The simple unweighted form of kappa (Cohen, 1960) was applied because of the non-ordered nature of the data (Ludbrook, 2002; Sim and Wright, 2005). The kappa statistic and confidence intervals (95% CIs) were calculated based on Altman (1991).

5.3.8.1.2 Reliability of the overall classification

An overall measure of reliability was constructed for the entire sample of participants and physiotherapists. Participants were examined by different pairs of physiotherapists; as a result, the classic two-rater form of κ may be not appropriate as in the case of the same physiotherapists rating the whole participant sample (Hubert, 1977).

Fleiss and associates (Fleiss, 1971; Fleiss and Davies, 1982; Hale and Fleiss 1993) proposed a version of kappa for the case where each participant was rated on a nominal scale by a fixed number of raters, selected from a larger pool of raters. This statistic is often referred to as the *generalised kappa coefficient*, although it generalises the two-rater π statistic of Scott (1955) rather than the κ statistic of Cohen (1960) (Gwet, 2008a; Gwet, 2008b). The generalised kappa coefficient was used in this chapter to describe the reliability of the overall symptom response classification because in contrast to other kappas proposed for multiple ratings per participant, it recognises that examiners rating one participant are not necessarily the same as those rating another. When using the generalised version of kappa, what is evaluated is the degree of inter-rater agreement with respect to a specific category λ versus the remaining categories (Λ-1) and this process is repeated for each λ. λ =1,2,…Λ (Fleiss, 1971). This yields a set of Λ interdependent kappa statistics each representing a distinct comparison (Donner and Eliasziw, 1997). The overall measure of agreement is then obtained by calculating a weighted average of pairwise agreement corrected for the amount expected by chance (Williams, 1976; Fleiss et al., 1979; Conger, 1980; Donner and Eliasziw, 1997).

The variance estimator proposed by Fleiss (1971), was used for testing the assumption of no agreement among raters than would be expected by chance[16]. It was not used for confidence interval construction (Gwet, 2008a) and therefore, precision of the observed agreement coefficient was not measured (Gardner and Altman, 1986). Fleiss and Davies (1982) provided a lower confidence bound for the generalised kappa based on the estimated large-sample variance of kappa. Because coverage using the above approach is not satisfactory for extreme values of kappa and for small sample sizes typical in reliability studies such as the study in this chapter (Fleiss and Cicchetti, 1978), Hale and Fleiss (1993) developed a lower bound interval, superior to the method proposed by Fleiss and Davies (1982). However, the lower bound interval by Hale and Fleiss (1993) was not estimated in this study because it is applicable to two-category classifications only.

5.3.8.1.3 Reliability of individual categories

For the reliability of individual categories, a maximum-likelihood estimator for kappa (Bloch and Kraemer, 1989) which Kraemer (1979) named the *intra-class kappa coefficient* was used. Formulae available for the standard error (Bloch and Kraemer, 1989; Garner, 1991) and sample size (Donner and Eliasziw, 1992) later extended to cases of more than two ratings (Altaye et al., 2001) or categories (Donner and Eliasziw, 1997) were used as described in Roberts (2008). In small sample sizes where the distribution of the kappa estimate tends to be non-symmetric, this approach has been more preferable (Garner, 1991).

[16] The null hypothesis (H_0) in this case would be: $H_{0:}\ \kappa = 0$.

Some question the use of an overall measure of agreement if the rater is not homogeneous[17] (Gwet, 2008a). The issue of rater heterogeneity was considered in the design of the study, by using physiotherapist pairs independently for each participant from a large pool of raters rather than using the same two raters for the whole sample (Roberts, 2008). This approach has practical advantages because it is not always possible to design a reliability study with the same raters evaluating throughout (Cicchetti and Sparrow, 1981). Calibrating ratings of observers so that the marginal distributions are identical could also have been used to ensure exchangeability of ratings (Becker and Agresti, 1992). However, this would have been difficult to achieve, and would pose a threat to the external validity of the study. Empirical support was also gained by inspecting for marginal homogeneity between paired ratings (Appendix 5.15); minor differences in the marginal distributions between ratings 1 and 2 were observed which supported the assumption of exchangeability of ratings (Roberts, 2008).

Kappa is a descriptive measure of agreement that is not based on a specific model of data distribution (Simon, 2006). Since all categories are treated with equal significance, the impact of a category on kappa can only be determined *a posteriori* according to the category's frequency (Simon, 2006). The investigation of heterogeneity in the pattern of agreement has been suggested to give more insights as to the degree of agreement in some pairs of categories rather than others (Roberts, 2008).

[17] This refers to marginal homogeneity between ratings.

Heterogeneity in the pattern of agreement can be investigated by considering the kappa coefficient for the indicator variable of each category (Roberts, 2008). The intraclass kappa coefficient (Bloch and Kraemer, 1989) described above or the approach suggested by Fleiss (1971) or Landis and Koch (1977a) are such methods. In general, estimated kappa coefficients for individual categories average agreement between the specified category and all others together (Roberts, 2008). The best method for calculating kappa for individual categories is unknown. However, the generalised version of kappa (Fleiss, 1971) can suffer from the same problems or paradoxes described for the original (Cohen, 1960) or other (Donner and Eliasziw, 1997) kappa reliability coefficients (Gwet, 2008b) and even be negatively biased for chance-generated data (Fleiss et al., 1979; Conger, 1980). Therefore, it was decided to present both kappas for individual categories proposed by Fleiss (1971) and Bloch and Kraemer (1989) for comparative purposes.

Another approach to investigating patterns of agreement is the one proposed by Tanner and Young (1985). Tanner and Young (1985) proposed a model of conditional independence that did not suffer from assigning negative agreement to chance-generated data. In contrast to most linear models treating rates in a symmetric manner, this model provided asymmetric interpretations which are appropriate for the design used in this chapter. Another advantage is that agreement is investigated from the perspective of a population model, rather than providing a test statistic (Kraemer et al., 2002). Although this approach is appealing, it was not followed for this data set. Such models are difficult to interpret and suffer from similar limitations due to marginal heterogeneity between raters (Roberts, 2008).

When considering the categories of 'CP', 'non-CP' and 'Other', the terminology that differentiates category 'CP' from 'non-CP' may differ from that differentiating 'CP' and 'Other' resulting in greater agreement between some pairs of categories than others (Roberts, 2008). Identification of pairs of categories that are easily confused can suggest changes that improve reliability, but a kappa statistic for a specific category gives only limited insight into this (Kraemer, 1979; Roberts, 2008).

Roberts and McNamee (1998) proposed a matrix of kappa-type coefficients that can be used to investigate agreement in distinguishing between pairs of categories. These kappa coefficients, called *inter-class kappa coefficients,* are interpreted in much the same way as the intra-class kappa coefficients (Roberts, 2008); a value equal to 1 implies that the two categories are not confused at all, while a value of zero implies that the two categories are indistinguishable (Roberts and McNamee, 1998). In contrast to alternative methods e.g. log-linear or latent class models of investigating patterns of agreement (Chapter 2), intra-class kappas are easier to interpret, and therefore are more likely to be adopted by researchers in the clinical field.

The intra-class kappa coefficient was used in this chapter *post hoc* to gain insights on agreement in distinguishing between pairs of categories. Unlike conventional kappas, intra-class kappas indicate where deficiencies in the measurement method or training of observers may lie (Kraemer, 1979) and thus, they can potentially be used to improve reliability of clinically induced symptom responses.

5.3.8.2 Other data

Information from performing the testing in the videotaped assessments was combined with the physiotherapists' recorded information in the assessment forms to identify discrepancies between physiotherapists' judgements and clarify possible reasons for variability in the symptom response classifications (i.e. based on changes in distal symptom location). This exploration addressed some variations due to measurement error and not variations between individuals within a population and thus, may partly explain reliability findings (De Vet et al., 2003). However, it was considered necessary in view of the fact that very few factors that could affect measurements are usually described in reliability papers (Van Genderen et al., 2003). Table 5.4 presents common sources of measurement error identified in the literature.

Table 5.4 Common sources of measurement error in reliability studies (Fritz et al., 2000a; Nordin et al., 2008; Trudelle-Jackson et al., 2008).

A. Biologic variation in the same participant: Stability of symptoms between assessments.
B. Variability in performance: Discrepancies in testing and recording e.g. starting and ending positions, instructions to participants, actual performance of tests, method of completing the pain diagrams.
C. Variability in judgements: Pain diagram versus physiotherapist's classification judgement.
D. Learning effect on participant.

5.3.9 Sample size

Literature on sample size estimation in reliability studies is limited, with scarce information on hypothesis testing of non-null hypotheses and interval estimation for samples of small to moderate size (Donner and Eliasziw, 1997). This resulted in a conservative approach for the calculation of sample size (Donner, email

communication 2006). The formula by Donner and Eliasziw (1992) and the succeeding table by Sim and Wright (2005) for binary outcomes were used. Donner and Eliasziw (1997) also provided formulas for questions involving multiple outcomes; however, these formulas were not used because they did not account for inferences regarding a primary trait of interest, i.e. centralisation.

The sample size calculation was performed with the aim to detect a statistically significant kappa coefficient with 80% power at a range of expected proportions (0.5 and 0.7)[18] and assuming that the null hypothesis value of kappa would be 0.4[19] (Sim and Wright, 2005). Currently, there is no agreement on the minimum acceptable value of kappa for physical tests (May et al., 2006) or potential prognostic indicators (Wainner, 2003). However, values equal and over 0.40 are pragmatic and relevant in the day-to-day clinical practice (Schneider et al., 2008). Based on reported values of kappa in published systematic reviews on the centralisation phenomenon before the study started (Aina et al., 2004; May et al., 2006), a $\kappa = 0.8$ was expected. Using the formula provided by Donner and Eliasziw (1992), the number of subjects required to detect a statistically significant value of $\kappa = 0.8$ ($p \leq 0.05$) on a dichotomous variable with 80% power and assuming the null hypothesis value of $\kappa_0 = 0.4$ was 42 and 48 individuals at a proportion of 0.50 and 0.70 respectively.

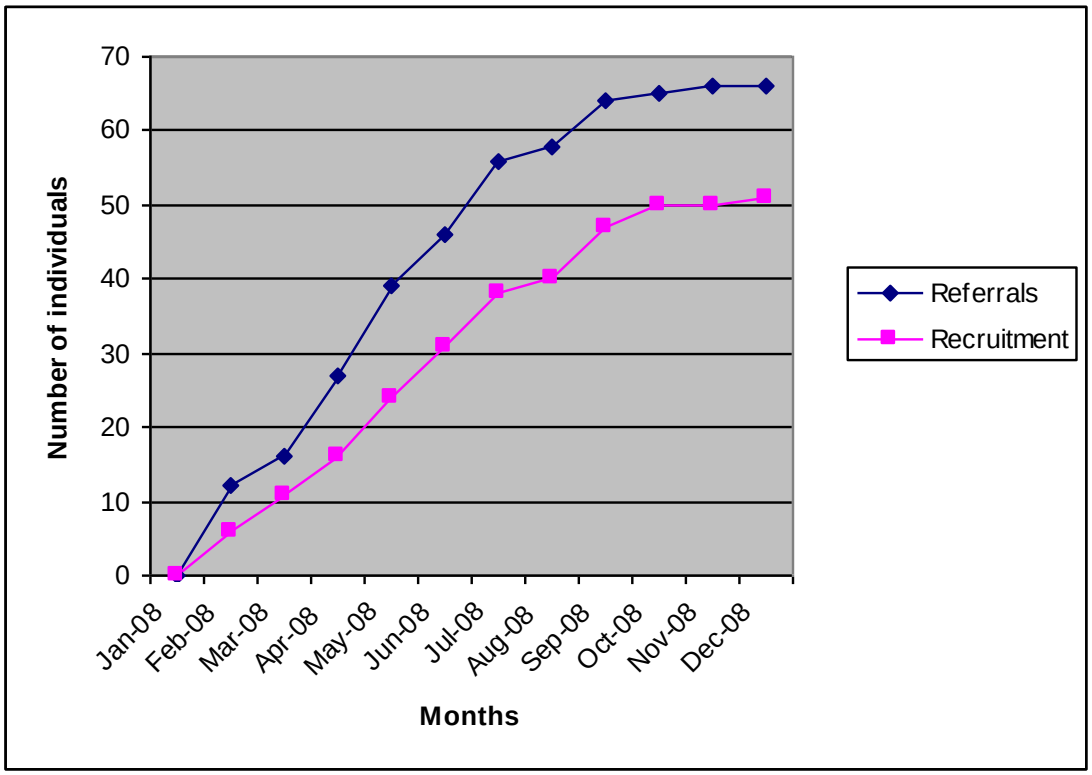
[18] See Chapter 1 for prevalence rates.
[19] In practice, agreement is usually better than expected by chance across clinicians and thus, a zero value for kappa in the null hypothesis would not have been meaningful unless its plausibility can be justified (Garner, 1991; Posner et al., 1990; Petersen, 1998).

5.4 RESULTS

5.4.1 Description of participants

Data collection took place from January 2008 until December 2008. In total, 51 individuals entered the study. Monthly referral and recruitment rates are presented in Figure 5.2.

Figure 5.2 Monthly referral and recruitment rates.



From the 51 individuals who consented to participate, 48 were considered for the reliability analysis (Figure 5.3) resulting in a total of 96 symptom response examinations. The characteristics of consenting individuals and the participants for the reliability analysis are shown in Table 5.5.

Figure 5.3 Reliability **s**tudy flowchart. SOHDs, Safety and Occupational Health Departments.

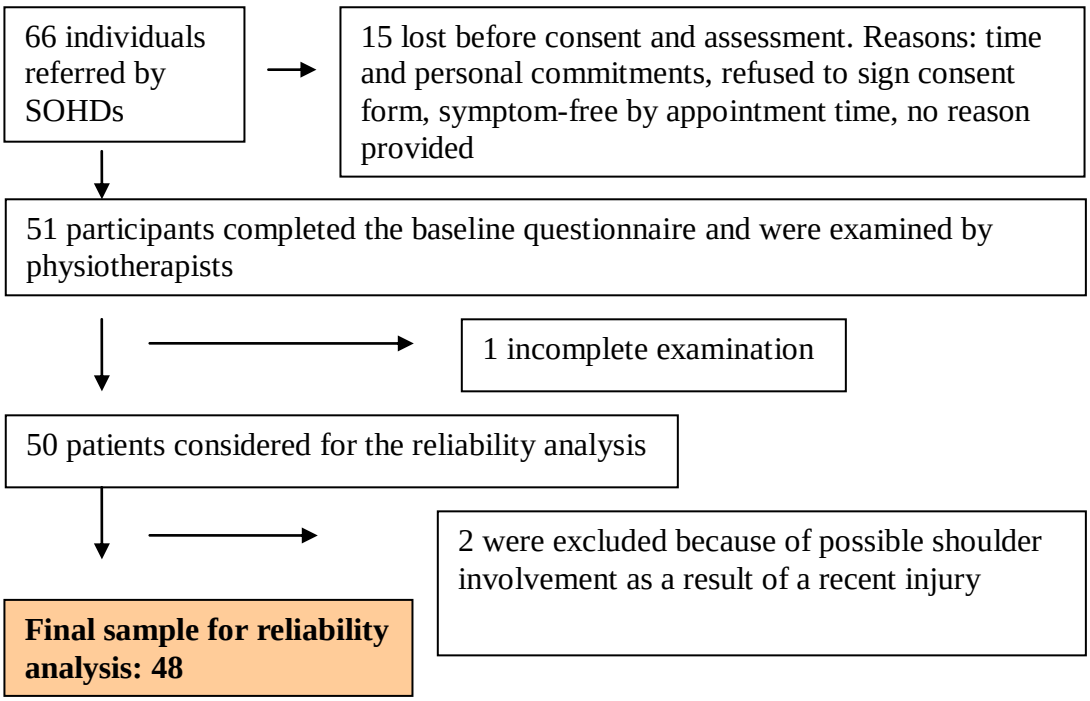


Table 5.5 Characteristics of participants at the time of the assessment.

Characteristic	All (n=51)	Reliability (n=48)
Age in years, mean (SD)	45.3 (12.7)	44.7 (12.6)
% Men	28	29
% Distal symptoms below the elbow	28	25
% Non-traumatic mode of onset	90	94
% participants with prior history of neck problems	84	83
Present pain intensity, mean (SD)	2.5 (1.9)	2.2 (1.7)
Average pain intensity last week, mean (SD)	3.6 (2.2)	3.4 (2.1)
Pain bothersomeness, median (range)	2.0 (2.0-3.0)	2.0 (2.0-3.0)
NDI mean % (SD)	19.4 (11.3)	18.7 (10.3)
DASH mean (SD)	18.1 (14.5)	16.9 (13.3)
DASH – W mean (SD)	21.7 (20.2)	20.3 (19.0)
% currently off work	0	0
TSK mean (SD)	34.8 (5.8)*	34.4 (5.7)[§]
PSEQ mean (SD)	52.3 (7.7)	52.7 (6.9)

Note: %, percentage; DASH, Disabilities Arm Shoulder Hand; DASH-W, Work component of the DASH questionnaire; NDI, Neck Disability Index; PSEQ, Pain Self Efficacy Questionnaire; SD, standard deviation; TSK, Tampa Scale of Kinesiophobia. Numbers rounded.
* Data on 44 / 51; [§] Data on 44 / 48.

5.4.2 Reliability

5.4.2.1 Reliability between pairs of physiotherapists

Reliability results for the overall symptom response classification between pairs of physiotherapists are presented in Table 5.6. Appendix 5.15 presents the raw data for these results. The reliability of symptom response classifications for individual pairs of physiotherapists ranged from 0.06 to 1.00 (95% CI ranging from -0.69 to 1.00) for changes based on location of ('CP_L', 'non-CP_L', '$Other_L$') and 0.04 to 0.61 (95%CI ranging from -0.32 to 1.00) ('B', 'W', 'NC') for changes based on location and / or intensity of the most distal symptoms (Table 5.6). Agreement was higher in the pair with more extensive symptom response training and experience (AGC / SLW).

5.4.2.2 Reliability of the overall symptom response classification

The reliability of overall symptom response classification using the generalised κ (Fleiss, 1971)[20] was 79%, $\kappa = 0.66$ ($p < 0.05$) based on changes in symptom location ('CP_L', 'non-CP_L', '$Other_L$'), and 58%, $\kappa = 0.33$ ($p < 0.05$) based on changes in symptom location and / or intensity ('B', 'W', 'NC').

5.4.2.3 Reliability of the individual symptom response categories

Results are presented below and in Table 5.7.

5.4.2.3.1 Fleiss' κ

Agreement in assigning participants to the 'CP_L', 'non-CP_L' and '$Other_L$' categories was $\kappa_{CPL} = 0.94$ ($p < 0.05$); $\kappa_{Non\text{-}CPL} = 0.36$ ($p > 0.05$); $\kappa_{OtherL} = 0.60$ ($p > 0.05$). For

[20] Agreement in assigning to the symptom response groups was statistically significant under the null hypothesis of no agreement beyond chance (Fleiss, 1971).

changes in symptom location and / or intensity, agreement in assigning participants to the 'Better', 'Worse' and 'No change' categories was: $\kappa_B = 0.47$ ($p > 0.05$); $\kappa_{W} = 0.10$ ($p > 0.05$); $\kappa_{NC} = 0.34$ ($p > 0.05$).

5.4.2.3.1 Roberts' κ

Agreement in assigning participants to the 'CP_L', 'non-CP_L' and '$Other_L$' categories was $\kappa_{CPL} = 0.40$ (0.07, 0.73), $p > 0.05$[19]; $\kappa_{Non\text{-}CPL} = 0.68$ (0.46, 0.90), $p < 0.05$[19]; $\kappa_{OtherL} = 0.64$ (0.36, 0.92) $p > 0.05$[19]. For changes in symptom location and / or intensity, agreement in assigning participants to the 'Better', 'Worse' and 'No change' categories was: $\kappa_B = 0.50$ (0.25, 0.75), $p > 0.05$[19]; $\kappa_{W} = 0.08$ (-0.07, 0.23), $p > 0.05$[19]; $\kappa_{NC} = 0.34$ (0.03, 0.65), $p > 0.05$[19].

5.4.2.4 Patterns of agreement

Results for patterns of agreement are displayed in Tables 5.8 and 5.9. When changes in distal symptom location were considered, the categories 'CP_L' and '$Other_L$' were mostly confused ($\kappa_{CPL/OtherL} = -0.05$, [-1.14, 1.04], $p > 0.05$[21]). The 'CP_L' and '$Other_L$' categories were characterised by the lowest prevalence rates (0.17 and 0.18, respectively). Differences between $\kappa_{CPL/OtherL}$ and $\kappa_{CPL/NonCPL}$ or $\kappa_{NonCPL/OtherL}$ were not statistically significant ($p > 0.025 / 3$ with Bonferroni correction). 'CP_L' /'$NonCP_L$' was less distinguishable than '$NonCP_L$'/'$Other_L$' ($\kappa = 0.52$, [0.16, 0.88], $p > 0.05$[19] versus $\kappa = 0.82$ [0.58, 1.00], $p < 0.05$[19]), but again this difference was not statistically significant ($p > 0.05$).

[21] This is for H_0: $\kappa = \kappa_0 = 0.40$, one-tailed test. Note: CI values lying beyond 0.4 would be expected to be statistically significant.

When changes in the distal symptom location and / or intensity were considered, the categories 'Better' and 'No change' were well distinguishable ($\kappa_{B/NC}$ = 0.73 [0.44, 1.02], $p < 0.05$). 'Better' / 'Worse' ($\kappa_{B/W}$ = 0.32 ([0.07, 0.71], $p > 0.05$, or 'Worse' / 'No change' ($\kappa_{W/NC}$ = -0.26 [-0.33, 0.85], $p > 0.05$) were poorly distinguished, with differences not being statistically significant ($p > 0.05$). Differences between all three categories were not statistically significant ($p > 0.05/3$ with Bonferroni correction).

5.4.2.5 Sources of error

Results related to potential sources of measurement error when considering changes in distal symptom location are presented in Table 5.10. Most sources of error were attributed to biologic variations of a participant's symptoms over time (6 / 10 cases) and subsequent performance variations (7 / 10 cases). There were also variations/errors in the interpretation of a positive test (3 / 10 cases).

Table 5.6 Reliability results of the overall symptom response classification for each pair of physiotherapists.

Physiotherapist pairs	Location							Location and / or intensity						
	n	CP_L	Non-CP_L	$Other_L$	%	κ	95% CI	n	B	W	NC	%	κ	95%CI
Pair 1	12	0.17	3.50	1.33	92	**0.86**	**(0.59, 1.00)**	12	0.67	0.75	2.92	50	**0.22**	**(-0.23, 0.66)**
Pair 2	19	0.42	8.21	0.63	74	**0.49**	**(0.10, 0.87)**	16	3.38	1.56	0.63	38	**0.04**	**(-0.32, 0.41)**
Pair 3	12	0.17	7.50	0.08	67	**0.06**	**(-0.69,0.81)**	12	3.00	0.75	0.75	67	**0.47**	**(0.04, 0.89)**
Pair 4	7	0.57	3.57	0.00	100	**1.00**	**(1.00, 1.00)**	7	4.29	0.14	0.00	86	**0.61**	**(-0.09, 1.00)**

Note: n= number of individuals with neck pain; %, percentage agreement; 95% CI, 95% confidence intervals; B, marginal proportion of the category 'better' determined by changes in distal symptom location and / or intensity; CP_L, marginal proportion of the category 'centralisation' determined by changes in distal symptom location; κ, kappa statistic; NC, marginal proportion of the category 'no change' determined by changes in distal symptom location and / or intensity; Non - CP_L, marginal proportion of the category 'non-centralisation' determined by changes in distal symptom location; Other, marginal proportion of the category 'other' determined by changes in distal symptom location, W, marginal proportion of the category 'worse' determined by changes in distal symptom location and / or intensity.

Table 5.7 Agreement for individual categories of the symptom response classification.

	Fleiss (1971)	Roberts (2008)
(n=48)	κ	κ
CP_L	**0.94**	**0.40**
Non-CP_L	**0.36**	**0.68**
$Other_L$	**0.60**	**0.64**
(n=45)		
B	**0.47**	**0.50**
W	**0.10**	**0.08**
NC	**0.34**	**0.34**

Note: n= number of individuals with neck pain; B, category 'better' determined by changes in distal symptom location and / or intensity; CP_L, category 'centralisation' determined by changes in distal symptom location; κ, kappa statistic; NC, category 'no change' determined by changes in distal symptom location and / or intensity; Non - CP_L, category 'non-centralisation' determined by changes in distal symptom location; Other, category 'other' determined by changes in distal symptom location; W, category 'worse' determined by changes in distal symptom location and / or intensity.

Table 5.8 Patterns of agreement when using changes in distal symptom location.

Frequencies		Physiotherapist 2			
		CP_L	**Non-CP_L**	**$Other_L$**	**Total**
Physiotherapist 1	**CP_L**	4	3	3	10
	Non-CP_L	2	28	0	30
	$Other_L$	0	2	6	8
	Total	6	33	9	48

		CP_L	**Non-CP_L**	**$Other_L$**
Full model parameter estimates (95%CI)				
CP_L	κ_{CPL}=0.40 (0.06, 0.74)			
Non-CP_L	$\kappa_{CPL/NonCPL}$=0.52 (0.16, 0.89)	κ_{NonCPL} =0.68 (0.46, 0.90)		
$Other_L$	$\kappa_{CPL/OtherL}$ =-0.05 (-1.17, 1.05)	$\kappa_{NonCPL/OtherL}$=0.82 (0.59, 1.06)	κ_{OtherL} =0.64 (0.36, 0.93)	
Marginal proportion π	0.17	0.66	0.18	

Note: 95% CI, 95% confidence intervals; CP_L, category 'centralisation' determined by changes in distal symptom location; κ, kappa statistic; Non - CP_L, category 'non-centralisation' determined by changes in distal symptom location; Other, category 'other' determined by changes in distal symptom location.

Table 5.9 Patterns of agreement when using changes in distal symptom location and / or intensity.

Frequencies		Physiotherapist 2			
		B	**W**	**NC**	**Total**
Physiotherapist 1	**B**	16	7	0	23
	W	1	4	7	12
	NC	3	1	6	10
	Total	20	12	13	45

	B	**W**	**NC**
Full model parameter estimates (95%CI)			
B	κ_B =0.50 (0.25, 0.75)		
W	$\kappa_{B/W}$ =0.32 (-0.07, 0.71)	κ_W =0.08(-0.07, 0.23)	
NC	$\kappa_{B/NC}$=0.73 (0.44, 1.02)	$\kappa_{W/NC}$=-0.26 (-0.33, 0.85)	κ_{NC} =0.34 (0.03, 0.65)
Marginal proportion π	0.48	0.27	0.26

Note: 95% CI, 95% confidence intervals; B, category 'better' determined by changes in distal symptom location and / or intensity; κ, kappa statistic; NC, category 'no change' determined by changes in distal symptom location and / or intensity; W, category 'worse' determined by changes in distal symptom location and / or intensity.

Table 5.10 Potential sources of error when considering changes in distal symptom location.

#	Classification (1, 2)	Authors comments
1	CP_L, non-CP_L	Error in interpretation of pain diagram for CP (C)
2	non-CP_L, CP_L	Variation in symptom location in the starting position and performance of testing: (A, B)
3	non- CP_L, $Other_L$	Error in interpretation of pain diagram for nonCP (C)
4	CP_L, non- CP_L	Variation in performance of testing (B)
5	CP_L, $Other_L$	Variation in symptom location in the starting position and performance of testing: (A, B)
6	non- CP_L, CP_L	Error in interpretation of pain diagram for CP (C)
7	CP_L, non- CP_L	Variation in symptom location in the starting position and performance of testing: (A, B)
8	CP_L, $Other_L$	Variation in symptom location in the starting position and performance of testing: (A, B)
9	non- CP_L, $Other_L$	Variation in symptom location in the starting position and performance of testing: (A, B)
10	CP_L, $Other_L$	Variation in symptom location in the starting position and subsequent effect of testing: (A, B)

Note: A. Biologic variation in the same participant; B. Variability in performance; C. Variability in judgements. CP, centralisation phenomenon; CP_L , category 'centralisation' determined by changes in distal symptom location; Non - CP_L , category 'non-centralisation' determined by changes in distal symptom location; Other, category 'other' determined by changes in distal symptom location.

5.5 DISCUSSION

5.5.1 Reliability

5.5.1.1 Reliability of the overall symptom response classification

The reliability of the overall classification was substantial for changes in symptom location (79%, $\kappa = 0.66$), but fair (58%, $\kappa = 0.33$) when intensity of the distal symptom was added. As expected, agreement between individual pairs of physiotherapists was also greater for symptom location than location and / or intensity considerations. This is a reasonable finding considering that changes in the location of the distal symptoms are usually more difficult to elicit and thus, they are more stable over time.

When looking at the reliability between individual pairs of physiotherapists, agreement ranged from poor to excellent. Reliability was consistently greater for the pair of physiotherapists with prior experience and formal extensive training in symptom response assessment. Some studies using highly trained clinicians have reported high kappa values when investigating agreement in symptom response classifications in cervical or LBP patients (Razmjou et al., 2000; Kilpikoski et al., 2002; Clare et al., 2005). In addition, physiotherapists who have experience of using a classification system including symptom responses produced higher reliability values than physiotherapist pairs who had no experience of the system (Fritz et al., 2006). Such findings should be interpreted with caution though since no formal testing was applied and the sample size of examined participants was very small in the latter as well as this study.

5.5.1.2 Reliability of the individual symptom response categories

The reliability of classifying according to changes in distal symptom location (CP_L) was fair (κ = 0.40 [0.06, 0.74], Roberts' κ). Reliability was lower from reported values in a similar study in LBP, but prevalence of centralisation was also lower. Kappa takes lower values when there is substantial symmetrical imbalance in marginal distributions in the presence of high percentage agreement (Sim and Wright, 2005). This situation, called limited variation, makes kappa susceptible for prevalence bias (Thompson and Walter, 1988; Brennan and Silman, 1992) and results in poor reliability despite the lack of a substantial difference between measurements (Haas, 1991; Byrt et al., 1993). Kilpikoski et al. (2002) found that reliability in eliciting centralisation was κ = 0.7 for judgements in people whose low back symptoms centralised. However, these authors used highly trained clinicians for

the assessment of patients resulting in higher proportions of positive findings (87% of participants had their symptoms centralised).

Reliability when classifying according to favourable changes in the distal symptom location and / or intensity (B) was moderate (κ = 0.50 [0.25, 0.75], Roberts κ) but better than that observed for changes in distal symptom location. However, prevalence of favourable changes in the distal symptom location and / or intensity (B) was also higher (0.48 versus 0.17). Previous studies in patients with LBP have reported similar kappa coefficients; Kilby and associates (1990) found κ = 0.51 with respect to the question 'do any repeated movements decrease, abolish or centralise the pain?' This is also in agreement with a study investigating a similar construct (i.e. directional preference) who found moderate agreement among 54 clinicians (κ = 0.46) (Dionne et al., 2006).

5.5.2 Potential sources of error

The inter-class kappa coefficients for changes in distal symptom location indicated that the CP_L and $Other_L$ categories were confused most often, followed by confusion between the non-CP_L and CP_L categories. Analysis of the videotaped data pointed to some possible reasons for discrepancies between physiotherapists in these categories. In the first case, biologic variations in observed symptoms and performance differences accounted for these discrepancies (cases 5, 8 and 10 in Table 5.10). Centralisation has been associated with a lasting change in distal symptoms, although a uniform timeframe for observed changes remains unclear (Chapter 5). In the latter case (cases 1, 2, 4, 6, 7 in Table 5.10), a combination of interpretation errors as well as performance variations between physiotherapists was responsible for error in these

measurements, a finding which is consistent with common suggested sources of measurement error (Viera and Garrett, 2005).

5.5.3 Study limitations

5.5.3.1 Generalisability

Findings in this chapter relate more to occupational NP and office workers since data were collected through occupational health in a university setting. With results generalisable in this group, any conclusions regarding the whole NP population should be made with caution. In addition, due to restrictions in the available statistical approaches, inferences made in this chapter can only be relevant to participants and not raters (Gwet, 2008a). This is because current methods for kappa –type indices account only for the variance due to the sampling of participants and not of raters (Gwet, 2008a)[22]. Some argue that using non-specialist physiotherapists increases the generalisability of findings (Werneke et al., 2008). Procedures usually involved active range of movement testing rather than more advanced and forceful techniques. These procedures were considered relevant in a recent Delphi consensus study for eliciting centralisation, described previously (Chapter 4). Different results might have occurred if varying degrees of proficiency in such techniques applied (Snodgrass et al., 2006; Snodgrass et al., 2007). However, this chapter concentrated on the reliability of the measurement method itself. The effect of characteristics of raters in reliability is addressed elsewhere (Chapter 6).

[22] The use of raters as an additional source of sampling variability received attention only recently for nominal scale agreement statistics, and methods accounting for this are still preliminary (Gwet, 2008a).

5.5.3.2 Sample size

The main purpose of most reliability studies is to test agreement against a benchmark of minimum reliability rather than zero agreement (Maclure and Willett, 1987; Maher and Latimer, 1992; Hale and Fleiss, 1993). The final sample size has 80% power to detect a statistically significant value of 0.40 or higher. This formula is based on model-based inferences for the kappa statistic (Donner and Eliasziw, 1992). However, there was limited precision in this chapter's results indicated by the observed wide confidence intervals (CIs) in most cases. This may restrict confidence in conclusions made in this chapter (Akobeng, 2008). Sample size calculation was based on the p value rather than CIs. Although calculations with the aim to provide a confidence interval of a desired width have been proposed as a better approach to sample size estimation (Sim and Wright, 2005) unfortunately, little data exist regarding such methods (Nam, 2000; Bartfay and Donner, 2001; Zou and Klar, 2005).

Based on computer simulation studies (Cicchetti and Fleiss, 1977; Cicchetti, 1981), some propose a minimum minimal N = 20 when the number of categories is three[23] (Cicchetti and Sparrow, 1981). However, *p* values and CI are sensitive to sample size (Viera and Garrett, 2005), the calculation of which depends on expected prevalence and kappa values. A low prevalence of centralisation (17%) was observed in this chapter when changes in distal symptom location only were considered. Expected prevalence rates required for sample size estimation (50 - 70%) were based on

[23] This is based on an empirically based formula for determining the approximate minimal sample sizes (N) for the valid application of the kappa statistics. The N varies as a function of $2k^2$, in which k refers to the number of categories of classification on a given qualitative scale of measurement.

available data at the time. Subsequent published data are consistent with our findings[24], but these could not be used at the initial planning stages.

In this chapter, multiple sets of secondary reliability calculations were performed on various sets of examination data on the same participants (some being *post hoc* comparisons). This can potentially violate the statistical assumption of independence of observations (Schneider et al., 2008). The complex and intensive nature of the statistical methodology required in such cases is very prohibitive (Schneider et al., 2008). Therefore, the statistical analysis was limited to providing the κ values, raw percentages and CIs for the sets of observations. When formal testing was performed to determine if differences in κ values achieved statistical significance, it was uncertain whether this was adequately powerful to detect such differences. Results should therefore be interpreted with caution and in the appropriate context of these statistical limitations (Schneider et al., 2008).

5.5.4 Implications for practice and future research

Prior reliability investigations on classifications tend to favour the assessment of the overall agreement in clinicians' judgements. High kappa values on the overall classification system are then regarded as evidence supporting the use of that classification in clinical practice. Although this is a valid approach when interested in

[24] Werneke et al. (2008) found a first visit prevalence of 15% in a predominantly middle-aged neck and LBP population. Prevalence increased with acuity and younger age at the first visit (Werneke et al., 2008). These findings were lower than Werneke and Hart's original paper (Werneke and Hart, 2004) where a rate of 45% was found, but similar (17-18%) to a later study using a first visit definition in LBP (Schmidt et al., 2007). Fritz and Brennan (2007) found 34.7% centralisation prevalence but these authors included only participants with referred or radicular symptoms in contrast to this study which mainly involved symptoms above the elbow (only 25% had symptoms below the elbow).

the reliability of classification, it does not provide any insights on the agreement in incorporated categories between clinicians.

If two raters are asked to judge the presence of a relatively rare occurrence, the fact that they agree on the more frequently occurring negative cases is of little use if there is strong disagreement about which few cases are positive. When looking at the results of the overall classification and individual categories, two opposing trends are identified: reliability in *generally* classifying according to distal symptom location only is greater than classifications according to location and / or intensity of the distal symptom, but reliability when classifying according to *favourable* changes in the distal symptom location and / or intensity is higher than that observed for changes in distal symptom location only. This can be explained if we consider that reliability in judgements of the overall classification lumps together the agreement on each of the categories when in fact the agreement may differ for each category. For this reason, reliability information on both individual categories and the overall classification needs to be readily available to clinicians, with future reliability studies reporting such information in their results (MaClure and Willett, 1987; Haas, 1991; Armitage and Berry, 2002).

Reliability in classifying using clinically induced changes in symptoms ranged from fair to substantial. However, the reliability in the classification of some categories (e.g. CP_L, W) was slight to fair. Some may argue that this corresponds to inadequate figures for clinical use. However, poor reliability does not only reflect the quality of the measurement or observation procedure; it also reflects the nature of the investigated population (Kraemer, 1979; de Vet et al., 2003). Many cases presented

very mild or minimal symptomatology usually assessed through the reproduction of the participant's symptoms by sufficient force in spinal testing rather than a clinically induced decrease in symptoms. On the other hand, chronic presentations may have required more time for changes in distal symptom location to occur. Nevertheless, the observed low prevalence rates of centralisation based on changes in distal symptom location only indicate that a homogeneous group was analysed resulting in relatively low, but still acceptable kappa values (Kraemer, 1979). What would be regarded as fair reliability using suggested standards (Landis and Koch, 1977b) might actually reflect not error but low prevalence, and may be near optimal for such a population in a clinical setting (Kraemer, 1979), especially when used as part of a package of examination procedures (Wainner, 2003). In research applications though, one would need to compensate for low reliability by increasing sample size or by using multiple observations per subject, but these strategies are not always practical and valid unless the sources of confusion between categories are delineated (Kraemer, 1979).

In this chapter, the combined use and reporting of the inter-class and the intra-class kappa is advocated to gain further insights into strategies for improvement in reliability. The need for such methods depends of course on the context and objectives of the study (Roberts, 2008). If the aim of the study is primarily pragmatic (i.e. the actual results are more important than the underlying mechanism), conventional kappa statistics may be sufficient (Donner and Eliasziw, 1997; Roberts, 2008). Our strategy, when supplemented by the videotaped information and the data from the assessment forms indicated that improvements in reliability in the CP_L category might occur through additional training to improve the elicitation of

centralisation in the light of the low prevalence and the nature of measurement error sources. What should also be considered is that not only the skill of the physiotherapist, but also the choice of technique may affect the classification outcome (Egwu, 2008). Nevertheless, a carryover effect can never be ruled out with confidence[25] (Schneider et al., 2008). In the case of nonCP_L and $Other_L$ categories, consistency in interpretation should also be considered.

The relatively recent nature of developments in the field of reliability methodology means that no single source of information has developed statistical alternatives in great depth. The choice of a statistical method for reliability analysis can influence the results and consequently, the conclusions made in a study in some cases. This was evident when exploring the use of Fleiss' over Roberts' kappa coefficient for the agreement in individual categories when symptom location was considered but not when symptom location/intensity. Differences in these two methods could perhaps account for this but the issue of influence by category prevalence should also be considered when using such approaches.

5.6 SUMMARY

Although the usefulness of centralisation has been investigated extensively in LBP, its reliability and prognostic value are still unclear in NP. This chapter presented the first part of the pilot investigation aiming to assess the inter-rater reliability of the identification of clinically induced symptom responses based on the centralisation phenomenon in patients with NP. Following a mixed methodology approach, it was

[25] This means that the performance of the first set of examination procedures may alter the results of the second physiotherapist.

found that: (a) reliability in *generally* classifying according to distal symptom location only was greater than classifications according to location and / or intensity of the distal symptom, but reliability when classifying according to *favourable* changes in the distal symptom location and / or intensity was higher than that observed for changes in distal symptom location only; (b) reliability was consistently greater for the pair of physiotherapists with prior experience and formal extensive training in symptom response assessment. Caution should be expressed about this chapter's findings in the light of the limited precision in the reliability coefficients. Information on both individual categories and the overall classification needs to be readily available to clinicians, with future reliability studies reporting such information in their results. The combined use and reporting of the inter-class and the intra-class kappa is also advocated to gain further insights into strategies for improvement in reliability.

SECTION 4

THESIS CONCLUSIONS AND FUTURE RESEARCH RECOMMENDATIONS

CHAPTER 6

General Discussion

6.1 AIMS OF CHAPTER

This thesis focused on symptom centralisation. Its primary objectives were to establish a standard operational definition of centralisation and evaluate its inter-rater reliability in NP. This was done in the context of available evidence on the reliability and prognostic value of clinically induced symptom responses in non-specific spinal pain (Chapters 2 and 3) and the most recent NP guidelines (Childs et al., 2008; Haldeman, 2008).

The aims of this chapter are to:

- Present a summary of the main findings of this thesis and discuss how these contribute to existing knowledge;
- Identify and highlight the limitations of the research presented in this thesis and discuss implications;
- Provide directions for future research.

6.2 INTRODUCTION

Chapter 1 introduced NP and provided information on its course, impact, risk and prognostic factors, and current assessment and treatment approaches. NP is a common condition in the general population and employees of various professions; however, little is known about the factors that predict future outcomes or the best available management approach for NP.

The management of NP cases usually starts with history taking and a physical examination. There is great variability in the way that different professional disciplines

and countries assess these cases and thus, in the way that NP is labeled and classified. More than 300 case definitions have been reported by the Task Force on Neck Pain (Guzman et al., 2008) and 9 neck pain (NP) classification approaches are available utilising different assessment criteria and perspectives. Some NP classifications have been developed to determine incidence and prevalence in health surveys (Waris et al., 1979; Viikari-Juntura, 1983; McCormack et al., 1990), other classifications to identify an impairment or function-based diagnosis (Spitzer et al., 1995; WHO, 2004), and some to determine treatment (Wang et al., 2003; Childs et al., 2004b; McKenzie and May, 2006; Guzman et al., 2008). There is no classification determining prognosis in NP.

Variability in case definitions, classification approaches and assessment procedures has often been attributed to the lack of communication between disciplines, different schools of thought and health systems (Kent and Keating, 2004; Terrier, 2004; Kent et al., 2009). This gap is difficult to bridge in the absence of a shared robust evidence base. Despite the fact that clinicians primarily focus on pain and impairment measures to guide clinical reasoning (Kent et al., 2009), there is little evidence to support the use of these physical measures in clinical practice (de Koning et al., 2008b; Rubinstein et al., 2008). Lack of information on useful clinical markers that could aid the prediction of patient outcomes has been reported to cause frustration to some clinicians (Axen et al., 2005).

In Chapter 1, clinically induced symptom responses were introduced. Clinically induced symptom responses, defined as immediate changes in the status of symptoms resulting from physical examination or intervention strategies, are frequently used to assess the underlying acuity and nature of the condition, establish a diagnosis, guide prognosis or

determine management strategies. Traditionally, clinically induced symptom responses have focused on provoking or altering symptoms with a variety of spine loading strategies. Again, there are different schools of thought about what symptom responses are important, depending on the system using them, and it is currently unknown what clinically induced responses can reliably predict spinal outcomes.

The focus of this thesis was on a commonly cited clinically induced symptom response, centralisation. In the UK, centralisation has been considered an important low back pain (LBP) clinical examination item (McCarthy et al., 2006). Although centralisation has been established as a useful and important physical sign in LBP, its definition, reliability and prognostic value is still unclear in NP.

6.3 MAIN FINDINGS AND CONTRIBUTION TO THE LITERATURE

Despite the recent publication of the NP guidelines (Childs et al., 2008; Haldeman, 2008), there is still no clear guidance on the operational definition of centralisation and its role in NP. Thus, the work done in producing an operational definition and testing for reliability in this thesis is based on the most contemporary information available. A number of principles, described in Chapter 1, were used to develop the operational definition of centralisation. These principles were set to ensure that the steps followed to develop definition of centralisation are in accordance with high current standards of test development.

6.3.1 Systematic reviews

The systematic reviews in Chapters 2 and 3 provided a comprehensive, up-to-date, objective and reliable overview of available evidence in spinal pain. They formed the basis on which subsequent research plans on centralisation were made. Clear and consistent measurements of potential prognostic and treatment indicators are required in research and clinical practice (Simon and Altman, 1994; Fritz and Wainner, 2001; Beattie and Nelson, 2007). Chapter 2 focused on studies investigating reliability of a range of clinically induced symptom responses. Previous attempts to summarise evidence on the reliability of various assessment procedures had been made, with most systematic reviews supporting the use of pain-related tests in the physical examination of NP and LBP over other assessment procedures (Seffinger et al., 2004; May et al., 2006; Hestboek and Leboeuf-Yde, 2000; van der Wurff et al, 2000; Stochkendahl et al., 2006). However, this was the first time that a comprehensive account of all the literature on clinically induced symptom responses was brought together without any restrictions to a spinal area or to any professional discipline.

Thirty-six studies were found evaluating the reliability of clinically induced symptom responses in the physical examination of spinal pain. Only six out of the thirty-six studies involved patients with NP; the remaining studies recruited patients with LBP. Symptom reproduction or changes in the intensity and / or location of symptoms in response to gross and segmental movement testing, palpation, non-organic signs, and neural testing were investigated most frequently. The clinically induced symptom responses demonstrating adequate reliability, and areas of improvement for the quality of research on the inter-rater reliability of these assessment procedures, were identified.

Most studies provided summary statistics within the context of independent preliminary investigations, but none explicitly explored reasons for inadequate reliability values or proposed strategies for improving reliability. This is a particularly important observation because findings from such studies may lead to the premature exclusion of potentially useful clinically induced symptom responses or the promotion of highly reliable, but clinically meaningless tests (Fritz and Wainner, 2001). Thus, attempts to determine and deal with potential sources of error were deemed necessary to improve the future use of clinically induced symptom responses.

Two commonly used thresholds of acceptable reliability (0.4 and 0.7) and methodological quality (0.5 and 0.6) were applied. Therefore, it was possible to explore the impact of different, but previously used, quality and reliability thresholds on the review conclusions. Changes in the threshold of acceptable reliability affected conclusions in both the lumbar spine, sacro-iliac joints (21 / 99) and the cervical spine (3 / 11). Lowering the threshold for adequate methodological quality from 60% to 50% had a smaller impact, and shifted levels of evidence from limited to moderate in 4 sacro-iliac procedures and 1 lumbar procedure. As expected, the selection of different reliability and study quality thresholds influenced the review conclusions but in this review, it was possible to identify where past disagreements may lie. This is an important addition to the existing literature, because the use of different methodologies and thresholds had previously resulted in variable conclusions about the same procedures. This may also attract more attention to the necessity of uniform and well-informed decisions when summarising and systematically appraising the literature.

The value of a test can not rely exclusively on reliability measures (Fritz and Wainner, 2001). In Chapter 3, the aim was to provide a comprehensive review of the quality of research on, and investigate the prognostic value of clinically induced responses in the conservative management of spinal pain. With the literature expanding considerably since the publication of previous reviews, and with articles reporting new information on clinically induced symptom responses, a systematic review of available evidence was deemed necessary. A systematic and reasoned approach was followed in the methodology, from the point of identifying relevant studies to appraising and analysing findings. In contrast to previous narrative or critical reviews of the literature (Wetzel and Donelson, 2003; Aina et al., 2004; Wessels et al., 2006; Berthelot et al., 2007), this was the first attempt to formally synthesise prognostic information on a much wider range of symptom responses. It was concluded that distal symptom changes in response to repeated movement testing and / or as a response to treatment offer promise as a predictor of LBP and NP outcomes, but further investigation is still required, particularly for longer term prediction across a range of outcomes.

6.3.2 Delphi study

Chapter 4 provided a definition of centralisation. This is the first study aiming to establish consensus on centralisation internationally among researchers and clinicians. This is also the first formal attempt to stimulate an in-depth discussion between different schools of thought on the issue centralisation and provide common directions for future research.

The Delphi panel included experts from various settings, countries and professional disciplines. Representatives were brought together from clinical practice, research and education, reflecting a range of experiences in using centralisation within the context of various health systems. The aim was to capture a wide spectrum of perspectives from a group of knowledgeable but also influential individuals with the ability to implement findings in clinical practice and research.

Centralisation was defined as the progressive and stable reduction of the most distal presenting pain towards the spine midline in response to standardised spinal loading strategies. Assessment in response to a single testing of active range of motion was considered inadequate by the Delphi panel. This is in agreement with the findings of the systematic literature reviews (Chapters 2 and 3). Although there was evidence supporting the reliability of changes in symptom location and / or intensity in response to single movements (strong evidence, $\kappa \geq 0.7$), their value in predicting outcomes was not demonstrated in NP.

Findings from Chapter 4 may also serve another purpose. They can be used by researchers who are considering studies on the centralisation phenomenon. In NP, the investigation of centralisation was voted as one of the most popular questions for future research by the Delphi panel. Indeed, moderate evidence supporting reliability of clinically induced symptom responses to procedures other than single movements has only been found for directional preference ($\kappa \geq 0.4$) (Chapter 2), and limited evidence from a small study ($n = 29$) only supports an association of treatment induced changes in pain location or intensity with neck symptoms in the short-term. There is hardly any

information on the reliability and role of centralisation in predicting long-term outcomes in NP. Other preliminary investigations have indicated the potential usefulness of this sign as a prognostic factor but these studies have included participants with both NP and LBP (Werneke et al, 1999; May et al., 2008a).

6.3.2 Reliability study

One of the reservations of potential participants in the Delphi study was that superiority of and preference for a particular definition should be demonstrated through research reports rather than anecdotal opinions. Indeed, current evidence-based practice requires that testing is sufficiently accurate and precise to allow clinicians to make a correct inference about the patient's condition (Rothstein et al., 2001; Cleland et al., 2008). Thus, after establishing consensus, the reliability of the developed definition had to be assessed and some elements of the operational definition tested against other suggested in the literature.

Chapter 5 involved an inter-rater reliability study on centralisation in NP. Agreement in the overall classification as well as individual symptom response categories involving the centralisation phenomenon was assessed. Five physiotherapists took part as the examiners for a reliability analysis after approximately 3 hours of training on eliciting and interpreting centralisation. A practical strategy of forming physiotherapist pairs independently for each participant from a large pool of raters rather than using the same two raters for the whole sample was followed. Experience in using centralisation or a relevant system was not a prerequisite for being an examiner, following relevant

recommendations by the Delphi panel (Chapter 4). The above steps were followed to increase external validity of findings. Nevertheless, reliability was consistently greater for the pair of physiotherapists with prior experience and formal extensive training in symptom response assessment.

The reliability of the overall symptom response classification was substantial for changes in symptom location (79%, $\kappa = 0.66$), but fair (58%, $\kappa = 0.33$) when intensity of the distal symptom was added. On the other hand, the inter-rater reliability of the category 'centralisation' was fair ($\kappa = 0.40$ [0.06, 0.74], Roberts' κ) when based on changes in distal symptom location, and moderate ($\kappa = 0.50$ [0.25, 0.75], Roberts κ) when changes in distal symptom location and / or intensity were considered (this refers to the category 'better'). Thus, two opposing trends were identified: reliability in *generally* classifying according to distal symptom location only was greater than classifications according to location and / or intensity of the distal symptom, but reliability when classifying according to *favourable* changes in the distal symptom location and / or intensity was higher than that observed for changes in distal symptom location only. Again, recommendations of the Delphi panel on the importance of including changes in distal symptom intensity rather than exclusively relying on changes in distal symptom location when considering the definition of centralisation were confirmed by the findings of this chapter.

The study in Chapter 5 introduced some novel methods of studying inter-rater reliability. For the first time, information was not only provided on the degree of agreement between physiotherapists in a reliability investigation but also on patterns of agreement

between different categories using simple kappa-type measures. In contrast to alternative methods such as the log-linear or latent class models, kappa-type coefficients were easier to interpret, and therefore are more likely to be adopted by researchers and clinicians in the field.

Very few factors that could affect measurements are usually described in reliability papers (Van Genderen et al., 2003) and there is also confusion around training issues and where they should be directed (Chapter 4). Information from the videotaped assessments were thus used and combined with recorded information in the assessment forms to identify possible reasons for variability in the physiotherapists' symptom response classifications. To the author's knowledge, only one study has used videotaped information in the past to explore potential sources of measurement error in an investigation of the reliability of range of motion measurements (Bush et al., 2000). It was possible to identify that training efforts aiming to improve reliability in identifying centralisation should target not only interpretation errors but also the skill of the physiotherapist in selecting the appropriate technique and in eliciting centralisation.

6.4 LIMITATIONS OF RESEARCH AND RECOMMENDATIONS FOR FUTURE RESEARCH

This thesis focused on centralisation and the development of a standard definition in NP. The operational definition of centralisation was developed in agreement with best practice for test development and research models intended to identify and validate subgroups in non-specific spinal syndromes. Evidence-based guidelines and relevant

literature as well as experts' opinions were considered at the initial stages of the development of the definition. Many issues around the operational definition of centralisation were raised and the most important gaps of knowledge were identified, agreed and prioritised. However, few of these issues and gaps could be addressed within the restricted timeframe and resources of this PhD project.

Nevertheless, documentation to a standard that promoted consistency and enabled replication of the testing procedure was sought when investigating the inter-rater reliability of the identification of centralisation and related symptom response classification in patients with NP. The support by the Delphi panel of a broader definition allowed for a multitude of different ways of testing to be included in the assessment procedure. This approach may offer some flexibility to clinicians assessing, classifying and managing different spinal pain presentations across different countries and is quite common in clinical practice, with the same therapist often combining testing and treatment pathways instead of exclusively following one only (Battié et al., 1994; Jette et al., 1994; Pinto et al., 2007). However, it may also lead to variations in testing which, depending on the prevalent health system of a country and as shown in Chapter 5, could account for discrepancies when classifying individuals with NP. Thus, the effect of procedural variations in testing on the outcomes of individuals who centralise requires urgent investigation before this broad definition of centralisation is adopted in clinical practice.

The operational definition in this thesis was developed with the view to be delivered within the context of current practice in terms of staffing and time. Training of

participating physiotherapists was short (3 hours) and did not necessarily involve clinicians with experience in eliciting and interpreting centralisation. Although the reliability of identifying centralisation was acceptable ($\kappa \geq 0.4$ when considering changes in distal symptom location, or changes in distal symptom location and / or intensity), high levels ($\kappa > 0.7$) of inter-rater reliability were not reported in Chapter 5, shown previously in studies using highly trained clinicians. Assessment times were also longer than what would normally be expected in the busy day-to-day clinical environment. However, reliability was greater for the pair of physiotherapists with prior experience and formal extensive training in symptom response assessment. Studies to date have not consistently concluded on the elements that lead to highly reliable tests and one reason may be the absence of formal statistical testing and the small sample sizes of examined participants per homogeneous expertise or experience of physiotherapists. Therefore, the optimal type and amount of training for eliciting and interpreting centralisation needs to be determined both in terms of its effect on patient outcomes and in terms of practicality and use for future studies in musculoskeletal therapy.

There is no widely agreed classification determining prognosis and treatment in NP. This thesis focused on establishing an operational definition of centralisation and evaluated its inter-rater reliability in NP, but this is only a preliminary step to meeting the goal of a standard definition of centralisation. The prognostic and management value of centralisation are still unknown, and in view of the significant amount of interest in sub-grouping in the clinical literature (Borkan et al., 2002; Childs et al., 2008), these issues need to be clarified internationally in the context of varying health systems and

classification trends. However, it is hoped that this investigation may pave the way for the standardisation of centralisation as a physical sign and stimulate interest for further study of potential sub-groups and classification of spinal syndromes.

6.5 SUMMARY

The aims of this chapter were to present the main findings of this thesis, discuss limitations of the research presented in this thesis and provide directions for future research.

This thesis focused on symptom centralisation. Its primary objectives were to establish a standard operational definition of centralisation and evaluate its inter-rater reliability in NP. The systematic reviews in Chapters 2 and 3 showed that the potential usefulness of centralisation has been demonstrated in LBP, however, concern has been expressed about the observed inconsistency in reported LBP definitions, and the scarcity of studies in NP. In a Delphi study (Chapter 4), centralisation was defined as the progressive and stable reduction of the most distal presenting pain towards the spine midline in response to standardised spinal loading strategies. The support by the Delphi panel of a broader definition allowed for a multitude of different ways of testing to be included in the assessment procedure which may offer some flexibility to clinicians assessing, classifying and managing different spinal pain presentations across different countries. . Although the reliability of identifying centralisation was acceptable, high levels of reliability were not demonstrated (Chapter 5). Thus, this thesis has indicated the urgent need for the optimal type and amount of training for eliciting and interpreting centralisation and the effect of procedural variations in testing on the outcomes of

individuals who centralise to be clarified. The prognostic and management value of centralisation are also unknown. This investigation may pave the way for the standardisation of centralisation as a physical sign and stimulate interest for further study of potential sub-groups and classification of spinal syndromes.

REFERENCES

Abbott, J., Flynn, T., Fritz, J., Hing, W., Reid, D., Whitman, J. (2009) Manual physical assessment of spinal segmental motion: intent and validity. *Manual Therapy,* 14, 36-44.

Abenhaim, L., Rossignol, M., Valat, J., Nordin, M., Avouac, B., Blotman, F., Charlot, J., Dreiser, R., Legrand, E., Rozenberg, S., Vautravers, P. (2000) The role of activity in the therapeutic management of back pain. Report of the International Paris Task Force on Back Pain. *Spine,* 25, S1-33.

Abraham, I., Killackey-Jones, B. (2002) Lack of evidence based research for idiopathic low back pain. *Archives of Internal Medicine,* 162, 142-144.

Abrams, D., Davidson, M., Harrick, J., Harcourt, P., Zylinski, M., Clancy, J. (2006) Monitoring the change: current trends in outcome measure usage in physiotherapy. *Manual Therapy,* 11, 46-53.

Agency For Health Care Policy And Research (AHCPR) (1994) *Acute low back problems in adults.* USA, Department of Health and Human Services. Available at:

http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=hstat6.chapter.25870#top [1st April 2009].

AGREE Collaboration (2003) *Appraisal of guidelines for research and evaluation: AGREE instrument manual.* Available at:

http://www.agreecollaboration.org/pdf/aitraining.pdf [21st May 2009].

Agresti, A. (1988) A model for agreement between ratings on an ordinal scale. *Biometrics,* 44, 539-548.

Agresti, A. (2002) *Categorical data analysis*. 2nd edition. New York, Wiley.

Agresti, A. (1992) Modelling patterns of agreement and disagreement. *Statistical Methods in Medical Research,* 1, 201-218.

Aina, A., May, S., Clare, H. (2004) The centralization phenomenon of spinal symptoms - a systematic review. *Manual Therapy,* 9, 134-143.

Akins, R., Tolson, H., Cole, B. (2005) Stability of response characteristics of a Delphi panel: application of bootstrap data expansion. *BMC Medical Research Methodology,* 5**,** 37.

Akobeng, A. (2008) Confidence intervals and p-values in clinical decision making. *Acta Paediatrica,* 97, 1004-1007.

Albright, J., Allman, R., Bonfiglio, R., Conill, A., Dobkin, B., Guccionne, A., Hasson, S., Russo, R., Shekelle, P., Susman, J., Ottawa Methods Group, Brosseau, L., Tugwell, P., Wells, G., Robinson, V., Graham, I., Shea, B., Mcgowan, J.,

Peterson, J., Tousignant, M., Poulin, L., Corriveau, H., Morin, M., Pelland, L., Laferriere, L., Casimiro, L., Trembray, L. (2001) Philadelphia Panel evidence-based clinical practice guidelines on selected rehabilitation interventions for low back pain. *Physical Therapy,* 81, 1641-1674.

Aldman, B., Gustavson, H., Nygren, A., Tinqvall, C. (1987) Child restraints: A prospective study of children as car passengers in road traffic accidents with respect to restraint effectiveness. *Acta Paediatrica Scandinavica,* 339, S1-22.

Alexandre, N. (2002) Predictors of compliance with short-term treatment among patients with back pain. *Pan American Journal of Public Health*, 12, 86-94.

Altman, D. (1991) *Practical statistics for medical research.* London, Chapman & Hall.

Altman, D. (2001) Systematic reviews of evaluations of prognostic variables. *British Medical Journal,* 323, 224-228.

Altman, D., Bland, J. (1983) Measurement in medicine: the analysis of method comparison studies. *The Statistician,* 32, 307-317.

Altman, D., Lyman, G. (1998) Methodological challenges in the evaluation of prognostic factors in breast cancer. *Breast Cancer Research and Treatment,* 52, 289-303.

Altman, D., Royston, P. (2006) The cost of dichotomising continuous variables. *British Medical Journal,* 332, 1080.

American Physical Therapy Association (APTA) (2001) Guide to Physical Therapist Practice, 2nd edition. *Physical Therapy,* 81, 9-746.

Anonymous (1966) Pain in the neck and arm: a multicentre trial of the effects of physiotherapy. *British Medical Journal*, 1, 253-258.

Ariens, G., Bongers, P., Douwes, M., Miedema, M., Hoogendoorn, W., Van Der Wal, G., Bouter, L., Van Mechelen, W. (2001) Are neck flexion, neck rotation, and sitting at work risk factors for neck pain? Results of a prospective cohort study. *Occupational and Environmental Medicine,* 58, 200-207.

Armitage, P., Berry, G. (2002) *Statistical Methods in Medical Research.* Oxford, Blackwell Science.

Armstrong, G. (1981) The intraclass correlation as a measure of interrater reliability of subjective judgements. *Nursing Research,* 30, 314-320.

Arnau, J., Vallano, A., Lopez, A., Pellise, F., Delgado, M., Prat, N. (2006) A critical review of guidelines for low back pain treatment. *European Spine Journal,* 15, 543-553.

Atlas, S., Deyo, R., Patrick, D., Convery, K., Keller, R., Singer, D. (1996) The Quebec Task Force classification for Spinal Disorders and the severity, treatment, and outcomes of sciatica and lumbar spinal stenosis. *Spine*, 21, 2885-2892.

Axen, I., Rosenbaum, A., Robech, R., Wren, T., Leboeuf-Yde, T. (2002) Can patient reactions to the first chiropractic treatment predict early favorable treatment outcome in persistent low back pain? *Journal of Manipulative and Physiological Therapeutics, 2*5, 450-454.

Axen, I., Jones J., Rosenbaum, A., Lovgren, P., Halasz, L., Larsen, K., Leboeuf-Yde, C. (2005a) The Nordic back pain subpopulation program: Validation and improvement of a predictive model for treatment outcome in patients with low back pain receiving chiropractic treatment. *Journal of Manipulative and Physiological Therapeutics*, 28, 381-385.

Axen, I., Rosenbaum, A., Robech, R., Larsen. K., Leboeuf-Yde, C. (2005b) The Nordic back pain subpopulation program: Can patient reactions to the first chiropractic treatment predict early favorable treatment outcome in nonpersistent low back pain? *Journal of Manipulative and Physiological Therapeutics*, 28, 153-158.

Baddeley, A. (1994) The Magical Number 7 - still magic after all these years. *Psychological Review*, 101, 353-356.

Baker, J., Lovell, K., Harris, N. (2006) How expert are the experts: An exploration of the concept of 'expert' within Delphi panel techniques. *Nurse Researcher*, 14, 59-70.

Bakkalbasi, N., Bauer, K., Glover, J., Wang, L. (2006) Three options for citation tracking: Google Scholar, Scopus and Web of Science. *Biomedical Digital Libraries*, 3, 7.

Banerjee, M. (1999) Beyond Kappa: a review of interrater agreement measures. *Canadian Journal of Statistics*, 27, 3-23.

Barker, M. (1990) A practical classification of spinal pains based on a study of patients seen in British general practice over a five-year period. In: Fairbank, J., Pynsent, P. (eds) *Back pain: classification of syndromes*. Manchester University Press, Manchester, 23-36.

Bartelson, B., Whinney, S. (2009) *Statistical Quick Reference Series Interrater reliability HIV/AIDS.* Available at:

www.uchsc.edu/ccfar/cores/QR04%20Interrater%20Reliability.doc [6th April 2009].

Bartfay, E., Donner, A. (2001) Statistical inferences for interobserver agreement studies with nominal outcome data. *The Statistician*, 50, 135-146.

Battie, M., Cherkin, D., Dunn, R., Clol, M., Wheeler, K. (1994) Managing low back pain: Attitudes and treatment preferences of physical therapists. *Physical Therapy*, 74, 219-226.

Beal, M. (1984) Palpatory examination of the spine: a comparison of the two methods and their relationship to visceral disease. *Manual Medicine,* 1, 25-32.

Becker, M., Agresti, A. (1992) Log-linear modelling of pairwise interobserver agreement on a categorical scale. *Statistics in Medicine,* 11, 101-114.

Bendebba, M., Torgerson, W., Long, D. (2000) A validated, practical classification procedure for many persistent low back pain patients. *Pain,* 87, 89-97.

Bekkering, G., Hendriks, H., Van Tulder, M., Knol, D., Simmonds, M., Oostendorp, R., Bouter, L. (2005) Prognostic factors for low back pain in patients referred for physiotherapy: Comparing outcomes and varying modelling techniques. *Spine,* 30, 1881-1886.

Bendix, A., Bendix, T., Haestrup, C. (1998) Can it be predicted which patients with chronic low back pain should be offered tertiary rehabilitation in a functional restoration program? A search for demographic, socioeconomic and physical predictors. *Spine,* 23, 1775-1783.

Benyamin, R., Singh, V., Parr, A., Conn, A., Diwan, S., Abdi, S. (2009) Systematic review of the effectiveness of cervical epidurals in the management of chronic neck pain. *Pain Physician,* 12, 137-157.

Berglund, A., Alfredsson, L., Cassidy, J., Jensen, I., Nygren, A. (2000) The association between exposure to rear-end collision and future neck or shoulder pain: a cohort study. *Journal of Clinical Epidemiology,* 53, 1089-1094.

Berglund, A., Alfredsson, L., Jensen, I., Bodin, L., Nygren, A. (2003) Occupant- and crash-related factors associated with the risk of whiplash injury. *Annals of Epidemiology,* 13, 66-72.

Berglund, A., Bodin, L., Jensen, I., Wiklund, A., Alfredson, L. (2006) The influence of prognostic factors on neck pain intensity, disability, anxiety and depression over a 2-year period in subjects with acute whiplash injury. *Pain,* 125, 244-256.

Bergstrom, G., Bodin, L., Jensen, I., Linton, S., Nygren, A. (2001a) Long-term, non-specific spinal pain: reliable and valid subgroups of patients. *Behavioural Research Therapy,* 27, 1783-1789.

Bergstrom, G., Jensen, I., Bodin, L., Linton, S., Nygren, A. (2001b) The impact of psychologically different patient groups on outcome after a vocational rehabilitation program for long-term spinal pain patients. *Pain,* 93, 229-237.

Berlin, J., University Of Pennsylvania Meta-Analysis Blinding Study Group (1997) Does blinding of readers affect the results of meta-analyses? *Lancet,* 350, 185-186.

Berthelot, J., Delecrin, J., Maugars, Y., Passuti, N. (2007) Contribution of centralisation phenomenon to the diagnosis, prognosis, and treatment of diskogenic low back pain. *Joint, Bone, Spine,* 74, 319-323.

Bertilson, B., Bring, J., Sjoblom, A., Sundell, K., Strender, L. (2006) Inter-examiner reliability in the assessment of low back pain (LBP) using the Kirkaldy-Willis classification (KWC). *European Spine Journal*, 15, 1695-1703.

Bertilson, B., Grunnesjo, M., Strender, L-E. (2003) Reliability of clinical tests in the assessment of patients with neck/shoulder problems – impact of history. *Spine,* 28, 2222-2231.

Bialosky, J., Bishop, M., Price, D., Robinson, M., George, S. (2008) The mechanisms of manual therapy in the treatment of musculoskeletal pain: a comprehensive model. *Manual Therapy*, 1-8.

Billis, E., McCarthy, C., Oldham, J. (2007) Subclassification of low back pain: a cross-country comparison. *European Spine Journal*, 16, 865-879.

Binkley, J., Finch, E., Hall, J., Black, T., Gowland, C. (1993) Diagnostic classification of patients with low back pain: report on a survey of physical therapy experts. *Physical Therapy,* 73, 138-155.

Binkley, J., Stratford, P., Gill, C. (1995) Interrater reliability of lumbar accessory motion mobility testing. *Physical Therapy,* 75, 786-792.

Black, N. (1994) Appropriateness of medical care in Europe: a commentary. *International Journal of Quality in Health Care*, 6, 231–232.

Bland, J., Altman, D. (1986) Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, i, 307-310.

Bloch, D., Kraemer, H. (1989) 2 x 2 kappa coefficients: measures of agreement or association. *Biometrics,* 45, 269-287.

Boersma, K., Linton, S. (2006) Expectancy, fear and pain in the prediction of chronic pain and disability: A prospective analysis. *European Journal of Pain*, 10, 551-557.

Boline, P., Haas, M., Meyer, J., Kassak, K., Nelson, C., Keating, J., Jr. (1993) Interexaminer reliability of eight evaluative dimensions of lumbar segmental abnormality: Part II. *Journal of Manipulative and Physiological Therapeutics,* 16, 363-374.

Boline, P., Keating, J. Brist, J., Denver, G. (1988) Interexaminer reliability of palpatory evaluations of the lumbar spine. *American Journal of Chiropractic Medicine,* 1, 5-11.

Borge, J., Leboeuf-Yde, C., Lothe, J. (2001) Prognostic values of physical examination findings in patients with chronic low back pain treated conservatively: a systematic literature review. *Journal of Manipulative and Physiological Therapeutics,* 24, 292-295.

Borghouts, J., Janssen, H., Boes, B., Muris, J., Metsemakers, J., Bouter, L. (1999a) The management of chronic neck pain in general practice. *Scandinavian Journal of Primary Health Care*, 17, 215-220.

Borghouts, J., Koes, B., Bouter, L. (1998) The clinical course and prognostic factors of non-specific neck pain: a systematic review. *Pain*, 77, 1-13.

Borghouts, J., Koes, B., Vondeling, H., Bouter, L. (1999b) Cost-of-illness of neck pain in the Netherlands in 1996. *Pain*, 80, 629-636.

Borkan, J., Koes, B., Reis, S., Cherkin, D. (1998) A report from the Second International Forum for Primary Care Research on Low Back Pain. Reexamining priorities. *Spine*, 23, 1992-1996.

Bossuyt, P., Reitsma, J., Bruns, D., Gatsonis, C., Glasziou, P., Irwig, L., Lijmer, J., Moher, D., Rennie, D., De Vet, H. (2003) Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD Initiative. *Clinical Chemistry*, 49**,** 1-6.

Bostick, G., Ferrari, R., Carroll, L., Russell, A., Buchbinder, R., Krawciw, D., Gross, D. (2009) A population-based survey of beliefs about neck pain from whiplash injury, work-related neck pain, and work-related upper extremity pain. *European Journal of Pain*, 13, 300-304.

Bot, S., Van Der Waal, J., Terwee, C., Van Der Windt, D., Scholten, R., Bouter, L., Dekker, J. (2005) Predictors of outcome in neck and shoulder symptoms: a cohort study in general practice. *Spine*, 30, E459-470.

Bouter, L., Van Tudler, M., Koes, B. (1998) Methodologic issues in low back pain research in primary care. *Spine*, 23, 2014-2020.

Boyd, N., Wolfson, C., Moskowitz, M., Carlile, T., Petitclerc, M., Ferri H., Fishell, E., Gregoire, A., Kiernan, M., Longley, J., Simor, I., Miller, A. (1982) Observer variation in the interpretation of xeromammograms. *Journal of the National Cancer Institute*, 68, 357-363.

Boyd, R., Massey, R., Duanne, L., Yates, D. (2002) Whiplash associated disorder in children attending the emergency department. *Emergency Medicine Journal*, 19, 311-314.

Bradish, C., Lloyd G., Aldam, C., Albert, J., Dyson, P., Doxey, N., Mitson, G. (1988) Do non organic signs help to predict the return to activity of patients with low back pain? *Spine*, 13, 557-560.

Brandt, L., Andersen, J., Lassen, C., Kryger, A., Overgaard, E., Vilstrup, I., Mikkelsen, S. (2004) Neck and shoulder symptoms and disorders among Danish computer workers. *Scandinavian Journal of Work, Environment and Health*, 30, 399-409.

Brennan, G., Fritz, J., Hunter, S., Thackeray, A., Delitto, A., Erhard, R. (2006) Identifying subgroups of patients with acute/subacute "nonspecific" low back pain. *Spine*, 31, 623-631.

Brennan, G., Parent, E. (2006) Re: Delphi study on centralization phenomenon. Email sent from Angeliki Chorti (physio_glaraki@yahoo.com), 5th December 2006.

Brennan, P., Silman, A. (1992) Statistical methods for assessing observer variability in clinical measures. *British Medical Journal*, 304, 1491-1494.

British Pain Society (2008) *Pain scales in multiple languages.* Available at: http://www.britishpainsociety.org/pub_pain_scales.htm [5th January 2009].

Browder, D., Childs, J., Cleland, J., Fritz, J. (2007) Effectiveness of an extension-oriented treatment approach in a subgroup of subjects with low back pain: a randomized clinical trial. *Physical Therapy*, 87, 1608-1618.

Bryner, P. (1996) Analysis of low back pain drawings: a critique. *Topics in Clinical Chiropractic*, 3, 26-31.

Buchbinder, R., Goel, V., Bombardier, C., Hogg-Johnson, S. (1996) Classification systems of soft tissue disorders of the neck and upper limb: do they satisfy methodological guidelines? *Journal of Clinical Epidemiology*, 49, 141-149.

Burdorf, A., Naaktgeboren, B., Post, W. (1998) Prognostic factors for musculoskeletal sickness absence and return to work among welders and metal workers. *Occupational and Environmental Medicine*, 55, 490-495.

Burton, A., Mcclune, T., Clarke, R., Main, C. (2004) Long-term follow-up of patients with low back pain attending for manipulative care: outcomes and predictors. *Manual Therapy*, 9, 30-35.

Burton, A., Tillotson, K. (1991) Prediction of the clinical course of low-back trouble using multivariate models. *Spine*, 16, 7-14.

Burton, A., Tillotson, K., Main, C., Hollis, S. (1995) Psychosocial predictors of outcome in acute and subchronic low back trouble. *Spine*, 20, 722-728.

Busanich, B., Verscheure, S. (2006) Does McKenzie therapy improve outcomes for back pain? *Journal of Athletic Training*, 41, 117-119.

Bush, K., Collins, N., Portman, L., Tillett, N. (2000) Validity and intertester reliability of cervical range of motion using inclinometer measurements. *Journal of Manual and Manipulative Therapy*, 8, 52-61.

Bybee, R. (2006a) Re: About your article in Manuelle Therapie. Email sent to Angeliki Chorti (physio_glaraki@yahoo.com), 9th August 2006.

Bybee, R. (2006b) Re: About your article in Manuelle Therapie. Email sent to Angeliki Chorti (physio_glaraki@yahoo.com), 7th November 2006.

Bybee, R., Hipple, L., Mcconnell, R., Crossland, P. (2005) The relationship between reported pain during movement and centralization of symptoms in low back pain patients. *Manuelle Therapie*, 9, 122-127.

Byrne, K., Doody, C., Hurley, D. (2006) Exercise therapy for low back pain: a small scale exploratory survey of current physiotherapy practice in the Republic of Ireland acute hospital setting. *Manual Therapy*, 11, 272-278.

Byrt, T., Bishop, J., Carlin, J. (1993) Bias, prevalence and kappa. *Journal of Clinical Epidemiology*, 46, 423-429.

Cailliet, R. (1988) Clinical application of low back mechanics in the diagnosis and treatment of pain syndromes. In: *Low back pain syndrome.* 2nd edition. Philadelphia, FA Davis, 49-57.

Carey, T., Garrett, J., Jackman, A. (2000) Beyond the good prognosis. Examination of an inception cohort of patients with chronic low back pain. *Spine*, 25, 115-120.

Carnes, D., Underwood, M. (2007) Chronic musculoskeletal pain. *British Journal of General Practice*, 57, 604-605.

Carroll, L., Hogg-Johnson, S., Van Der Velde, G., Haldeman, S., Holm, L., Carragee, E., Hurwitz, E., Cote, P., Nordin, M., Peloso, P., Guzman, J., Cassidy, J. (2008a) Course and prognostic factors for neck pain in the general population: results of the Bone and Joint Decade 2000-2010 Task Force on Neck Pain and Its Associated Disorders. *Spine*, 33, S75-82.

Carroll, L., Hogg-Johnson, S., Cote, P., Van Der Velde, G., Holm, L., Carragee, E., Hurwitz, E., Peloso, P., Cassidy, J., Guzman, J., Nordin, M., Haldeman, S. (2008b) Course and prognostic factors for neck pain in workers: results of the Bone and Joint Decade 2000-2010 Task Force on Neck Pain and Its Associated Disorders. *Spine*, 33, S93-100.

Carroll, L., Holm, L., Hogg-Johnson, S., Cote, P., Cassidy, J., Haldeman, S., Nordin, M., Hurwitz, E., Carragee, E., Van Der Velde, G., Peloso, P., Guzman, J. (2008c) Course and prognostic factors for neck pain in whiplash-associated disorders (WAD): results of the Bone and Joint Decade 2000-2010 Task Force on Neck Pain and Its Associated Disorders. *Spine*, 33, S83-92.

Carroll, L., Hurwitz, E., Cote, P., Hogg-Johnson, S., Carragee, E. J., Nordin, M., Holm, L., Van Der Velde, G., Cassidy, J., Guzman, J., Peloso, P., Haldeman, S. (2008d) Research priorities and methodological implications: the Bone and Joint Decade 2000-2010 Task Force on Neck Pain and Its Associated Disorders. *Spine*, 33, S214-220.

Cassidy, J., Carroll, L., Cote, P., Frank, J. (2007) Does multidisciplinary rehabilitation benefit whiplash recovery? Results of a population –based incidence cohort study. *Spine*, 32, 126-131.

Cassidy, J., Carroll, L., Cote, P., Lemstra, M., Berglund, A., Nygren, A. (2000) Effect of eliminating compensation for pain and suffering on the outcome of insurance claims for whiplash injury. *New England Journal of Medicine*, 342, 1179-1186.

Cassidy, J., Cote, P. (2008) Is it time for a population health approach to neck pain? *Journal of Manipulative and Physiological Therapeutics*, 31, 442-446.

Cassidy, J., Carroll, L., Cote, P., Berglund, A., Nygren, A. (2003) Low back pain after traffic collisions: a population-based cohort study. *Spine*, 28, 1002-1009.

Cassou, B., Derriennic, F., Monfort, C., Norton, J., Touranchet, A. (2002) Chronic neck and shoulder pain, age, and working conditions: longitudinal results from a large random sample in France. *Occupational and Environmental Medicine*, 59, 537-544.

Centre For Reviews And Dissemination (CRD) (2008) *Systematic reviews: CRD guidance for undertaking reviews in health care.* University of York, CRD.

Cervical Range-Of-Motion (CROM) Device. Performance Attainment Assoc, St Paul, Minnesota, USA.

Charlton, E. (1995) Ethical guidelines for pain research in Humans. *Pain*, 63, 277-278.

Charter, R. (2003) Combining reliability coefficients: possible application to meta-analysis and reliability generalization. *Psychological Reports*, 93, 643-647.

Chene, G., Thompson, S. (1996) Methods for summarising the risk associations of quantitative variables in epidemiologic studies in a consistent form. *American Journal of Epidemiology*, 144, 610-621.

Cherkin, D., Deyo, R., Street, J., Barlow, W. (1996) Predicting poor outcome for back pain seen in primary care using patients' own criteria. *Spine*, 21, 2900-2907.

Cherkin, D., Deyo, R., Wheeler, K., Ciol, M. (1995) Physician views about treating low back pain. The results of a national survey. *Spine*, 20, 1-9.

Childs, J. (2006) Re: Inquiry about the Delitto classification system and treatment – based classifications. Email sent to Angeliki Chorti (physio_glaraki@yahoo.com), 24th July 2006.

Childs, J., Cleland, J., Elliott, J., Teyhen, D., Wainner, R., Whitman, J., Sopky, B., Godges, J., Flynn, T. (2008) Neck Pain: Clinical Practice Guidelines linked to the International Classification of Functioning, Disability and Health from the orthopaedic section of the American Physical Therapy Association. *Journal of Orthopaedic and Sports Physical Therapy*, 38, A1-34.

Childs, J., Fritz, J., Flynn, T., Irrgang, J., Johnson, K., Majkowski, G., Delitto, A. (2004a) A clinical prediction rule to identify patients with low back pain most likely

to benefit from spinal manipulation: a validation study. *Annals of Internal Medicine*, 141, 920-928.

Childs, J., Fritz, J., Piva, S., Whitman, J. (2004b) Proposal of a classification system for patients with neck pain. *Journal of Orthopaedic and Sports Physical Therapy*, 34, 686-700.

Chiradejnant, A., Latimer, J., Maher, C. (2002) Forces applied during manual therapy to patients with low back pain. *Journal of Manipulative and Physiological Therapeutics*, 25, 362-369.

Chou, R., Qaseem, A., Snow, V., Casey, D., Cross, J. Jr., Shekelle, P., Owens, D., Clinical Efficacy Assessment Subcommittee Of The American College Of Physicians, American College Of Physicians, American Pain Society Low Back Pain Guidelines Panel (2007) Diagnosis and treatment of low back pain: a joint clinical practice guideline from the American College of Physicians and the American Pain Society. *Annals of Internal Medicine*, 147, 478-491.

Christensen, H., Vach, W., Manniche, C., Haghfelt, T., Hartvigsen, L., Hoilund-Carlsen, P. (2003) Palpation for muscular tenderness in the anterior chest wall: an observer reliability study. *Journal of Manipulative and Physiological Therapeutics*, 26**,** 469-475.

Christiansen, D., Larsen, K., Kudsk Jensen, O., Vinther Nielsen, C. (2009) Pain responses in repeated end-range spinal movements and psychological factors in sick-listed patients with low back pain: is there an association? *Journal of Rehabilitation Medicine*, 41, 545-549.

Cibulka, M., Delitto, A., Koldehoff, R. (1988) Changes in innominate tilt after manipulation of the sacroiliac joint in patients with low back pain. An experimental study. *Physical Therapy*, 68, 1359-1363.

Cicchetti, D. (1981) Testing the normal approximation and minimal sample size requirements of weighted kappa when the number of categories is large. *Applied Psychological Measurement*, 5, 101-104.

Cicchetti, D., Feinstein, A. (1990) High agreement but low Kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, 43, 543-549.

Cicchetti, D., Fleiss, J. (1977) Comparison of the null distributions of weighted kappa and the C ordinal statistic. *Applied Psychological Measurement*, 1, 195-201.

Cicchetti, D., Sparrow, S. (1981) Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. *American Journal of Mental Deficiency*, 86, 127-137.

Cieza, A., Stucki, G., Weigl, M., Disler, P., Jackel, W., Van Der Linden, S., Kostanjsek, N., De Bie, R. (2004) ICF core sets for low back pain. *Journal of Rehabilitation Medicine*, 44, S69-74.

Clare, H., Adams, R., Maher, C. (2004a) A systematic review of efficacy of McKenzie therapy for spinal pain. *Australian Journal of Physiotherapy*, 50, 209-216.

Clare, H., Adams, R., Maher, C. (2005) Reliability of McKenzie classification of patients with cervical or lumbar pain. *Journal of Manipulative and Physiological Therapeutics*, 28, 122-127.

Clare, H., Adams, R., Maher, C. (2004b) Reliability of McKenzie spinal pain classification using patient assessment forms. *Physiotherapy*, 90, 114-119.

Clark-Carter, D. (2004) *Quantitative Psychological Research: A student's handbook*. Hove & New York, Psychology Press.

Cleland, J., Childs, J., Fritz, J., Whitman, J. (2006a) Interrater reliability of the history and physical examination in patients with mechanical neck pain. *Archives of Physical Medicine and Rehabilitation*, 87, 1388-1395.

Cleland, J., Childs, J., Fritz, J., Whitman, J., Eberhart, S. (2007) Development of a clinical prediction rule for guiding treatment of a subgroup of patients with neck pain: use of thoracic spine manipulation, exercise and patient education. *Physical Therapy*, 87, 9-23.

Cleland, J., Childs, J., Palmer, J., Eberhart, S. (2006b) Slump stretching in the management of non-radicular low back pain: a pilot clinical trial. *Manual Therapy*, 11, 279-286.

Cleland, J., Noteboom, J., Whitman, J., Allison, S. (2008) A primer on selected aspects of evidence – based practice relating to questions of treatment, Part 1: Asking questions, finding evidence, and determining validity. *Journal of Orthopaedic and Sports Physical Therapy*, 38, 476-484.

Clinical Standards Advisory Group (CSAG) (1994) *Back pain.* London, HMSO.

Cohen, J. (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46.

Cohen, J. (1968) Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70, 213-220.

Cole, D., Hudak, P. (1996) Prognosis of nonspecific work-related musculoskeletal disorders of the neck and upper extremity. *American Journal of Industrial Medicine*, 29, 657-668.

Comeaux, Z., Eland, D., Chila, A., Pheley, A., Tate, M. (2001) Measurement challenges in physical diagnosis: refining inter-rater palpation, perception and communication. *Journal of Bodywork and Movement Therapies*, 5, 245-253.

Conger, A. (1980) Integration and generalization of Kappas for multiple raters. *Psychological Bulletin*, 88, 322-328.

Cook, C., Hegedus, E., Ramey, K. (2005) Physical therapy exercise intervention based on classification using the patient response method: a systematic review of the literature. *Journal of Manual and Manipulative Therapy*, 13, 152-162.

Copeland, J., Taylor, W., Dean, S. (2008) Factors influencing the use of outcome measures for patients with low back pain: a survey of New Zealand physical therapists. *Physical Therapy,* 88, 1492-505.

Coste, J., Delecoeuillerie, G., Cohen De Lara, A., Le Parc, J., Paolaggi, J. (1994) Clinical course and prognostic factors in acute low back pain: an inception cohort study in primary care practice. *British Medical Journal*, 308, 577-580.

Coste, J., Paolaggi, J., Spira, A. (1992a) Classification of nonspecific low back pain. I. Psychological involvement in low back pain. A clinical, descriptive approach. *Spine*, 17, 1028-1037.

Coste, J., Paolaggi, J., Spira, A. (1992b) Classification of nonspecific low back pain. II. Clinical diversity of organic forms. *Spine*, 17, 1038-1042.

Coste, J., Spira, A., Ducimetiere, P., Paolaggi, J. (1991) Clinical and psychological diversity of non-specific low back pain. A new approach towards the classification of clinical subgroups. *Journal of Clinical Epidemiology*, 44, 1233-1245.

Cote, P., Cassidy, J., Carroll, L. (2001a) The treatment of neck and low back pain. Who seeks care? Who goes where? *Medical Care,* 39, 956-967.

Cote, P., Cassidy, J., Carroll, L., Frank, J., Bombardier, C. (2001b) A systematic review of the prognosis of acute whiplash and a new conceptual framework to synthesize the literature. *Spine*, 26, E445-458.

Cote, P., Cassidy, J., Carroll, L., Kristman, V. (2004) The annual incidence and course of neck pain in the general population: a population – based cohort study. *Pain*, 112, 267-273.

Cote, P., Hogg-Johnson, S., Cassidy, J., Carroll, L., Frank, J., Bombardier, C. (2007) Early aggressive care and delayed recovery from whiplash: isolated finding or reproducible result? *Arthritis and Rheumatism,* 57, 861-868.

Cote, P., Hogg-Johnson, S., Cassidy, J., Carroll, L., Frank, J., Bombardier, C. (2005) Initial patterns of clinical care and recovery from whiplash injuries: a population – based cohort study. *Archives of Internal Medicine*, 165, 2257-2263.

Cote, P., Van Der Velde, G., Cassidy, J., Carroll, L., Hogg-Johnson, S., Holm, L., Carragee, E., Haldeman, S., Nordin, M., Hurwitz, E., Guzman, J., Peloso, P. (2008) The burden and determinants of neck pain in workers: results of the Bone and Joint Decade 2000-2010 Task Force on Neck Pain and Its Associated Disorders. *Spine*, 33, S60-74.

Counsell, C. (1997) Formulating questions and locating primary studies for inclusion in systematic reviews. *Annals of Internal Medicine*, 127, 380-387.

Couper, M. (1984) The Delphi technique: characteristics and sequence model. *Advances in Nursing Science*, 7, 72-77.

Croft, P., Lewis, M., Hannaford, P. (2003) Is all chronic pain the same? A 25-year follow-up study. *Pain*, 105, 309-317.

Croft, P., Lewis, M., Papageorgiou, A., Thomas, E., Jayson, M., Macfarlane, G., Silman, A. (2001) Risk factors for neck pain: a longitudinal study in the general population. *Pain*, 93, 317-325.

Crooks, P., Davies S. (eds) (1998) *Research into practice: Essential skills for reading and applying research.* London, Ballière Tindall, in association with the RCN.

Cross, H. (2005) Consensus methods: a bridge between clinical reasoning and clinical research? *International Journal of Leprosy*, 73, 28-31.

Cutler, R., Fishbain, D., Cole, B., Steele-Rosomoff, R., Rosomoff, H. (2001) Identifying patients at risk for loss to follow-up after pain center treatment. *Pain Medicine*, 2, 46-51.

Cyriax, J. (1982) The lumbar region: examination. In: *Textbook of orthopaedic medicine.* 8th edition. London, Bailliere Tindall, 253-279.

Dalkey, N. (1969) The Delphi method: an experimental study of group opinion. RAND.

Damush, T., Weinberger, M., Clark, D., Tierney, W., Rao, J., Perkins, S., Verel, K. (2002) Acute low back pain self-management intervention for urban primary care patients: rationale, design, and predictors of participation. *Arthritis and Rheumatism*, 47, 372-379.

Danish Institute For Health Technology Assessment (1999) Low back pain: frequency, management, and prevention from an HTA perspective. Denmark, Danish Institute for Health Technology Assessment.

Davis, M., Fleiss, J. (1982) Measuring agreement of multinomial data. *Biometrics*, 38, 1047-1051.

Deeks, J., Dinnes, J., D'amico, R., Sowden, A., Sakarovitch, C., Song, F., Petticrew, M., Altman, D., International Stroke Trial Collaborative & European Carotid Surgery Trial Collaborative (2003) Evaluating non-randomised intervention studies. *Health Technology Assessment*, 7, 1-173.

Deeks, J., Higgins, J., Altman, D. On Behalf Of The Cochrane Statistical Methods Group (2008) Chapter 9: Analysing data and undertaking meta-analyses. In: Higgins, J., Green, S. (eds) *Cochrane Handbook for Systematic Reviews of Interventions.* Version 5.0.1 [updated September 2008] The Cochrane Collaboration. Available at: http://www.cochrane-handbook.org/ [14th April 2009].

De Koning, C., Van Den Heuvel, S., Staal, J., Smits – Engelsman, B., Hendriks, E. (2008a) Clinimetric evaluation of active range of motion measures in patients with non-specific neck pain: a systematic review. *European Spine Journal*, 17, 905-921.

De Koning, C., Van Den Heuvel, S., Staal, J., Smits – Engelsman, B., Hendriks, E. (2008b) Clinimetric evaluation of methods to measure muscle functioning in patients with non-specific neck pain: a systematic review. *BMC Musculoskeletal Disorders*, 9, 142.

Delbecq, A., Van De Ven, A., Gustafson, D. (1975) *Group techniques for program planning: a guide to nominal group and Delphi processes.* Glenview, Ill: Scott, Foresman and Company. Available at: https://netfiles.umn.edu/users/avandeve/www/Group%20Techniques/COVER1.HTM [24th October 2008].

Delitto, A. (2006) Re: Inquiry about the Delitto classification system and treatment-based classifications. Email sent to Angeliki Chorti (physio_glaraki@yahoo.com), 25th July 2006.

Delitto, A., Cibulka, M. T., Erhard, R., Bowling, R., Tenhula, J. (1993) Evidence for use of an extension-mobilization category in acute low back syndrome: a prescriptive validation pilot study. *Physical Therapy*, 73**,** 216-222.

Delitto, A., Erhard, R., Bowling, R. (1995) A treatment-based classification approach to low back syndrome: identifying and staging patients for conservative treatment. *Physical Therapy*, 75, 470-489.

Delitto, A., Shulman, A., Rose, S., Strube, M., Erhard, R., Bowling, R., Smith, D. (1992) Reliability of a clinical examination to classify patients with low back syndrome. *Physical Therapy Practice*, 1**,** 1-9.

Department of Health (DoH) (2001) *The expert patient: a new approach to chronic disease management in the 21st century.* London, HMSO.

Department of Health (DoH) (2008) Common core principles to support self-care: a guide to support implementation. Available at: http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_084505?IdcService=GET_FILE&dID=164571&Rendition=Web [28th May 2008].

Derosa, C., Porterfield, J. (1992) A physical therapy model for the treatment of low back pain. *Physical Therapy*, 72, 261-269.

De Vet, H., Heymans, M., Dunn, K., Pope, D., Van Der Beek, A., Macfarlane, G., Bouter, L., Croft, P. (2002) Episodes of low back pain, a proposal for uniform definitions to be used in research. *Spine*, 27, 2409-2416.

De Vet, H., Terwee, C., Bouter, L. (2003a) Current challenges in clinimetrics. *Journal of Clinical Epidemiology*, 56, 1137-1141.

De Vet, H., Terwee, C., Knol, D., Bouter, L. (2006) When to use agreement versus reliability measures. *Journal of Clinical Epidemiology,* 59, 1033-1039.

De Vet, H., Van Tulder, M., Bouter, L. (2003b) Levels of evidence: intellectual aid or absolute judgement? *Journal of Clinical Epidemiology*, 56, 917-918.

Devillé, W., Buntinx, F., Bouter, L., Montori, V., De Vet, H., Van Der Windt, D., Bezemer, P. (2002) Conducting systematic reviews of diagnostic studies: didactic guidelines. *BMC Medical Research Methodology*, 2.

Deyo, R. (1993) Practice variations, treatment fads, rising disability: do we need a new clinical research paradigm? *Spine*, 18, 2153-2162.

Deyo, R. (2004) Treatments for back pain: Can we get past trivial effects? *Annals of Internal Medicine*, 141, 957-958.

Deyo, R., Diehl, A. (1988) Psychosocial predictors of disability in patients with low back pain. *Journal of Rheumatology*, 15, 1557-1564.

Dickersin, K., Berlin, J. (1992) Meta-analysis: state-of-the-science. *Epidemiologic Reviews*, 14, 154-176.

Dickersin, K., Min, Y. (1993) Publication bias: the problem that won't go away. *Annals of the New York Academy of Science*, 703, 135-146.

Dinnes, J., Deeks, J., Kirby, J., Roderick, P. (2005) A methodological review of how heterogeneity has been examined in systematic reviews of diagnostic test accuracy. *Health Technology Assessment*, 9.

Dionne, C., Bourbonnais, R., Fremont, P., Rossignol, M., Stock, S., Larocque, I. (2005) A clinical return-to-work rule for patients with back pain. *Canadian Medical Association Journal*, 172, 1559-1567.

Dionne, C., Bybee, R., Tomaka, J. (2006) Inter-rater reliability of McKenzie assessment in patients with neck pain. *Physiotherapy*, 92, 75-82.

Dionne, C., Dunn, K., Croft, P., Nachemson, A., Buchbinder, R., Walker, B., Wyatt, M., Cassidy, J., Rossignol, M., Leboeuf-Yde, C., Hartvigsen, J., Leino-Arjas, P., Latza, U., Reis, S., Del Real, M-T., Kovacs, F., Oberg, B., Cedrasi, C., Bouter, L., Koes, B., Picavet, S., Van Tulder, M., Burton, K., Foster, N., Macfarlane, G., Thomas, E., Underwood, M., Waddell, G., Shekelle, P., Volinn, E., Von Korff, M. (2008) A consensus approach toward the standardisation of back pain definitions for use in prevalence studies. *Spine*, 33, 95-103.

Dionne, C., Von Korff, M., Koepsell, T., Deyo, R., Barlow, W., Checkoway, H. (1999) A comparison of pain, functional limitation, and work status indices as outcome measures in back pain research. *Spine*, 24, 2339-2345.

Dionne, C., Koepsell, T., Von Korff, M., Deyo, R., Barlow, W., Checkoway, H. (1997) Predicting long term functional limitations among back pain patients in primary care settings. *Journal of Clinical Epidemiology*, 50, 31-43.

Donahue, M., Riddle, D., Sullivan, M. (1996) Intertester reliability of a modified version of McKenzie's lateral shift assessments obtained on patients with low back pain. *Physical Therapy*, 76**,** 706-716.

Donelson, R. (2008) Is your client's back pain "rapidly reversible"? Improving low back care at its foundation. *Professional Case Manager*, 13, 87-96.

Donelson, R., Aprill, C., Medcalf, R., Grant, W. (1997) A prospective study of centralization of lumbar and referred pain. A predictor of symptomatic discs and anular competence. *Spine*, 22, 1115-1122.

Donelson, R., Grant, W., Kamps, C., Medcalf, R. (1991) Pain response to sagittal end-range spinal motion. A prospective, randomized, multicentered trial. *Spine*, 16, S206-212.

Donelson, R., Silva, G., Murphy, K. (1990) Centralization phenomenon. Its usefulness in evaluating and treating referred pain. *Spine*, 15, 211-213.

Donner, A. (2006) *Re: Inquiry on your articles on interobserver agreement*. Email sent to Angeliki Chorti (physio_glaraki@yahoo.com), 4th September 2006.

Donner, A. (1998) Sample size requirements for the comparison of two or more coefficients of inter-observer agreement. *Statistics in Medicine*, 17, 1157-1168.

Donner, A., Eliasziw, M. (1992) A goodness-of-fit approach to inference procedures for the kappa statistic: confidence interval construction, significance-testing and sample size estimation. *Statistics in Medicine*, 11, 1511-1519.

Donner, A., Eliasziw, M. (1997) A hierarchical approach to inferences concerning interobserver agreement for multinomial data. *Statistics in Medicine*, 16, 1097-1106.

Donner, A., Klar, N. (1996) The statistical analysis of kappa statistics in multiple samples. *Journal of Clinical Epidemiology*, 49, 1053-1058.

Doxey, N., Dzioba, R., Mitson, G., Lacroix, J. (1988) Predictors of outcome in back surgery candidates. *Journal of Clinical Psychology*, 44, 611-622.

Dreisinger, T. (2006) *Re: MII Research Committee*. Email sent to Angeliki Chorti (physio_glaraki@yahoo.com), 20th September 2006.

Duncan, E., Nicol, M., Ager, A. (2004) Factors that constitute a good cognitive behavioural treatment manual: a Delphi study. *Behavioural and Cognitive Psychotherapy*, 32, 199-213.

Dworkin, R., Richlin, D., Handlin, D., Brand, L. (1986) Predicting treatment response in depressed and non-depressed chronic pain patients. *Pain*, 24, 343-353.

Edwards, B. (1994) Clinical assessment: the use of combined movements in assessment and treatment. In: Twomey, L., Taylor, J., (eds) *Physical Therapy of the low back.* 2nd edition. New York, Churchill Livingstone Inc., 197-220.

Egger, M., Smith, D. (1998) Bias in location and selection of studies. *British Medical Journal*, 316, 61-66.

Egger, M., Smith, G., Altman, D. (2001) *Systematic reviews in health care. Meta-analysis in context.* London, BMJ Books.

Egger, M., Zellweger-Zahner, T., Schneider, M., Junker, C., Lengeler, C., Antes, G. (1997) Language bias in randomised controlled trials published in English and German. *Lancet*, 350, 326-329.

Egwu, M. (2008) Relative therapeutic efficacy of some vertebral mobilization techniques in the management of unilateral cervical spondylosis: a comparative study. *Journal of Physical Therapy Science*, 20, 103-108.

Enthoven, P., Skargren, E., Oberg, B. (2004) Clinical course in patients seeking primary care for back or neck pain: a prospective 5-year follow-up of outcome and health care consumption with subgroup analysis. *Spine*, 29, 2458-2465.

Enthoven, P., Skargren, E., Kjellman, G., Oberg, B. (2003) Course of back pain in primary care: a prospective study of physical measures. *Journal of Rehabilitation Medicine*, 35, 168-173.

Erffmeyer, R., Erffmeyer, E., Lane, I. (1986) The Delphi technique: an empirical evaluation of the optimal number of rounds. *Group and Organization Management*, 11, 120-128.

Erhard, R., Delitto, A., Cibulka, M. (1994) Relative effectiveness of an extension program and a combined program of manipulation and flexion and extension exercises in patients with acute low back syndrome. *Physical Therapy*, 74, 1093-1100.

Eriksen, W. (2004) Do people who were passive smokers during childhood have increased risk of long-term work disability? A 15-month prospective study of nurses' aides. *European Journal of Public Health*, 14, 296-300.

Eriksen, W., Natvig, B., Knardahl, S., Bruusgaard, D. (1999) Job characteristics as predictors of neck pain. A 4-year prospective study. *Journal of Occupational and Environmental Medicine*, 41, 893-902.

Farmer, C., Wells, J., Lund, A. (2002) *Effects of head restraint and seat redesign on neck injury risk in rear –end crashes.* Arlington, VA, Insurance Institute for Highway Safety.

Farmer, C., Wells, J., Werner, J. (1999) Relationship of head restraint positioning to driver neck injury in rear-end crashes. *Accident Analysis and Prevention*, 31, 719-728.

Feinstein, A., Cicchetti, D. (1990) High agreement but low Kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, 43, 543-549.

Fejer, R., Hartvigsen, J. (2008) Neck pain and disability due to neck pain: what is the relation? *European Spine Journal*, 17, 80-88.

Fejer, R., Kyvik, K., Hartvigsen, J. (2006) The prevalence of neck pain in the world population: a systematic critical review of the literature. *European Spine Journal*, 15, 834-848.

Ferrari, R., Russell, A. (2003) Regional musculoskeletal conditions: neck pain. *Best Practice and Research Clinical Rheumatology*, 17, 57-70.

Ferreira, P., Ferreira, M., Maher, C., Refshauge, K., Herbert, R., Latimer, J. (2002) Effect of applying different "levels of evidence" criteria on conclusions of Cochrane reviews of interventions for low back pain. *Journal of Clinical Epidemiology*, 55, 1126-1129.

Fink, A., Kosecoff, J., Chassin, M., Brook, R. (1984) Consensus methods: characteristics and guidelines for use. *American Journal of Public Health*, 74, 979-983.

Fleiss, J. (1986) Analysis of data from multi clinic trials. *Controlled Clinical Trials*, 7, 267-275.

Fleiss, J. (1971) Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76, 378-382.

Fleiss, J. (2003) *Statistical methods for rates and proportions*, 3rd edition. New York, Wiley.

Fleiss, J. (1986) *The design and analysis of clinical experiments*, New York, Wiley.

Fleiss, J., Cicchetti, D. (1978) Inference about weighted kappa in the non-null case. *Applied Psychological Measurement*, 2, 113-117.

Fleiss, J., Cuzick, J. (1979) The reliability of dichotomous judgements: unequal numbers of judges per subject. *Applied Psychological Measurement*, 3, 537-542.

Fleiss, J., Davies, M. (1982) Jack knifing functions of multinomial frequencies, with an application to a measure of concordance. *American Journal of Epidemiology*, 115, 841-845.

Fleiss, J., Nee, J., Landis, J. (1979) Large sample variance of kappa in the case of different sets of raters. *Psychological Bulletin*, 86, 974-977.

Fletcher, R., Fletcher, S., Wagner, E. (1996) *Clinical epidemiology: the essentials.* Baltimore, London, Williams & Wilkins.

Flynn, T., Fritz, J., Whitman, J., Wainner, R., Magel, J., Rendeiro, D., Butler, B., Garber, M., Allison, S. (2002) A clinical prediction rule for classifying patients with low back pain who demonstrate short-term improvement with spinal manipulation. *Spine*, 27**,** 2835-2843.

Foster, N., Thompson, K., Baxter, D., Allen, J. (1999) Management of nonspecific low back pain by physiotherapists in Britain and Ireland: A descriptive questionnaire of current clinical practice. *Spine*, 24, 1332-1342.

Frances, A., Kahn, D., Carpenter, D., Frances, G., Docherty, J. (1998) A new method of developing expert consensus practice guidelines. *American Journal of Managed Care*, 4, 1023-1029.

French, S., Green, S., Forbes, A. (2000) Reliability of chiropractic methods commonly used to detect manipulable lesions in patients with chronic low-back pain. *Journal of Manipulative and Physiological Therapeutics*, 23, 231-238.

Fritz, J. (2004) Invited commentary. *Physical Therapy*, 84, 332-333.

Fritz, J. (2006) Re: [Fwd: Inquiry about the Delitto classification system and treatment-based classifications]. Email sent to Angeliki Chorti (physio_glaraki@yahoo.com), 22nd September 2006.

Fritz, J. (1998) Use of a classification approach to the treatment of 3 patients with low back syndrome. *Physical Therapy*, 78, 766-777.

Fritz, J., Brennan, G. (2007) Preliminary examination of a proposed treatment-based classification system for patients receiving physical therapy interventions for neck pain. *Physical Therapy*, 87, 513-524.

Fritz, J., Brennan, G., Clifford, S., Hunter, S., Thackeray, A. (2006) An examination of the reliability of a classification algorithm for subgrouping patients with low back pain. *Spine*, 31, 77-82.

Fritz, J., Childs, J., Flynn, T. (2005) Pragmatic application of a clinical prediction rule in primary care to identify patients with low back pain with a good prognosis following a brief spinal manipulation intervention. *BMC Family Practice*, 6, 29.

Fritz, J., Cleland, J., Childs, J. (2007a) Subgrouping patients with low back pain: evolution of a classification approach to physical therapy. *Journal of Orthopaedic and Sports Physical Therapy*, 37, 290-302.

Fritz, J., Delitto, A., Erhard, R. (2003) Comparison of a classification-based physical therapy with therapy based on clinical practice guidelines for patients with acute low back pain. *Spine*, 28, 1363-1372.

Fritz, J., Delitto, A., Vignovic, M., Busse, R. (2000a) Interrater reliability of judgments of the centralization phenomenon and status change during movement testing in patients with low back pain. *Archives of Physical Medicine and Rehabilitation,* 81**,** 57-61.

Fritz, J., George, S. (2002) Identifying psychosocial variables in patients with acute work-related low back pain: the importance of fear-avoidance beliefs. *Physical Therapy,* 82, 973-983.

Fritz, J., George, S. (2000) The use of a classification approach to identify subgroups of patients with acute low back pain. Interrater reliability and short-term treatment outcomes. *Spine,* 25, 106-114.

Fritz, J., Lindsay, W., Matheson, J., Brennan, G., Hunter, S., Moffit, S., Swalberg, A., Rodriguez, B. (2007b) Is there a subgroup of patients with low back pain likely to benefit from mechanical traction? Results of a randomized clinical trial and subgrouping analysis. *Spine,* 32, E793-800.

Fritz, J., Piva, S. (2003) Physical impairment index: reliability, validity, and responsiveness in patients with acute low back pain. *Spine*, 28**,** 1189-1194.

Fritz, J., Wainner, R. (2001) Examining diagnostic tests: an evidence-based perspective. *Physical Therapy*, 81, 1546-1564.

Fritz, J., Wainner, R., Hicks, G. (2000b) The use of nonorganic signs and symptoms as a screening tool for return-to-work in patients with acute low back pain. *Spine*, 25, 1925-1931.

Gallagher, R., Rauh, V., Haugh, L., Milhous, R., Callas, P., Langelier, R., Mcclallen, J., Frymoyer, J. (1989) Determinants of return-to-work among low back pain patients. *Pain*, 39, 55-67.

Gamsa, A. (1994) The role of psychological factors in chronic pain. II. A critical appraisal. *Pain*, 57, 17-29.

Gardner, M., Altman, D. (1986) Confidence intervals rather than p values: estimation rather than hypothesis testing. *British Medical Journal,* 292, 746-750.

Garner, J. (1991) The standard error of Cohen's Kappa. *Statistics in Medicine,* 10, 767-775.

George, S. (2006) Re: Inquiry about your response on the centralization issue in JOSPT 2005. Email sent to Angeliki Chorti (physio_glaraki@yahoo.com), 2nd August 2006.

George, S., Bialosky, J., Donald, D. (2005) The centralisation phenomenon and fear-avoidance beliefs as prognostic factors for acute low back pain: a preliminary investigation involving patients classified for specific exercise. *Journal of Orthopaedic and Sports Physical Therapy*, 35, 580-588.

George, S., Fritz, J. (2005) Author response. *Journal of Orthopaedic and Sports Physical Therapy*, 35, 845-847.

George, S., Hirsh, A. (2005) Distinguishing patient satisfaction with treatment delivery from treatment effect: a preliminary investigation of patient satisfaction with symptoms after physical therapy treatment of low back pain. *Archives of Physical Medicine and Rehabilitation*, 86, 1338-1344.

Gerr, F., Marcus, M., Ensor, C., Kleinbaum, D., Cohen, S., Edwards, A., Gentry, E., Ortiz, D., Monteilh, C. (2002) A prospective study of computer users: I. Study design and incidence of musculoskeletal symptoms and disorders. *American Journal of Industrial Medicine*, 41, 221-235.

Gerr, F., Marcus, M., Monteilh, C., Hannan, L., Ortiz, D., Kleinbaum, D. (2005) A randomised controlled trial of postural interventions for prevention of musculoskeletal symptoms among computer users. *Occupational and Environmental Medicine*, 62, 478-487.

Gerth, W., Smith, M. (1991) *The Delphi technique: Background for use in probability estimation*. Human Health Economics MHHD, Merk and Co.

Gerwin, R., Shannon, S., Hong, C., Hubbard, D., Gevirtz, R. (1997) Interrater reliability in myofascial trigger point examination. *Pain*, 69**,** 65-73.

Goldstein, M., Morgenstern, H., Hurwitz, E., Yu, F. (2002) The impact of treatment confidence on pain and related disability among patients with low-back pain: results from the University of California, Los Angeles, low-back pain study. *Spine Journal*, 2, 391-399.

Goodman, C. (1987) The Delphi technique: a critique. *Journal of Advanced Nursing*, 12, 729-734.

Gorbach, C., Schmid, M., Elfering, A., Hodler, J., Boos, N. (2006) Therapeutic efficacy of facet joint blocks. *American Journal of Roentgenology*, 186, 1228-1233.

Gore, D., Sepic S., Gardner G., Murray M. (1987) Neck pain: a long-term follow-up of 205 patients. *Spine*, 12, 1-5.

Gracey, J., Mcdonough, S., Baxter, G. (2002) Physiotherapy management of low back pain: a survey of current practice in Northern Ireland. *Spine*, 27, 406-411.

Greatorex, J., Dexter, T. (2000) An accessible analytical approach for investigating what happens between the rounds of a Delphi study. *Journal of Advanced Nursing*, 32, 1016-1024.

Greenhalgh, T. (1997) Papers that go beyond numbers (qualitative research). *British Medical Journal*, 315, 740-743.

Greenhalgh, S., Selfe, J. (2006) *Red flags: a guide to identifying serious pathology of the spine*. Edinburgh, Churchill Livingstone.

Grimmer, K., Sheppard, L., Pitt, M., Magarey, M., Trott, P. (1999) Differences in stakeholder expectations in the outcome of physiotherapy management of acute low back pain. *International Journal of Quality in Health Care*, 11, 155-162.

Gross, S. (1986) The kappa coefficient of agreement for multiple observers when the number of subjects is small. *Biometrics*, 42, 883-893.

Grotle, M., Brox, J., Veierod, M., Glomsrod, B., Lonn J., Vollestad, N. (2005) Clinical course and prognostic factors in acute low back pain: Patients consulting primary care for the first time. *Spine*, 30, 976-982.

Grzegorzewski, P. (2006) The coefficient of concordance for vague data. *Computational Statistics and Data Analysis*, 51, 314-322.

Guggenmoos-Holzmann, I., Vonk, R. (1998) Kappa-like indices of observer agreement viewed from a latent class perspective. *Statistics in Medicine*, 17, 797-812.

Gun, R., Osti, O., O'riordan, A., Mpelasoka, F., Eckerwall, C., Smyth J. (2005) Risk factors for prolonged disability after whiplash injury: a prospective study. *Spine*, 30, 386-391.

Guzman, J., Hurwitz, E., Carroll, L., Haldeman, S., Cote, P., Carragee, E., Peloso, P., Van Der Velde, G., Holm, L., Hogg-Johnson, S., Nordin, M., Cassidy, J. (2008) A new conceptual model of neck pain. Linking onset, course, and care: The Bone and Joint Decade 2000-2010 Task Force on Neck Pain and its Associated Disorders. *Spine*, 33, S14-23.

Gwet, K. (2008a) Variance estimation of nominal scale inter-rater reliability with random selection of raters. *Psychometrika*, 73, 407-430.

Gwet, K. (2008b) Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61, 29-48.

Haas, M. (1995) How to evaluate intraexaminer reliability using an interexaminer reliability study design. *Journal of Manipulative and Physiological Therapeutics*, 18, 10-15.

Haas, M. (1991) Statistical methodology for reliability studies. *Journal of Manipulative and Physiological Therapeutics*, 14**,** 119-132.

Haas, M., Nyiendo, J., Aickin, M. (2002) One-year trend in pain and disability relief recall in acute and chronic ambulatory low back pain patients. *Pain*, 95, 83-91.

Haazen, I., Vlaeyen, J., Kole-Snijders, A., Van Eek, H., Van Es, F. (1994) Behavioral rehabilitation of chronic low back pain: Searching for predictors of treatment outcome. *Journal of Rehabilitation Sciences*, 7, 34-43.

Hagel, B., Fick, G., Meeuwisse, W. (2003) Injury risk in men's Canada West University football. *American Journal of Epidemiology*, 157, 825-833.

Hagen, E., Svensen, E., Eriksen, H. (2005) Predictors and modifiers of treatment effect influencing sick leave in subacute low back pain patients. *Spine*, 30, 2717-2723.

Hahne, A., Keating, J., Wilson, S. (2004) Do within-session changes in pain intensity and range of motion predict between-session changes in patients with low back pain? *Australian Journal of Physiotherapy*, 50, 17-23.

Haldeman, S. (2008) The Bone and Joint Decade 2000-2010 Task Force on Neck Pain and its Associated Disorders: Executive summary. *Spine*, 33**,** S5-7.

Haldorsen, E., Indahl, A., Ursin, H. (1998) Patients with low back pain not returning to work. A 12-month follow-up study. *Spine*, 23, 1202-1208.

Hale, C., Fleiss, J. (1993) Interval estimation under two study designs for kappa with binary classification. *Biometrics*, 49, 523-534.

Halpern, M. (2001) Functional assessment taxonomy relevant to low-back impairments. *Journal of Occupational Rehabilitation*, 11, 201-215.

Hamberg-Van Reenen, H., Ariens, G., Blatter, B., Twisk, J., Van Mechelen, W., Bongers, P. (2006) Physical capacity in relation to low back, neck, or shoulder pain in a working population. *Occupational and Environmental Medicine*, 63, 371-377.

Hancock, M., Maher, C., Latimer, J., Spindler, M., Mcauley, J., Laslett, M., Bogduk, N. (2007) Systematic review of tests to identify the disc, SIJ, or facet joint as the source of low back pain. *European Spine Journal*, 16, 1539-1550.

Harper, A., Harper, D., Lambert, I., Andrews, H., Lo, S., Ross, F., Straker, L. (1992) Symptoms of impairment, disability and handicap in low back pain: a taxonomy. *Pain*, 50, 189-195.

Hartemink, N., Boshuizen, H., Nagelkerke, N., Jacobs, M., Van Houwelingen, H. (2006) Combining risk estimates from observational studies with different exposure cut points: a meta-analysis on body mass index and diabetes type 2. *American Journal of Epidemiology*, 163, 1042-1052.

Hartling, L., Pickett, W., Brison, R. (2002) Derivation of a clinical decision rule for whiplash associated disorders among individuals involved in rear-end collisions. *Accident Analysis and Prevention*, 34, 531-539.

Hartman, D. (1977) Considerations in the choice of interobserver reliability estimates. *Journal of Applied Behavior Analysis*, 10, 103-116.

Hasson, F., Keeney, S., Mckenna, H. (2000) Research guidelines for the Delphi survey technique. *Journal of Advanced Nursing*, 32, 1008-1015.

Haswell, K., Williams, M. (2004) Interexaminer reliability of symptom-provoking active sidebend, rotation and combined movement assessments of patients with low back pain. *Journal of Manual and Manipulative Therapy*, 12, 11-20.

Hay, E., Dunn, K., Hill, J., Lewis, M., Mason, E., Konstantinou, K., Sowden, G., Somerville, S., Vohora, K., Whitehurst, D., Main, C. (2008) A randomised clinical trial of subgrouping and targeted treatment for low back pain compared with best current care. The STarT Back Trial Study Protocol. *BMC Musculoskeletal Disorders*, 9, 58.

Hayden, J., Cote, P., Bombardier, C. (2006) Evaluation of the quality of prognosis studies in systematic reviews. *Annals of Internal Medicine*, 144, 427-437.

Hayden, J., Cote, P., Steenstra, I., Bombardier, C. For The QUIPS-LBP Working Group (2008) Identifying phases of investigation helps planning, appraising, and applying the results of explanatory prognosis studies. *Journal of Clinical Epidemiology*, 61, 552-560.

Haynes, B., Wilczynski, N., Mckibbon, A., Walker, C., Sinclair, J. (1994) Developing optimal search strategies for detecting clinically sound studies in MEDLINE. *Journal of the American Medical Informatics Association*, 1, 447-458.

Hefford, C. (2008) McKenzie classification of mechanical spinal pain: profile of syndromes and directions of preference. *Manual Therapy*, 13, 75-81.

Heikkila, H., Heikkila, E., Eisemann, M. (1998) Predictive factors for the outcome of a multidisciplinary pain rehabilitation programme on sick-leave and life satisfaction in patients with whiplash trauma and other myofascial pain: a follow-up study. *Clinical Rehabilitation*, 12, 487-496.

Heinrich, I., O'hare, H., Sweetman, B., Anderson, J. (1985) Validation aspects of an empirically derived classification for non-specific low back pain. *Statistician*, 34, 215-230.

Heiss, D., Fitch, D., Fritz, J., Sanchez, W., Roberts, K., Buford, J. (2004) The interrater reliability among physical therapists newly trained in a classification system for acute low back pain. *Journal of Orthopaedic and Sports Physical Therapy*, 34**,** 430-439.

Helmer, O. (1977) Problems in futures research: Delphi and causal cross-impact analysis. *Futures*, 9, 17-31.

Hendriks, H., Bekkering, G., Van Ettekoven, H., Blandsma, J., Van Der Wees, P., De Bie, R. (2000) Development and implementation of national practice guidelines: a prospect for continuous quality improvement in physiotherapy, an introduction to the method of guideline development (part 1). *Physiotherapy*, 86, 535-547.

Heneweer, H., Aufdemkampe, G., Van Tulder, M., Kiers, H., Stappaerts, K., Vanhees, L. (2007) Psychosocial variables in patients with (sub) acute low back

pain: an inception cohort in primary care physical therapy in the Netherlands. *Spine*, 32, 586-592.

Herbert, R., Jamtvedt, G., Mead, J., Birger - Hagen, K. (2005) Outcome measures measure outcomes, not effects of intervention. *Australian Journal of Physiotherapy*, 51, 3-4.

Hestbaek, L., Leboeuf-Yde, C., Manniche, C. (2003) Low back pain: what is the long-term course? A review of studies of general patient populations *European Spine Journal*, 12**,** 149-165.

Hestbaek, L., Leboeuf-Yde, C. (2000) Are chiropractic tests for the lumbo-pelvic spine reliable and valid? A systematic critical literature review. *Journal of Manipulative and Physiological Therapeutics*, 23, 258-275.

Hicks, C. (2004) *Research methods for clinical therapists: applied project design and analysis*. Edinburgh, Churchill Livingstone.

Hicks, G. (2002) *Predictive validity of clinical variables used in the determination of patient prognosis following a lumbar stabilization program*. PhD Thesis, University of Pittsburgh.

Hicks, G., Fritz, J., Delitto, A., Mcgill, S. (2005) Preliminary development of a clinical prediction rule for determining which patients with low back pain will respond to a stabilization exercise program. *Archives of Physical Medicine and Rehabilitation*, 86, 1753-1762.

Hicks, G., Fritz, J., Delitto, A., Mishock, J. (2003) Interrater reliability of clinical examination measures for identification of lumbar segmental instability. *Archives of Physical Medicine and Rehabilitation*, 84, 1858-1864.

Higginbottom, G. (2005) Sampling issues in qualitative research. *Nurse Researcher*, 12, 7-19.

Hill, J., Lewis, M., Papageorgiou, A., Dziedzic, K., Croft, P. (2004) Predicting persistent neck pain: a 1-year follow-up of a population cohort. *Spine*, 29, 1648-1654.

Hinton, R., Lincoln, A., Almquist, J., Douoguih, W., Sharma, K. (2005) Epidemiology of lacrosse injuries in high school-aged girls and boys: a 3-year prospective study. *American Journal of Sports Medicine*, 33, 1305-1314.

Hitch, B., Mugatroyd, J. (1983) Professional communications in cancer care: a Delphi survey of hospital nurses. *Journal of Advanced Nursing*, 8, 413-422.

Hoehler, F. (2000) Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity. *Journal of Clinical Epidemiology*, 53, 499-503.

Hogg-Johnson, S., Van Der Velde, G., Carroll, L., Holm, L., Cassidy, J., Guzman, J., Cote, P., Haldeman, S., Ammendolia, C., Carragee, E., Hurwitz, E., Nordin, M.,

Peloso, P. (2008) The burden and determinants of neck pain in the general population: results of the Bone and Joint Decade 2000-2010 Task Force on Neck Pain and Its Associated Disorders. *Spine*, 33, S39-51.

Hollerwoger, D. (2006) Methodological quality and outcomes of studies addressing manual cervical spine examinations: a review. *Manual Therapy*, 11, 93-98.

Holm, L., Carroll, L., Cassidy, J., Hogg-Johnson, S., Cote, P., Guzman, J., Peloso, P., Nordin, M., Hurwitz, E., Van Der Velde, G., Carragee, E., Haldeman, S. (2008) The burden and determinants of neck pain in whiplash-associated disorders after traffic collisions: results of the Bone and Joint Decade 2000-2010 Task Force on Neck Pain and Its Associated Disorders. *Spine*, 33, S52-59.

Hopwood, M., Abram, S. (1993) Factors associated with failure of lumbar epidural steroids. *Regional Anesthesia*, 18, 238-243.

Horneij, E., Hemborg, B., Johnsson, B., Ekdahl, C. (2002) Clinical tests on impairment level related to low back pain: a study of test reliability. *Journal of Rehabilitation Medicine*, 34, 176-182.

Hoving, J., De Vet, H., Koes, B., Mameren, H., Deville, W., Van Der Windt, D., Assendelft, W., Pool, J., Scholten, R., Korthals-De Bos, I., Bouter, L. (2006) Manual therapy, physical therapy, or continued care by the general practitioner for patients with neck pain: long-term results from a pragmatic randomized clinical trial. *Clinical Journal of Pain*, 22, 370-377.

Hripcsak, G., Heitjan, D. (2002) Measuring agreement in medical informatics reliability studies. Available at:

www.dbmi.columbia.edu/hripcsak/Publication/2002_JBI_Hripcsak_Kappa.doc [3rd May 2009].

Hsieh, C, Hong, C., Adams, A., Platt, K., Danielson, C., Hoehler, F., Tobis, J. (2000) Interexaminer reliability of the palpation of trigger points in the trunk and lower limb muscles. *Archives of Physical Medicine and Rehabilitation*, 81, 258-264.

Hubka, M., Phelan, S. (1994) Interexaminer reliability of palpation for cervical spine tenderness. *Journal of Manipulative and Physiological Therapeutics*, 17, 591-595.

Hudak, P., Cole, D., Haines, A. (1996) Understanding prognosis to improve rehabilitation: the example of lateral elbow pain. *Archives of Physical Medicine and Rehabilitation*, 77, 586-593.

Humphreys, K. (1990) Back Pain Syndromes – A Chiropractic Perspective. In: Fairbank, J., Pynsent, P. (eds) *Back pain: classification of syndromes.* Manchester. Manchester University Press, 73-81.

Hunt, D., Zuberbier, O., Kozlowski, A., Berkowitz, J., Schultz, I., Milner, R., Crook, J., Turk, D. (2002) Are components of a comprehensive medical assessment predictive of work disability after an episode of OLPT? *Spine*, 27, 2715-2719.

Hurri, H. (1989) The Swedish back school on chronic low back pain II: Part II Factors predicting the outcome. *Scandinavian Journal of Rehabilitation Medicine*, 21, 41-44.

Hurwitz, E, Coulter, I., Adams, A., Genovese, B., Shekelle, P. (1998) Use of chiropractic services from 1985 through 1991 in the United States and Canada. *American Journal of Public Health*, 88, 771-776.

Hurwitz, E, Carragee, E, Van Der Velde, G., Carroll, L., Nordin, M., Guzman, J., Peloso, P., Holm, L., Cote, P., Hogg-Johnson, S., Cassidy, J., Haldeman, S. (2008) Treatment of neck pain: noninvasive interventions: results of the Bone and Joint Decade 2000-2010 Task Force on Neck Pain and Its Associated Disorders. *Spine*, 33, S123-152.

Hurwitz, E., Goldstein, M., Morgenstern, H., Chiang, L. (2006) The impact of psychosocial factors on neck pain and disability outcomes among primary care patients: results from the UCLA Neck Pain Study. *Disability and Rehabilitation*, 28, 1319-1329.

Huyse F., Lyons, J., Stiefel, F., Slaets, J., De Jonge, P., Fink, P., Gans, R., Guex, P., Herzog, T., Lobo, A., Smith, G., Van Schijndel, R. (1999) "INTERMED": a method to assess health service needs. I. Development and reliability. *General Hospital Psychiatry*, 21, 39-48.

International Association For The Study Of Pain (IASP) Task Force (1994) *Classification of chronic pain*. Seattle, IASP Press.

Ijzelenberg, W., Burdorf, A. (2004) Patterns of care for low back pain in a working population. *Spine*, 29, 1362-1368.

Infante-Rivard, C., Lortie, M. (1997) Relapse and short sickness absence for back pain in the six months after return to work. *Occupational and Environmental Medicine*, 54, 328-334.

Ingemarsson, A., Nordholm, L., Sivik, T. (1997) Risk of long-term disability among patients with back pain. *Scandinavian Journal of Rehabilitation Medicine*, 29, 205-212.

Jackson, D. (2001) How is low back pain managed? Retrospective study of the first 200 patients with low back pain referred to a newly established community – based physiotherapy department. *Physiotherapy*, 87, 573-581.

Jairath, N., Weinstein, J. (1994) The Delphi methodology: a useful administrative approach. *Canadian Journal of Nursing Administration*, 7, 29-42.

Jakobsson, L., Norin, H., Bunketorp, O. (2003) Whiplash-associated disorders in frontal impacts: influencing factors and consequences. *Traffic Injury Prevention*, 4, 153-161.

Janis, I. (1982) *Groupthink. psychological studies of policy decisions and fiascoes* 2nd edition. Boston, Houghton – Mifflin.

Jellema, P., Van Der Horst, H., Vlaeyen, J., Stalman, W., Bouter, L., Van Der Windt, D. (2006) Predictors of outcome in patients with (sub)acute low back pain differ across treatment groups. *Spine*, 31, 1699-1705.

Jensen. C. (2003) Development of neck and hand-wrist symptoms in relation to duration of computer use at work. *Scandinavian Journal of Work, Environment and Health*, 29, 197-205.

Jensen, I., Bradley, L., Linton, S. (1989) Validation of an observation method of pain assessment in non-chronic back pain. *Pain*, 39, 267-274.

Jill, N., Blackman, M. (2004) Reproducibility of clinical data II: categorical outcomes. *Pharmaceutical Statistics*, 3, 109-122.

Johnston, V., Jimmieson, N., Jull, G., Souvlis, T. (2008) Contribution of individual, workplace, psychosocial and physiological factors to neck pain in female office workers. *European Journal of Pain*, 13, 985-991 .

Johnston, W., Allan, B., Neff, D., Rosen, M., Sills, L., Thomas, S. (1983) Inter-examiner study of palpation in detecting location of spinal segmental dysfunction. *Journal of the American Osteopathic Association*, 82, 839-845.

Jonsson, B., Persson, J., Kilborn, A. (1988) Disorders of the cervicobrachial region among female workers in the electronics industry: a two-year follow-up. *International Journal of Industrial Ergonomics*, 3, 1-12.

Jordan, K. (2000) Assessment of published reliability studies for cervical spine range-of-motion measurement tools. *Journal of Manipulative and Physiological Therapeutics*, 23, 180-195.

Jordan, K. (1996) *Length of disability in the rehabilitation of chronic spinal disorder patients: treatment outcomes and predictor variables*. PhD Thesis, University of Texas South Western Medical Center at Dallas.

Jull, G. (2004) Editorial: Classification systems in clinical practice. *Physiotherapy Research International*, 9, iii-iv.

Jull, G., Sterling, M. (2009) Bring back the biopsychosocial model for neck pain disorders. *Manual Therapy*, 14, 117-118.

Jull, G., Treleaven, J., Versace, G. (1994) Manual examination: Is pain provocation a major cue for spinal dysfunction? *Australian Journal of Physiotherapy*, 40, 159-165.

Kamper, S., Rebbeck T., Maher, C., Mcauley, J., Sterling M. (2008) Course and prognostic factors of whiplash: a systematic review and meta-analysis. *Pain*, 138, 617-629.

Kaptain, G., Shaffrey, C., Alden, T., Young, J., Laws, E. Jr., Whitehill, R. (1999) Secondary gain influences the outcome of lumbar but not cervical disc surgery. *Surgical Neurology*, 52, 217-223.

Karas, R., McIntosh, G., Hall, H., Wilson, L., Melles, T. (1997) The relationship between nonorganic signs and centralisation of symptoms in the prediction of return to work for patients with low back pain. *Physical Therapy*, 77, 354-360.

Kasch, H., Bach, F., Jensen, T. (2001) Handicap after acute whiplash injury: a 1-year prospective study of risk factors. *Neurology*, 56, 1637-1643.

Keating, J., Jr., Bergmann, T., Jacobs, G., Finer, B., Larson, K. (1990) Interexaminer reliability of eight evaluative dimensions of lumbar segmental abnormality. *Journal of Manipulative and Physiological Therapeutics*, 13, 463-470.

Keefe, F., Bradley, L., Crisson, J. (1990) Behavioral assessment of low back pain: identification of pain behaviour subgroups. *Pain*, 40, 153-160.

Keeney, S., Hasson, F., Mckenna, H. (2001) A critical review of the Delphi panel technique as a research methodology for nursing. *International Journal of Nursing Studies*, 38, 195-200.

Kelsey, J, Githens, P., Walter, S., Southwick, W., Weil, U., Holford, T., Ostfeld, A., Calogero, J., O'connor, T., White, A. (1984) An epidemiological study of acute prolapsed cervical intervertebral disc. *Journal of Bone, Joint, Surgery* (American volume), 66, 907-914.

Kent, P., Keating, J. (2008) Can we predict poor recovery from recent-onset nonspecific low back pain? A systematic review. *Manual Therapy*, 13, 12-28.

Kent, P., Keating, J. (2005) Classification in non-specific low back pain: what methods do primary care clinicians currently use? *Spine*, 30, 1433-1440.

Kent, P., Keating, J. (2004) Do primary –care clinicians think that non-specific low back pain is one condition? *Spine*, 29, 1022-1031.

Kent, P., Keating, J., Buchbinder, R. (2008) Searching for a conceptual framework for nonspecific low back pain. *Manual Therapy*, 14, 387-396.

Kent, P., Keating, J., Taylor, N. (2009) Primary care clinicians use variable methods to assess acute non-specific low back pain and usually focus on impairments. *Manual Therapy*, 14, 88-100.

Kernan, W., Feinstein, A., Brass, L. (1991) A methodological appraisal of research on prognosis after transient ischemic attacks. *Stroke*, 22, 1108-1116.

Key, J., Clift, A., Condie, F., Harley, C. (2008a) A model of movement dysfunction provides a classification system guiding diagnosis and therapeutic care in spinal pain and related musculoskeletal syndromes: a paradigm shift – Part 1. *Journal of Bodywork and Movement Therapies*, 12, 7-21.

Key, J., Clift, A., Condie, F., Harley, C. (2008b) A model of movement dysfunction provides a classification system guiding diagnosis and therapeutic care in spinal pain and related musculoskeletal syndromes: a paradigm shift – Part 2. *Journal of Bodywork and Movement Therapies*, 12, 105-120.

Key, J., Hodgson, S., Omar, R., Jensen, T., Thompson, S., Boobis, A., Davies, D., Elliott, P. (2006) Meta-analysis of studies of alcohol and breast cancer with consideration of the methodological issues. *Cancer Causes Control*, 17, 759-770.

Kilby, J., Stigant, M., Roberts, A. (1990) The reliability of back pain assessment by physiotherapists, using a "McKenzie algorithm". *Physiotherapy*, 76, 579-583.

Kilpikoski, S., Alen, M., Paatelma, M., Simonen, R., Heinonen, A., Videman, T. (2009) Outcome comparison among working adults with centralizing low back pain: Secondary analysis of a randomized controlled trial with 1-year follow-up. *Advances in Physiotherapy*, 1651-1948.

Kilpikoski, S., Airaksinen, O., Kankaanpaa, M., Leminen, P., Videman, T., Alen, M. (2002) Interexaminer reliability of low back pain assessment using the McKenzie method. *Spine*, 27, E207-214.

Kirkness, C., Korner-Bitensky, N. (2002) Prevalence of outcome measure use by physiotherapists in the management of low back pain. *Physiotherapy Canada*, 54, 249-257.

Kjellman, G., Oberg, B. (2002) A randomized clinical trial comparing general exercise, McKenzie treatment and a control group in patients with neck pain. *Journal of Rehabilitation Medicine*, 34, 183-190.

Klaber-Moffett, J., Jackson, D., Gardiner, E., Torgerson, D., Coulton, S., Eaton, S., Mooney, M., Pickering, C., Green, A., Walker, L., May, S., Young, S. (2006) Randomized trial of two physiotherapy interventions for primary care neck and back pain patients: 'McKenzie vs brief physiotherapy pain management. *Rheumatology*, 45, 1514-1521.

Klapow, J., Slater, M., Patterson, T., Atkinson, J., Weickgenant, A., Grant, I., Garfin, S. (1995) Psychosocial factors discriminate multidimensional clinical groups of chronic low back pain patients. *Pain*, 62, 349-355.

Klapow, J., Slater, M, Patterson, T., Doctor, J., Atkinson, J., Garfin, S. (1993) An empirical evaluation of multidimensional clinical outcome in chronic low back pain patients. *Pain*, 55, 107-118.

Klenerman, L., Slade, P., Stanley, I., Pennie, B., Reilly, J., Atchison, L., Troup, J., Rose, M. (1995) The prediction of chronicity in patients with an acute attack of LBP in a general practice setting. *Spine*, 20, 478-484.

Koes, B., Van Tulder, M., Ostelo, R., Kim Burton, A., Waddell, G. (2001) Clinical guidelines for the management of low back pain in primary care: an international comparison. *Spine*, 26, 2504-2513.

Koes, B., Van Tulder, M., Thomas, S. (2006) Diagnosis and treatment of low back pain. *British Medical Journal*, 332, 1430-1434.

Kool, J., Oesch, P., De Bie, R. (2002) Predictive tests for non-return to work in patients with chronic low back pain. *European Spine Journal*, 11, 258-266.

Koopman, F., Edelaar, M., Slikker, R., Reynders, K., Van Der Woude, L., Hoozemans, M. (2004) Effectiveness of a multidisciplinary occupational training program for chronic low back pain: a prospective cohort study. *American Journal of Physical Medicine and Rehabilitation*, 83, 94-103.

Kopp, J., Alexander, A., Turocy, R., Levrini, M., Lichtman, D. (1986) The use of lumbar extension in the evaluation and treatment of patients with acute herniated nucleus pulposus. A preliminary report. *Clinical Orthopedics and Related Research*, 202, 211-218.

Korbon, G., Degood, D., Schroeder, M., Schwartz, D., Shutty, M. Jr. (1987) The development of a Somatic Amplification Rating Scale for low-back pain. *Spine*, 12, 787-791.

Korhonen, T., Ketola, R., Toivonen, R., Luukkonen, R., Hakkanen, M., Viikari-Juntura, E. (2003) Work related and individual predictors for incident neck pain among office employees working with video display units. *Occupational and Environmental Medicine*, 60, 475-482.

Kraemer, H. (1980) Extension of the kappa coefficient. *Biometrics*, 36, 207-216.

Kraemer, H. (1979) Ramifications of a population model for k as a coefficient of reliability. *Psychometrika*, 44, 461-472.

Kraemer, H., Bloch, D. (1988) Kappa coefficients in epidemiology: an appraisal of a reappraisal. *Journal of Clinical Epidemiology*, 41, 959-968.

Kraemer, H., Periyakoil, V., Noda, A. (2002) Kappa coefficients in medical research. *Statistics in Medicine*, 21, 2109-2129.

Kraft, M., Kullgren, A., Tingwall, C., Bostrom, O., Fredriksson, R. (2000) How crash severity in rear impacts influences short- and long-term consequences to the neck. *Accident Analysis and Prevention*, 32, 187-195.

Kraft, M., Kullgren, A., Ydenius, A., Tingvall, C. (2002) Influence of crash pulse characteristics on whiplash associated disorders in rear impacts – crash recording in real life crashes. *Traffic Injury Prevention*, 3, 141-149.

Krause, N., Ragland, D. (1994) Occupational disability due to low back pain: a new interdisciplinary classification based on a phase model of disability. *Spine*, 19, 1011-1020.

Krippendorff, K. (1980) *Content Analysis: An Introduction to its methodology*. Thousand Oaks, California, Sage.

Kuorinka, I., Jonsson, B., Kilborn, A., Vinterberg, H., Biering-Sorensen, F., Andersson, G., Jorgensen, K. (1987) Standardised Nordic questionnaires for the analysis of musculoskeletal symptoms. *Applied Ergonomics*, 18, 233-237.

Kuper, A., Lingard, L., Levinson, W. (2008a) Critically appraising qualitative research. *British Medical Journal*, 337, 1035. Available at: http://www.bmj.com/cgi/content/full/337/aug07_3/a1035 [15th September 2008].

Kuper, A., Reeves, S., Levinson, W. (2008b) An introduction to reading and appraising qualitative research. *British Medical Journal*, 337, 404-407.

Kyzas, P., Loizou, K., Ioannidis, J. (2005) Selective reporting biases in cancer prognostic factor studies. *Journal of the National Cancer Institute*, 97, 1043-1055.

Lancourt, J., Kettelhut, M. (1992) Predicting RTW for lower back pain patients. *Spine*, 17, 629-640.

Lacroix, J., Powell, J., Lloyd, G., Doxey, N., Mitson, G., Aldam, C. (1990) Low-back pain. Factors of value in predicting outcome. *Spine*, 15, 495-499.

Landis, J., Koch, G. (1977a) A one-way components of variance model for categorical data. *Biometrics*, 33, 671-679.

Landis, J., Koch, G. (1977b) The measurement of observer agreement for categorical data. *Biometrics*, 33, 159-174.

Langworthy, J., Breen, A. (1997) Rationalizing back pain: the development of a classification system through cluster analysis. *Journal of Manipulative and Physiological Therapeutics*, 20, 303-310.

Lanier, D., Stockton, P. (1988) Clinical predictors of outcomes of acute episodes of low back pain. *Journal of Family Practice*, 27, 483-489.

Lankhorst, G., Van De Stadt, R., Vogelaar, T., Van Der Korst, J., Prevo, A. (1982) Objectivity and repeatability of measurements in low back pain. *Scandinavian Journal of Rehabilitation Medicine*, 14, 21-26.

Laslett, M., Aprill, C., Mcdonald, B., Oberg, B. (2006a) Clinical predictors of lumbar provocation discography: a study of clinical predictors of lumbar provocation discography. *European Spine Journal*, 15, 1473-1484.

Laslett, M., Mcdonald, B., Aprill, C., Tropp, H., Oberg, B. (2006b) Clinical predictors of screening lumbar zygapophyseal joint blocks: development of clinical prediction rules. *Spine Journal*, 6, 370-379.

Laslett, M., Oberg, B., Aprill, C., Mcdonald, B. (2005) Centralization as a predictor of provocation discography results in chronic low back pain, and the influence of disability and distress on diagnostic power. *Spine Journal*, 5, 370-380.

Laslett, M., Van Wijmen P. (1999) Low back and referred pain: diagnosis and proposed new system of classification. *New Zealand Journal of Physiotherapy*, 27, 5-14.

Laslett, M., Williams, M. (1994) The reliability of selected pain provocation tests for sacroiliac joint pathology. *Spine*, 19, 1243-1249.

Laslett, M., Young, S., Aprill, C., Mcdonald, B. (2003) Diagnosing painful sacroiliac joints: a validity study of a McKenzie evaluation and sacroiliac provocation tests. *Australian Journal of Physiotherapy*, 49, 89-97.

Laupacis, A., Sekar, N., Stiell, I. (1997) Clinical prediction rules. A review and suggested modifications of methodological standards. *Journal of the American Medical Association*, 277, 488-494.

Laupacis, A., Wells, G., Richardson, W., Tugwell, P. (1994) Users' guides to the medical literature. V. How to use an article about prognosis. Evidence-Based Medicine Working Group. *Journal of the American Medical Association*, 272**,** 234-237.

Leboeuf, C. (1989) Orthopedic examination procedures: a reliability and consistency study. *Journal of the Australian Chiropractic Association*, 19, 75-77.

Leboeuf, C. (1991) The reliability of specific sacro-occipital technique diagnostic tests. *Journal of Manipulative and Physiological Therapeutics*, 14, 512-517.

Leboeuf, C., Gardner, V., Carter, A., Scott, T., (1989) Chiropractic examination procedures: a reliability and consistency study. *Journal of the Australian Chiropractic Association*, 19, 101-104.

Leboeuf-Yde, C., Axen, I., Jones, J., Rosenbaum, A., Lovgren, P., Halasz, L., Larsen, K. (2005) The Nordic Back Pain Subpopulation program: the long-term outcome pattern in patients with low back pain treated by chiropractors in Sweden. *Journal of Manipulative and Physiological Therapeutics*, 28, 472-478.

Leboeuf-Yde, C., Lauritsen, J., Lauritzen, T. (1997) Why has the search for causes of low back pain largely been inconclusive? *Spine*, 22, 877-881.

Lehmann, T., Spratt K., Lehmann, K. (1993) Predicting long-term disability in low back injured workers presenting to a spine consultant. *Spine*, 18, 1103-1112.

Li, L., Bombardier, C. (2001) Physical therapy management of low back pain: an exploratory survey of therapist approaches. *Physical Therapy*, 81, 1018-1028.

Liddle, S., Gracey, J., Baxter, G. (2007) Advice for the management of low back pain: a systematic review of randomised controlled trials. *Manual Therapy*, 12, 310-327.

Light, R. (1971) Measures of response agreement for qualitative data: some generalizations and alternatives. *Psychological Bulletin*, 76, 365-377.

Lijmer, J., Mol, B., Heisterkamp, S., Bonsel, G., Prins, M., Van Der Meulen, J., Bossuyt, P. (1999) Empirical evidence of design-related bias in studies of diagnostic tests. *Journal of the American Medical Association*, 282, 1061–1066.

Lindsay, D., Meeuwisse, W., Mooney, M., Summersides, J. (1995) Interrater reliability of manual therapy assessment techniques. *Physiotherapy Canada*, 47, 173-180.

Linstone, H., Turoff, M. (eds) (1975) *The Delphi method: techniques and applications*. Reading, MA, Addison-Wesley.

Lindstrom, I., Ohlund, C., Nachemson, A. (1995) Physical performance, pain, pain behavior and subjective disability in patients with subacute low back pain. *Scandinavian Journal of Rehabilitation Medicine*, 27, 153-160.

Lloyd, D., Troup, J. (1983) Recurrent back pain and its prediction. *Journal of the Society of Occupational Medicine*, 33, 66-74.

Lock, C., Allgar, V., Jones, K., Marples, G., Chandler, C., Dawson, P. (1999) Prevalence of back, neck and shoulder problems in the inner city: implications for the provision of physiotherapy services in primary healthcare. *Physiotherapy Research International*, 4, 161-169.

Loisel, P., Vachon, B., Lemaire, J., Durand, M., Poitras, S., Stock, S., Tremblay, C. (2002) Discriminative and predictive validity assessment of the Quebec Task Force Classification. *Spine*, 27, 851-857.

Long, A. (2006) Re: research. Email sent to Angeliki Chorti (physio_glaraki@yahoo.com), 18th September 2006.

Long, A. (1995) The centralisation phenomenon. Its usefulness as a predictor or outcome in conservative treatment of chronic low back pain (a pilot study). *Spine*, 20, 2513-2520.

Long, A., Donelson, R., Fung, T. (2004) Does it matter which exercise? A randomized control trial of exercise for low back pain. *Spine*, 29, 2593-2602.

Long, A., May, S., Fung, T. (2008a) Specific directional exercises for patients with low back pain: a case series. *Physiotherapy Canada*, 60, 307-317.

Long, A., May, S., Fung, T. (2008b) The comparative value of directional preference and centralization: a useful tool for frontline clinicians? *Journal of Manual and Manipulative Therapy*, 16, 248-254.

Lorish, T., Rizzo, T. Jr., Ilstrup, D., Scott, S. (1992) Injuries in adolescent and preadolescent boys at two large wrestling tournaments. *American Journal of Sports Medicine*, 20, 199-202.

Lundberg, G., Gerdle, B. (1999) The relationships between spinal sagittal configuration, joint mobility, general low back mobility and segmental mobility in female homecare personnel. *Scandinavian Journal of Rehabilitation Medicine,* 31, 197-206.

Ludbrook, J. (2002) Statistical techniques for comparing measurers and methods of measurement: a critical review. *Clinical and Experimental Pharmacology & Physiology*, 29, 527-536.

Luime, J. (2005) Study protocol: Pilot study Early Rehabilitation of Neck and Shoulder Pain. Coventry, University of Warwick.

Luime, J., Koes, B., Miedem, H., Verhaar, J., Burdorf, A. (2005) High incidence and recurrence of shoulder and neck pain in nursing home employees was demonstrated during a 2-year follow-up. *Journal of Clinical Epidemiology*, 58, 407-413.

Luime, J., Kuiper, J., Koes, B., Verhaar, J., Miedema, H., Burdorf, A. (2004) Work – related risk factors for the incidence and recurrence of shoulder and neck complaints among nursing-home and elderly-care workers. *Scandinavian Journal of Work, Environment and Health*, 30, 279-286.

Lyle M., Manes, S., Mcguinness, M., Ziaei, S., Iversen, M. (2005) Relationship of physical examination findings and self-reported symptom severity and physical function in patients with degenerative lumbar conditions. *Physical Therapy*, 85, 120-133.

Macdonald, R. (1990) An osteopathic view of back pain classification. In: Fairbank, J., Pynsent, P. *Back pain: Classification of Syndromes.* Manchester, Manchester University Press, 62-72.

Macfarlane, G., Thomas, E., Croft, P., Papageorgiou, A., Jayson, M., Silman, A. (1999) Predictors of early improvement in low back pain amongst consulters to general practice: the influence of pre-morbid and episode –related factors. *Pain*, 80, 113-119.

Machado, L., De Souza, M., Ferreira, P., Ferreira, M. (2006) The McKenzie method for low back pain: a systematic review of the literature with a meta-analysis approach. *Spine,* 31, E254-262.

Maclure, M., Willett, W. (1987) Misinterpretation and misuse of the kappa statistic. *American Journal of Epidemiology*, 126, 161-169.

Magee, D. (2002) *Orthopedic physical assessment*. Philadelphia, Saunders.

Magnusson, M., Pope, M., Hansson, T. (1995) Does hyperextension have an unloading effect on the intervertebral disc? *Scandinavian Journal of Rehabilitation Medicine*, 27, 5-9.

Maher, C., Adams, R. (1994) Reliability of pain and stiffness assessments in clinical manual lumbar spine examination. *Physical Therapy*, 74, 801-809.

Maher, C., Latimer, J. (1992) Pain or resistance – the manual therapists' dilemma. *Australian Journal of Physiotherapy*, 38, 257-260.

Main, C., Watson, P. (1995) Screening for patients at risk of developing chronic incapacity. *Journal of Occupational Rehabilitation*, 5, 207-217.

Main, C., Wood, P., Hollis, S., Spanswick, C., Waddell, G. (1992) The distress and risk assessment method. A simple patient classification to identify distress and evaluate the risk of poor outcome. *Spine*, 17, 42-52.

Maitland, G. (1986) Selection of techniques. In: *Vertebral Manipulation.* 5th edition. London, Butterworth & Co Ltd, 115-143.

Maitland, G., Edwards, B. (1986) Examination. In: *Vertebral Manipulation.* 5th edition. London, Butterworth & Co Ltd, 43-92.

Mallen, C., Peat, G., Thomas, E., Dunn, K., Croft, P. (2007) Prognostic factors for musculoskeletal pain in primary care: a systematic review. *British Journal of General Practice*, 57, 655-661.

Manca, A., Dumville, J., Torgeson, D., Klaber-Moffett, J., Mooney, M., Jackson, D., Eaton, S. (2007) Randomised trial of two physiotherapy interventions for primary care back and neck pain patients: cost effectiveness analysis. *Rheumatology*, 46, 1495-1501.

Marcus, M., Gerr, F., Monteilh, C., Ortiz, D., Gentry, E., Cohen, S., Edwards, A., Ensor, C., Kleinbaum, D. (2002) A prospective study of computer users: II. Postural risk factors for musculoskeletal symptoms and disorders. *American Journal of Industrial Medicine*, 41, 236-249.

Marras, W., Pamianpour, M., Ferguson, S., Kim, J., Crowell, R., Bose, S., Simon, S. (1995) The classification of anatomic- and symptom – based low back disorders using motion measure models. *Spine*, 20, 2531-2546.

Matyas, T., Bach, T. (1985) The reliability of selected techniques in clinical arthrometrics. *Australian Journal of Physiotherapy*, 31, 175-199.

May, S. (2003) An outcome audit for musculoskeletal patients in primary care.

Physiotherapy Theory and Practice, 19, 189-198.

May, S. (2006) Re: PhD talk. Email sent to Angeliki Chorti (physio_glaraki@yahoo.com), 31st May 2006.

May, S. (2006) Re: Topics for discussion. Email sent to Angeliki Chorti (physio_glaraki@yahoo.com), 17th March 2006.

May, S., Donelson, R. (2008) Evidence – informed management of chronic low back pain with the McKenzie method. *Spine Journal*, 8, 134-141.

May, S., Gardiner, E., Young, S., Klaber-Moffett, J. (2008a) Predictor variables for a positive long-term functional outcome in patients with acute and chronic neck and back pain treated with a McKenzie approach: a secondary analysis. *Journal of Manual and Manipulative Therapy*, 16, 155-160.

May, S., Littlewood, C., Bishop, A. (2006) Reliability of procedures used in the physical examination of non-specific low back pain: a systematic review. *Australian Journal of Physiotherapy*, 52, 91-102.

May, S., Rosedale, R. (2007) A case of a potential manipulation responder whose back pain resolved with flexion exercises. *Journal of Manipulative and Physiological Therapeutics*, 30, 539-542.

May, S., Rosedale, R., Hoyt, K., Marley, J. (2008b) Clinical prediction rules and sub-classification for low back pain: is there a trump card? *International Journal of Mechanical Diagnosis and Therapy*, 3, 14-18.

McCarthy, C., Arnall, F., Strimpakos, N., Freemont, A., Oldham, J. (2004) The biopsychosocial classification of non-specific low back pain: a systematic review. *Physical Therapy Reviews*, 9, 17-30.

McCarthy, C., Gittins, M., Roberts, C., Oldham, J. (2007a) The reliability of the clinical tests and questions recommended in international guidelines for low back pain. *Spine*, 32, 921-926.

McCarthy, C., Rushton, A., Billis, V., Arnall, F., Oldham, J. (2006) Development of a clinical examination in non-specific low back pain: a Delphi technique. *Journal of Rehabilitation Medicine*, 38, 263-267.

McCarthy, M., Grevitt, M., Silcocks, P., Hobbs, G. (2007b) The reliability of the Vernon and Mior neck disability index, and its validity compared with the short form -36 health survey questionnaire. *European Spine Journal*, 16, 2111-2117.

McCombe, P., Fairbank, J., Cockersole, B., Pynsent, P. (1989) 1989 Volvo Award in clinical sciences. Reproducibility of physical signs in low-back pain. *Spine*, 14, 908-918.

McConnell, D., Beal, M., Dinnar, U., Goodridge, J., Johnston, W., Karni, Z., Upledger, J., Blum, J. (1980) Low agreement of findings in neuromusculoskeletal

examinations by a group of osteopathic physicians using their own procedures. *Journal of the American Osteopathic Association*, 79, 441-450.

McCormack, R., Inman, R., Wells, A., Berntsen, C., Imbus, H. (1990) Prevalence of tendinitis and related disorders of the upper extremity in a manufacturing workforce. *Journal of Rheumatology*, 17, 958-964.

McIntosh, G., Frank, J., Hogg - Johnson, S., Bombardier, C., Hall, H. (2000) Prognostic factors for time receiving workers' compensation benefits in a cohort of patients with low back pain. *Spine*, 25, 147-157.

McIntosh, G., Hall, H., Boyle, C. (2008) *Outcomes based on a low back pain classification system.* International Society for the Study of the Lumbar Spine. Geneva, Switzerland.

McKenna, H. (1994) The Delphi technique: a worthwhile approach for nursing? *Journal of Advanced Nursing*, 19, 1221-1225.

McKenzie, R. (1990) *The cervical and thoracic spine: mechanical diagnosis and therapy,* Waikanae, New Zealand, Spinal Publications.

McKenzie, R. (1981) *The lumbar spine: mechanical diagnosis and therapy,* Waikanae, New Zealand, Spinal Publications.

McKenzie, R., May, S. (2006) *The cervical and thoracic spine: mechanical diagnosis and therapy*, Waikanae, New Zealand, Spinal publications.

McKenzie, R., May, S. (2003) *The lumbar spine: mechanical diagnosis and therapy,* Waikanae, New Zealand, Spinal publications.

McNeill, T., Sinkora G., Leavitt F. (1986) Psychologic classification of low-back pain patients: a prognostic tool. *Spine*, 11, 955-999.

McPartland, J., Goodridge, J. (1997) Counterstrain and traditional osteopathic examination of the cervical spine compared. *Journal of Bodywork and Movement Therapies,* 1, 173-178.

McShane, L., Altman, D., Sauerbrei, W., Taube, S., Gion, M., Clark, G. For The Statistics Subcommittee On The Nci-Eortc Working Group On Cancer Diagnostics (2005) REporting recommendations for tumour MARKer prognostic studies (REMARK). *British Journal of Cancer*, 93, 387-391.

Mead, D., Moseley, L. (2001) The use of Delphi as a research approach. *Nurse Researcher,* 8, 4-23.

Mendelson, G., Selwood, T., Kranz, H., Loh, T., Kidson, M., Scott, D. (1983) Acupuncture treatment of chronic back pain. A double-blind placebo-controlled trial. *American Journal of Medicine,* 74, 49-55.

Mercer, C., Jackson, A., Hettinga, D., Barlos, P., Ferguson, S., Greenhalgh, S. (2006) *Clinical guidelines for the physiotherapy management of persistent low back pain. Part 1: Exercise*. London, Chartered Society of Physiotherapy.

Mercer, S., Jull, J. (1996) Morphology of the cervical intervertebral disc: implications for McKenzie's model of the disc derangement syndrome. *Manual Therapy*, 2, 76-81.

Mercer, S., Rivett, D. (2004) Editorial: Clinical anatomy serving manual therapy. *Manual Therapy*, 9, 59.

Michaelson, P., Sjolander, P., Johansson, H. (2004) Factors predicting pain reduction in chronic back and neck pain after multimodal treatment. *Clinical Journal of Pain*, 20, 447-454.

Miettinen, T., Airaksinen, O., Lindgren, K., Leino, E. (2004) Whiplash injuries in Finland – the possibility of some sociodemographic and psychosocial factors to predict the outcome after one year. *Disability and Rehabilitation*, 26, 1367-1372.

Mikhail, C., Korner-Bitensky, N., Rossignol, M., Durnas, J. (2005) Physical therapists' use of interventions with high evidence of effectiveness in the management of a hypothetical typical patient with acute low back pain. *Physical Therapy*, 85, 1151-1167.

Miller, G. (1956) The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, 63, 81-87.

Miro, J., Nieto, R., Huguet, A. (2008) Predictive factors of chronic pain and disability in whiplash: a Delphi poll. *European Journal of Pain*, 12, 30-47.

Moffroid, M., Haugh, L., Henry, S., Short, B. (1994) Distinguishable groups of musculoskeletal low back pain patients and asymptomatic control subjects based on physical measures of the NIOSH Low Back Atlas. *Spine*, 19, 1350-1358.

Moher, D., Cook, D., Eastwood, S., Olkin, I., Rennie, D., Stroup, D. (1999) Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. Quality of Reporting of Meta-analyses. *Lancet*, 354, 1896-900.

Mokkink, L., Terwee, C., Knol, D., Stratford, P., Alonso, M., Patrick, D., Bouter, L. (2008) *COSMIN Checklist & Feedback report.* Amsterdam, VU University Medical Centre.

Mooney, V. (1989) The classification of low back pain. *Annals of Medicine*, 21, 321-325.

Mullen, P. (2003) Delphi: myths and reality. *Journal of Health Organization and Management*, 17, 37-52.

Muller, C., Monrad, T., Biering - Sorensen, F., Darre, E., Deis, A., Kryger, P. (1999) The influence of previous low back trouble, general health and working conditions on future sick-listing because of low back trouble. A 15-year follow-up study of risk indicators for self-reported sick-listing caused by low back trouble. *Spine*, 24, 1562-1570.

Murphy, K., Cornish, R. (1984) Prediction of chronicity in acute low back pain. *Archives of Physical Medicine & Rehabilitation*, 65, 334-337.

Murphy, D., Hurwitz, E. (2007) A theoretical model for the development of a diagnosis – based clinical decision rule for the management of patients with spinal pain. *BMC Musculoskeletal Disorders*, 8, 75.

Murphy, D., Hurwitz, E., Nelson, C. (2008) A diagnosis-based clinical decision rule for spinal pain: review of the literature. *Chiropractic & Osteopathy*, 16, 7.

Murphy, L., Reinsch, S., Najm, W., Dickerson, V., Seffinger, M., Adams, A., Mishra, S. (2003a) Searching biomedical databases on complementary medicine: the use of controlled vocabulary among authors, indexes and investigators. *BMC Complementary and Alternative Medicine*, 3, 3.

Murphy, L., Reinsch, S., Najm, W., Dickerson, V., Seffinger, M., Adams, A., Mishra, S. (2003b) Spinal palpation: the challenges of information retrieval using available databases. *Journal of Manipulative and Physiological Therapeutics*, 26, 374-382.

Murphy, M., Black, N., Lamping, D., Mckee, C., Sanderson, C., Askham, J., Marteau, T. (1998) Consensus development methods, and their use in clinical guideline development. *Health Technology Assessment*, 2, 1-88.

Murry, J., Hammons, J. (1995) Delphi: a versatile methodology for conducting qualitative research. *Review of Higher Education*, 18, 423-436.

Myburgh, C., Larsen, A., Hartvigsen, J. (2008) A systematic, critical review of manual palpation for identifying myofascial trigger points: evidence and clinical significance. *Archives of Physical Medicine and Rehabilitation*, 89, 1169-1176.

Nachemson, A, Vingard, E. (2000) Assessment of Patients with Neck and Back Pain: A Best Evidence Synthesis. In: Nachemson A., Jonsson E. *Neck and back pain. The scientific evidence of causes, diagnosis, and treatment*. Philadelphia, Lippincott Williams & Wilkins.

Najm, W., Seffinger, M., Mishra, S., Dickerson, V., Adams, A., Reinsch, S., Murphy, L., Goodman, A. (2003) Content validity of manual spinal palpatory exams - A systematic review. *BMC Complementary and Alternative Medicine*, 3, 1.

Nam, J. (2000) Interval estimation of the kappa coefficient with binary classification and an equal marginal probability model. *Biometrics*, 56, 583-585.

Nederhand, M., Hermens, H., Ijzerman, M., Turk, D., Zilvold, G. (2003) Chronic neck pain disability due to an acute whiplash injury. *Pain*, 102, 63-71.

Nederhand, M., Ijzerman, M., Hermens, H., Turk, D., Zilvold, G. (2004) Predictive value of fear avoidance in developing chronic neck pain disability: consequences for clinical decision making. *Archives of Physical Medicine and Rehabilitation,* 85, 496-501.

Nelson, M, Allen, P., Clamp, S., De Dombal, F. (1979) Reliability and reproducibility of clinical findings in low-back pain. *Spine,* 4, 97-101.

Newton, W., Curtis, P., Witt, P., Hobler, K. (1997) Prevalence of subtypes of low back pain in a defined population. *Journal of Family Practice,* 45, 331-335.

New Zealand Guideline Group (2004) *New Zealand Acute Low Back Pain Guide*, Wellington, ACC.

Nice, D., Riddle, D., Lamb, R., Mayhew, T., Rucker, K. (1992) Intertester reliability of judgments of the presence of trigger points in patients with low back pain. *Archives of Physical Medicine and Rehabilitation*, 73, 893-898.

Nicholas, M. (2008) Pain management in musculoskeletal conditions. *Best Practice and Research Clinical Rheumatology*, 22, 451-470.

Niemisto, L., Sarna, S., Lahtinen-Suopanki, T., Lindgren, K., Hurri, H. (2004) Predictive factors for 1-year outcome of chronic low back pain following manipulation, stabilizing exercises, and physician consultation or physician consultation alone. *Journal of Rehabilitation Medicine,* 36, 104-109.

Njoo, K., Van Der Does, E. (1994) The occurrence and inter-rater reliability of myofascial trigger points in the quadratus lumborum and gluteus medius: a prospective study in non-specific low back pain patients and controls in general practice. *Pain*, 58, 317-323.

Nordin, M. (2009) Personal communication, Occupational and Industrial Orthopaedic Centre, New York University Hospital of Joint Diseases, New York, USA, June 2009.

Nordin, M., Carragee, E., Hogg-Johnson, S., Schecter-Weiner, S., Hurwitz, E., Peloso, P., Guzman, J., Van Der Velde, G., Carroll, L., Holm, L., Cote, P., Cassidy, J., Haldeman, S. (2008) Assessment of neck pain and its associated disorders: Results of the Bone and Joint Decade 2000-2010 Task Force on Neck Pain and its Associated Disorders *Spine*, 33, S101-122.

Nordin, M., Skovron M., Hiebert, R., Weiser, S., Brisson, P., Campello, M., Harwood, K., Crane, M., Lewis, S. (1997) Early predictors of delayed return to work in patients with low back pain. *Journal of Musculoskeletal Pain*, 5, 5-27.

Norris, S., Watt, I. (1983) The prognosis of neck injuries resulting from rear-end vehicle collisions. *Journal of Bone, Joint, Surgery* (British volume), 65, 608-611.

Nunnally, J.(1967) *Psychometric theory.* New York, London, McGraw-Hill.

Nyiendo, J., Haas, M., Goodwin, P. (2000) Patient characteristics, practice activities, and one-month outcomes for chronic, recurrent low-back pain treated by chiropractors and family medicine physicians: a practice-based feasibility study. *Journal of Manipulative and Physiological Therapeutics*, 23, 239-245.

Obelieniene, D., Schrader, H., Bovim, G., Miseviciene, I., Sand, T. (1999) Pain after whiplash: a prospective controlled inception cohort study. *Journal of Neurology, Neurosurgery and Psychiatry*, 66, 279-283.

O'hearn, M. (1997) Physical therapy treatment of low back pain: a report of outcomes according to three types of patient classification. *Journal of Rehabilitation Outcomes Measurement*, 1, 23-31.

Ohlund, C., Lindstrom, I., Eek, C., Areskoug, B., Nachemson, A. (1996) The causality field (extrinsic and intrinsic factors) in industrial subacute low back pain patients. *Scandinavian Journal of Medicine and Science in Sports*, 6, 98-111.

Ombregt, L., Bisschop, P., Ter Veer, H., Van De Velde, T. (1995) *A System of Orthopaedic Medicine*. London, WB Saunders Ltd.

Ostergren, P., Hanson, B., Balogh, I., Ektor-Andersen, J., Isacsson, A., Orbaek, P., Winkel, J., Isacsson, S. (2005) Incidence of shoulder and neck pain in a working population: effect modification between mechanical and psychosocial exposures at work? Results from a one year follow up of the Malmo shoulder and neck study cohort. *Journal of Epidemiology and Community Health*, 59, 721-728.

O'Sullivan, P. (2005) Diagnosis and classification of chronic low back pain disorders: maladaptive movement and motor control impairments as underlying mechanism. *Manual Therapy*, 10, 242-255.

O'Sullivan, P., Dankaerts, W., Burnett, A., Farrell, G., Jefford, E., Naylor, C., O'Sullivan, K. (2006) Effect of different upright sitting postures on spinal-pelvic curvature and trunk muscle activation in a pain-free population. *Spine*, 31, E707-712.

Ozguler, A., Gueguen, A., Leclerc, A., Landre, M-F., Piciotti, M., Le Gall, S., Morel-Fatio, M., Boureau, F. (2002) Using the Dallas Pain Questionnaire to classify individuals with low back pain in a working population. *Spine*, 27, 1783-1789.

Parascandola, J. (1993) A behavioral approach to determining the prognosis in patients with lumbosacral radiculopathy. *Rehabilitation Nursing*, 18, 314-317.

Parent, E. (2008) Preliminary development of a clinical prediction rule to identify patients with low back pain responding to extension exercises. *International Journal of Mechanical Diagnosis and Therapy*, 3, 33-34.

Parent, E., Videman, T., Battie, M. (2006) The effect of lumbar flexion and extension on disc contour abnormality measured quantitatively on magnetic resonance imaging. *Spine*, 31, 2836-2842.

Parente, F. Anderson-Parente, J. (1987) Delphi Inquiry Systems. In: Wright G., Ayton, P. (eds) *Judgemental Forecasting.* Chichester, John Wiley.

Patijn, J. (2004) Reproducibility and validity studies of diagnostic procedures in Manual/Musculoskeletal Medicine. Protocol formats. FIMM (International Federation for Manual/Musculoskeletal Medicine) Scientific Committee. Available at: http://www.dsmm.org/Videnskab/Validityprotocol.doc [November 2007].

Pedersen, P. (1981) Prognostic indicators in low back pain. *Journal of the Royal College of General Practitioners*, 31, 209-216.

Pengel, L, Herbert, R., Maher, C., Refshauge, K. (2003) Acute low back pain: systematic review of its prognosis. *British Medical Journal*, 327, 323.

Pennie, B., Agambar, L. (1991) Patterns of injury and recovery in whiplash. *Injury*, 22, 57-59.

Petersen, I. (1998) Using the kappa coefficient as a measure of reliability or reproducibility. *Chest*, 114, 946-947.

Petersen, T. (2003) Non specific low back pain. Classification and treatment. PhD Thesis, Lund University. Available at: http://www.ffy.dk/graphics/PDF-filer/Afhandlinger/phd_afhandling/Non-specific_low_back_pain_phdafhandling_Tom_Petersen_2003.pdf [30th September 2008].

Petersen, T., Laslett, M., Manniche, C., Ekdahl, C., Jacobsen, S. (2003) Diagnostic classification on non-specific low back pain. A new system integrating patho-anatomic and clinical categories. *Physiotherapy Theory and Practice*, 19, 213-237.

Petersen, T., Olsen, S., Laslett, M., Thorsen, H., Manniche, C., Ekdahl, C., Jacobsen, S. (2004) Inter-tester reliability of a new diagnostic classification system for patients with non-specific low back pain. *Australian Journal of Physiotherapy*, 50, 85-91.

Petersen, T., Thorsen, H., Manniche, C. (1999) Classification of non-specific low back pain: a review of the literature on classification systems relevant to physiotherapy. *Physical Therapy Reviews*, 4, 265-281.

Petticrew, M., Roberts, H. (2006) *Systematic reviews in the social sciences.* Malden, Blackwell Publishing.

Petty, N. (2006) *Neuromusculoskeletal examination and assessment: a handbook for therapists*. Edinburgh, Elsevier/ Churchill Livingstone.

Philadelphia Panel (2001) Philadelphia Panel evidence-based clinical practice guidelines on selected rehabilitation interventions for neck pain *Physical Therapy*, 81, 1701-1717.

Phillips, D., Twomey, L. (1996) A comparison of manual diagnosis with a diagnosis established by a uni-level lumbar spinal block procedure. *Manual Therapy*, 1, 82-87.

Pill, J. (1971) The Delphi method: substance, content, a critique and an annotated bibliography. Socioeconomic Planning Science; 5:57-71. Cited In: Pope, C., Mays, N. (Ed.) *Qualitative Research in Health Care.* London, BMJ.

Pincus, T., Santos, R., Breen, A., Burton A., Underwood M., Multinational Musculoskeletal Inception Cohort Study Collaboration. (2008) A review and proposal for a core set of factors for prospective cohorts in low back pain: a consensus statement. *Arthritis and Rheumatism*, 59, 14-24.

Pinto, D., Cleland, J., Palmer, J., Eberhart, S. (2007) Management of low back pain: a case series illustrating the pragmatic combination of treatment and mechanism – based classification systems. *Journal of Manual and Manipulative Therapy,* 15, 111-122.

Piva, S., Erhard, R, Childs, J., Browder, D. (2006) Inter-tester reliability of passive intervertebral and active movements of the cervical spine. *Manual Therapy*, 11, 321-330.

Poitras, S., Blais, R., Swaine, B., Rossignol, M. (2005) Management of work-related low back pain: a population –based survey of physical therapists. *Physical Therapy,* 85, 1168-1181.

Polatin, P., Cox, B., Gatchel, R., Mayer, T. (1997) A prospective study of Waddell signs in patients with chronic low back pain. When they may not be predictive. *Spine,* 22, 1618-1621.

Polatin, P., Gatchel, R., Barnes, D., Mayer, H., Arens, C., Mayer, T. (1989) A psychosociomedical prediction model of response to treatment by chronically disabled workers with low-back pain. *Spine,* 1, 956-961.

Pool, J., Hoving, J., De Vet, H., Van Mameren, H., Bouter, L. (2004) The interexaminer reproducibility of physical examination of the cervical spine. *Journal of Manipulative and Physiological Therapeutics,* 27, 84-90.

Pope, C., Mays, N. (2000) *Qualitative research in health care*. London, BMJ.

Posner, K., Sampson, P., Caplan, R., Ward, R., Cheney, F. (1990) Measuring interrater reliability among multiple raters: an example of methods for nominal data. *Statistics in Medicine,* 9, 1103-1115.

Potter, N., Rothstein, J. (1985) Intertester reliability for selected clinical tests of the sacroiliac joint. *Physical Therapy,* 65**,** 1671-1675.

Potter, R., Jones, J., Boardman, A. (2000) A prospective study of primary care patients with musculoskeletal pain: the identification of predictive factors for chronicity. *British Journal of General Practice*, 50, 225-227.

Powell, C. (2003) The Delphi technique: myths and realities. *Journal of Advanced Nursing*, 41, 376-382.

Powers, C., Beneck, G., Kulig, K., Landel, R., Fredericson, M. (2008) Effects of a single session of posterior – to – anterior spinal mobilization and press-up exercise on pain response and lumbar spine extension in people with nonspecific low back pain. *Physical Therapy*, 88, 485-493.

Procter, S., Hunt, M. (1994) Using the Delphi survey technique to develop a professional definition of nursing for analysing nursing workload. *Journal of Advanced Nursing*, 19, 1003-1014.

Radanov, B., Di Stefano, G., Schnidrig, A., Ballinari, P. (1991) Role of psychosocial stress in recovery from common whiplash. *Lancet*, 338, 712-715.

Radanov, B., Sturzenegger, M. (1996a) Predicting recovery from common whiplash. *European Neurology*, 36, 48-51.

Radanov, B., Sturzenegger, M. (1996b) The effect of accident mechanisms and initial findings on the long-term outcome of whiplash injury. *Journal of Musculoskeletal Pain*, 4, 47-59.

Radanov, B., Sturzzenegger, M., Di Stefano, G. (1995) Long-term outcome after whiplash injury: a two-year follow-up considering features on injury mechanism and somatic, radiologic and psychosocial findings. *Medicine*, 74, 281-297.

Radanov, B., Sturzzenegger, M., Di Stefano, G., Schnidrig, A. (1994) Relationship between early somatic, radiological, cognitive and psychosocial findings and outcome during a one-year follow-up in 117 patients suffering from common whiplash. *British Journal of Rheumatology*, 33, 442-448.

Rainville, J., Carlson, N., Polatin, P., Gatchel, R., Indahl, A. (2000) Exploration of physicians' recommendations for activities in chronic low back pain. *Spine*, 25, 2210-2220.

Razmjou, H., Kramer, J., Yamada, R. (2000) Intertester reliability of the McKenzie evaluation in assessing patients with mechanical low back pain. *Journal of Orthopaedic and Sports Physical Therapy*, 30, 368-383.

Reid, N. (1988) The Delphi technique: its contribution to the evaluation of professional practice. In Ellis, R. (ed) *Professional Competence and Quality Assurance in the Caring Professions.* London, Chapman and Hall.

Rennie, D. (1981) Consensus statements. *New England Journal of Medicine*, 304, 665-666.

Rezaian, S., Spector, I., Collins, A. (1993) A new clinical classification of low back pain: a practical guide for diagnosing etiology by patient history. *Journal of Neurological and Orthopaedic Medicine and Surgery*, 14, 205-209.

Richter, M., Ferrari, R., Otte, D., Kuensebeck, H., Blauth, M., Krettek, C. (2004) Correlation of clinical findings, collision parameters, and psychological factors in the outcome of whiplash associated disorders. *Journal of Neurology and Neurosurgical Psychiatry*, 75, 758-764.

Riddle, D. (1998) Classification and low back pain: a review of the literature and critical analysis of selected systems. *Physical Therapy*, 78, 708-737.

Riddle, D., Rothstein J. (1993) Intertester reliability of McKenzie's classifications of the syndrome types present in patients with low back pain. *Spine*, 18, 1333-1344.

Rigby, A. (2000) Statistical methods in epidemiology. V. Towards an understanding of the kappa coefficient. *Disability and Rehabilitation*, 22, 339-344.

Riley, R., Abrams, K., Sutton, A., Lambert, P., Jones, D., Heney, D., Burchill, S. (2003) Reporting of prognostic markers: current problems and development of guidelines for evidence-based practice in the future. *British Journal of Cancer*, 88, 1191-1198.

Roberts, C. (2008) Modelling patterns of agreement for nominal scales. *Statistics in Medicine*, 27, 810-830.

Roberts, C., Mcnamee, R. (1998) A matrix of kappa –type coefficients to assess the reliability of nominal scales. *Statistics in Medicine*, 17, 471-488.

Robinson, H., Brox, J., Robinson, R., Bjelland, E., Solem, S., Telje, T. (2007) The reliability of selected motion- and pain provocation tests for the sacroiliac joint. *Manual Therapy*, 12, 72-79.

Roland, M., Morris, R. (1983) A study on the natural history of low-back pain. Part II: development of guidelines for trials of treatment in primary care. *Spine*, 8, 145-150.

Rothstein, J. (2004) The difference between knowing and applying. *Physical Therapy*, 84, 310-311.

Rothstein, J., Campbell, S., Ekhternach, J., Jette, A., Knecht, H., Rose, S. (1991) Task Force on Standards for Measurement in Physical Therapy. Standards for tests and measurements in physical therapy practice. *Physical Therapy*, 71, 589-622.

Rowe, G. Wright, G., Bolger, F. (1991) Delphi: a reevaluation of research and theory. *Technological Forecasting and Social Change*, 39, 235-251.

Royston, P., Altman, D., Sauerbrei, W. (2006) Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in Medicine*, 25, 127-141.

Rubinstein, S., Van Tulder, M. (2008) A best-evidence review of diagnostic procedures for neck and low-back pain. *Best Practice and Research: Clinical Rheumatology*, 22, 471-482.

Sackett, D., Haynes, R., Tugwell, P. (1991) *Clinical Epidemiology: a basic science for clinical medicine.* Boston, Little, Brown and Company.

Sahrmann, S. (2002) *Diagnosis and Treatment of Movement Impairment Syndromes.* Saint Louis, MO. Mosby Publishers.

Sampson, M., Mcgowan, J., Tetzlaff, J., Cogo, E., Moher, D. (2008) No consensus exists on search reporting methods for systematic reviews. *Journal of Clinical Epidemiology,* 61, 748-754.

Sandstrom, J. (1986) Clinical and social factors in rehabilitation of patients with chronic low back pain. *Scandinavian Journal of Rehabilitation Medicine*, 18, 35-43.

Schectman, J., Schroth, W., Verme, D., Voss, J. (2003) Randomized controlled trial of education and feedback for implementation of guidelines for acute low back pain. *Journal of General Internal Medicine*, 18, 773-780.

Scheider, M., Erhard, R., Brach, J., Tellin, W., Imbarlina, F., Delitto, A. (2008) Spinal palpation for lumbar segmental mobility and pain provocation: an interexaminer reliability study. *Journal of Manipulative and Physiological Therapeutics*, 31, 465-473.

Schellingerhout, J., Verhagen, A., Heymans, M., Pool, J., Vonk, F., Koes, B., De Vet, H. (2008) Which subgroups of patients with non-specific neck pain are more likely to benefit from spinal manipulation therapy, physiotherapy, or usual care? *Pain*, 139, 670-680 .

Schenk, R., Jazefczyk, C., Kopf, A. (2003) A randomized trial comparing interventions in patients with lumbar posterior derangement. *Journal of Manual and Manipulative Therapy*, 11, 95-102.

Schibye, B., Skov, T., Ekner, D., Christiansen, J., Sjogaard, G. (1995) Musculoskeletal symptoms among sewing machine operators. *Scandinavian Journal of Work, Environment and Health*, 21, 427-434.

Schiottz-Christensen, B., Nielsen, G. (1999) Long term prognosis of acute low back pain in patients seen in general practice: a 1-year prospective follow-up study. *Family Practice*, 16, 223-232.

Schmidt, I., Rechter, L., Hansen, V. K., Andreasen, J., Overvad, K. (2008) Prognosis of subacute low back pain patients according to pain response. *European Spine Journal*, 17, 57-63.

Scholten-Peeters, G., Verhagen, A., Bekkering, G., Van Der Windt, D., Barnsley, L., Oostendorp, R., Hendriks, E. (2003) Prognostic factors of whiplash-associated disorders: a systematic review of prospective cohort studies. *Pain*, 104**,** 303-322.

Schultz, I., Crook, J., Berkowitz, J., Meloche, G., Milner, R., Zuberbier, O., Meloche, W. (2002) Biopsychosocial multivariate predictive model of occupational low back disability. *Spine*, 27, 2720-2725.

Schuster, C. (2004) A note on the interpretation of weighted kappa and its relations to other rater agreement statistics for metric scales. *Educational and Psychological Measurement*, 64, 243.

Scientific Advisory Committee Of The Medical Outcomes Trust (2002) Assessing health status and quality –of –life instruments: attributes and review criteria. *Quality of Life Research*, 11, 193-205.

Scott, W. (1955) Reliability of content analysis: the case of nominal scale coding. *Public Opinion Quarterly,* XIX, 321-325.

Seale, C. (1999) *The quality of qualitative research.* London, Sage Publications.

Seferlis, T., Nemeth, G., Carlsson, A. (2000) Prediction of functional disability, recurrences, and chronicity after 1 year in 180 patients who required sick leave for acute low-back pain. *Journal of Spinal Disorders*, 13, 470-477.

Seffinger, M., Najm, W., Mishra, S., Adams, A., Dickerson, V., Murphy, L., Reinsch, S. (2004) Reliability of spinal palpation for diagnosis of back and neck pain: a systematic review of the literature. *Spine*, 29, E413-425.

Seymour, R., Walsh, T., Blankenberg, C., Pickens, A., Rush, H. (2002) Reliability of detecting a relevant lateral shift in patients with lumbar derangement: a pilot study. *Journal of Manual and Manipulative Therapy*, 10, 129-135.

Shi, J., Copas, J. (2004) Meta-analysis for trend estimation. *Statistics in Medicine,* 23, 3-19.

Shrout, P. (1998) Measurement reliability and agreement in psychiatry. *Statistical Methods in Medical Research*, 7, 301-317.

Shrout, P., Fleiss, J. (1979) Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin,* 86, 420-428.

Sieben, J., Vlaeyen, J., Portegijs, P., Warmenhoven, F., Sint, A., Dautzenberg, N., Romeijnders, A., Arntz, A., Knottnerus, A. (2009) General practitioners' treatment orientations towards low back pain: influence on treatment behaviour and patient outcome. *European Journal of Pain*, 13, 412-418.

Siivola, S., Levoska, S., Latvala, K., Hoskio, E., Vanharanta, H., Keinanen-Kiukaanniemi, S.(2004) Predictive factors for neck and shoulder pain: a longitudinal study in young adults. *Spine*, 29, 1662-1669.

Sikorski, J. (1985) A rationalized approach to physiotherapy for low back pain. *Spine*, 10, 571-579.

Sim, J., Wright, C. (2005) The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical Therapy*, 85, 257-268.

Sim, J., Wright, C. (2000) *Research in Health Care: Concepts, Designs and Methods*, Cheltenham, Nelson Thornes.

Simon, P. (2006) Including omission mistakes in the calculation of Cohen's kappa and an analysis of the coefficient's paradox features. *Educational and Psychological Measurement*, 66, 765-777.

Simon, R., Altman, D. (1994) Statistical aspects of prognostic factor studies in oncology. *British Journal of Cancer*, 69, 979-985.

Skargren, E., Oberg, B. (1998) Predictive factors for 1-year outcome of low back and neck pain in patients treated in primary care: comparison between the treatment strategies chiropractic and physiotherapy. *Pain*, 77, 201-207.

Skytte, L., May, S., Petersen, P. (2005) Centralization: its prognostic value in patients with referred symptoms and sciatica. *Spine*, 30, E293-299.

Slavin, R. (1995) Best evidence synthesis: an intelligent alternative to meta-analysis. *Journal of Clinical Epidemiology*, 48, 9-18.

Smythe, H. (1994) The C6-7 syndrome--clinical features and treatment response. *Journal of Rheumatology*, 21, 1520-1526.

Snodgrass, S., Rivett, D., Robertson, V. (2007) Manual forces applied during cervical mobilization. *Journal of Manipulative and Physiological Therapeutics*, 30, 17-25.

Snodgrass, S., Rivett, D., Robertson, V. (2006) Manual forces applied during posterior –to – anterior spinal mobilization: a review of the evidence. *Journal of Manipulative and Physiological Therapeutics*, 29, 316-329.

Soanes, C., Stevenson, A. (2003) *The New Oxford English Dictionary*. 2nd edition. Oxford, Oxford English Press.

Sobel, J., Sollenberger, P., Robinson, R., Polatin, P., Gatchel, R. (2000) Cervical nonorganic signs: a new clinical tool to assess abnormal illness behavior in neck pain patients: a pilot study. *Archives of Physical Medicine and Rehabilitation*, 81, 170-175.

Soderlund, A., Linberg, P. (1999) Long-term functional and psychological problems in whiplash associated disorders. International *Journal of Rehabilitation Research*, 22, 77-84.

Soderlund, A., Olerud, C., Lindberg, P. (2000) Acute whiplash-associated disorders (WAD): the effects of early mobilization and prognostic factors in long-term symptomatology. *Clinical Rehabilitation*, 14, 457-467.

Spitzer, W., Leblanc, F., Dupuis, M., Abenhaim, L., Bloch, R., Bombardiier, C. (1987) Scientific approach to the activity assessment and management of activity related spinal disorders. *Spine*, 12, S1-55

Spitzer, W., Skovron, M., Salmi, L., Cassidy, J., Duranceau, J., Suissa, S., Zeiss, E. (1995) Scientific monograph of the Quebec Task Force on Whiplash-Associated Disorders: redefining "whiplash" and its management. *Spine*, 20, S1–73.

Spratt, K. (2002) Statistical relevance. In: FARDON, D. *Orthopaedic Knowledge Update Spine*. Rosemont, IL, American Academy of Orthopaedic Surgeons, 497-505.

Spratt, K., Lehmann, T., Weinstein, J., Sayre, H. (1990) A new approach to the low-back physical examination. Behavioral assessment of mechanical signs. *Spine,* 15, 96-102.

SPSS 14.0 SPSS Inc. Chicago, Illinois.

SPSS 15.0 SPSS Inc. Chicago, Illinois.

Staal, J., Hlobil, H., Van Tulder, M., Waddell, G., Burton, A., Koes, B., Van Mechelen, W. (2003) Occupational health guidelines for the management of low back pain: an international comparison. *Occupational and Environmental Medicine*, 60, 618-626.

Stanton, W., Jull, G. (2003) Cervicogenic headache: locus of control and success of treatment. *Headache*, 43, 956-961.

Sterling, M. (2004) A proposed new classification system for whiplash associated disorders – implications for assessment and management. *Manual Therapy*, 9, 60-70.

Sterner, Y., Toolanen, G., Gerdle, B., Hildingsson, C. (2003) The incidence of whiplash trauma and the effects of different factors on recovery. *Journal of Spinal Disorders Technology*, 16, 195-199.

Stevens, B., Mckenzie, R. (1988) Mechanical diagnosis and self-treatment of the cervical spine. In: Grant, E. (ed) Physical therapy of the cervical and thoracic spine. Churchill Livingstone, Edinburgh, 271-289.

Stiefel, F., Dejonge, P., Huyse, F., Guex, P., Slaets, J., Lyons, J., Spagnoli, J., Vannotti, M. (1999a) "INTERMED": a method to assess health service needs. II. Results on its validity and clinical use. *General Hospital Psychiatry*, 21, 49-56.

Stiefel, F., Dejonge, P., Huyse, F., Slaets, J., Guex, P., Lyons, J., Vannotti, M., Fritsch, C., Moeri, R., Leyvraz, P., So, A., Spagnoli, J. (1999b) "INTERMED" – an assessment and classification system for case complexity: Results in patients with low back pain. *Spine*, 24, 378-384.

Stochkendahl, M., Christensen, H., Hartvigsen, J., Vach, W., Haas, M., Hestbaek, L., Adams, A., Bronfort, G. (2006) Manual examination of the spine: a systematic

critical literature review of reproducibility. *Journal of Manipulative and Physiological Therapeutics,* 29, 475-485.

Stokes, F. (1997) Using the Delphi technique in planning of a research project on the occupational therapists' role in enabling people to make vocational choices following illness/ injury. *British Journal of Occupational Therapy,* 60, 263-267.

Straus, S., Richardson, W., Glasziou, P., Haynes, R. (2005) *Evidence-Based Medicine: How to practice and teach EBM.* 3rd edition. New York, Elsevier/Churchill Livingstone, 101-114.

Streiner, D., Norman, G. (2003) *Health Measurement Scales: a practical guide to their development and use*. Oxford, Oxford University Press.

Strender, L., Lundin, M., Nell, K. (1997a) Interexaminer reliability in physical examination of the neck. *Journal of Manipulative and Physiological Therapeutics,* 20, 516-520.

Strender, L., Sjoblom, A., Sundell, K., Ludwig, R., Taube, A. (1997b) Interexaminer reliability in physical examination of patients with low back pain. *Spine,* 22, 814-820.

Strong, J., Ashton, R., Stewart, A. (1994) Chronic low back pain: toward an integrated psychosocial assessment model. *Journal of Consulting and Clinical Psychology,* 62, 1058-1063.

Strong, J., Large, R., Ashton, R., Stewart, A. (1995) A New Zealand replication of the IPAM clustering model for low back patients. *Clinical Journal of Pain,* 11, 296-306.

Stroup, D., Berlin, J., Morton, S., Olkin, I., Williamson, G., Rennie, D., Moher, D., Becker, B., Sipe, T., Thacker, S. (2000) Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis Of Observational Studies in Epidemiology (MOOSE) group. *Journal of the American Medical Association,* 283, 2008-2012.

Sufka, A., Hauger, B., Trenary, M., Bishop, B., Hagen, A., Lozon, R., Martens, B. (1998) Centralization of low back pain and perceived functional outcome. *Journal of Orthopaedic and Sports Physical Therapy,* 27, 205-212.

Suissa, S. (2003) Risk factors of poor prognosis after whiplash injury. *Pain Research and Management,* 8, 69-75.

Suissa, S., Harder, S., Veilleux, M. (2001) The relation between initial symptoms and signs and the prognosis of whiplash. *European Spine Journal,* 10, 44-49.

Sumsion, T. (1998) The Delphi technique: an adaptive research tool. *British Journal of Occupational Therapy,* 61, 153-156.

Sweetman, B., Heinrich, I., Anderson, J. (1992) Clinical tests and patterns of low back pain. *Journal of Orthopaedic Rheumatology*, 5, 209-222.

Swinkels, R., Bouter, L., Oostendorp, R., Van Den Ende, C. (2005) Impairment measures in rheumatic disorders for rehabilitation medicine and allied health care: a systematic review. *Rheumatology International*, 25, 501-512.

Tanner, M., Young, M. (1985) Modelling agreement among raters. *Journal of the American Statistical Association,* 80, 175-180.

Taylor, W., Simpson, R., Gow, D., Mcnaughton, H. (2001) Rehabilitation that works-vocational outcomes following rehabilitation for occupational musculoskeletal pain. *New Zealand Medical Journal*, 114, 185-187.

Terrier, B. (2004) FIMM Academy for Manual Medicine: upgrade or sidegrade? *Manuelle Medizin*, 42, 7-8.

Thomas, E., Silman, A., Croft, P., Papageorgiou, A., Jayson, M., Macfarlane, G. (1999) Predicting who develops chronic low back pain in primary care: a prospective study. *British Medical Journal*, 318, 1662-1667.

Thompson, W., Walter, S. (1988) A reappraisal of the kappa coefficient. *Journal of Clinical Epidemiology*, 41, 949-958.

Tierney, J., Stewart, L., Ghersi, D., Burdett, S., Sydes, M. (2007) Practical methods for incorporating summary time-to-event data into meta-analysis. *Trials*, 8, 16.

Tong, H., Geisser, M., Ignaczak, A. (2006) Ability of early response to predict discharge outcomes with physical therapy for chronic low back pain. *Pain Practice*, 6, 166-170.

Travell, J., Simons, D. (1983) *Myofascial Pain and Dysfunction: The Trigger Point Manual.* Baltimore, Williams and Wilkins.

Troup, J. (1981a) Straight leg Raising (SLR) and the qualifying tests for increased root tension: their predictive value after back and sciatic pain. *Spine*, 6, 526-527.

Troup, J., Martin, J., Lloyd, D. (1981b) Back pain in industry: a prospective survey. *Spine*, 6, 61-69.

Tseng, Y., Wang, W., Chen, W., Hou, T., Chen, T., Lieu, F. (2006) Predictors for the immediate responders to cervical manipulation in patients with neck pain. *Manual Therapy*, 11, 306-315.

Tubach, F., Leclerc, A., Landre, M., Pietri-Taleb, F. (2002) Risk factors for sick leave due to low back pain: a prospective study. *Journal of Occupational and Environmental Medicine*, 44, 451-458.

Tuchin, P. (2008) Treatment provider preference of Australian office workers for common neurologic and musculoskeletal conditions. *Journal of Chiropractic Medicine*, 7, 79-85.

Tuckett, A. (2005) Qualitative research sampling: the very real complexities. *Nurse Researcher*, 12, 47-61.

Turner, P., Whitfield, A., Brewster, S., Halligan, M., Kennedy, J. (1996) The assessment of pain: an audit of physiotherapy practice. *Australian Journal of Physiotherapy*, 42, 55-62.

Tuttle, N. (2005) Do changes within a manual therapy treatment session predict between-session changes for patients with cervical spine pain? *Australian Journal of Physiotherapy*, 51, 43-48.

Tuttle, N., Laasko, L., Barrett, R. (2006) Change in impairments in the first two treatments predicts outcome in impairments, but not in activity limitations, in subacute neck pain: an observational study. *Australian Journal of Physiotherapy*, 52, 281-285.

Underwood, M., Morton, V., Farrin, A., On Behalf Of The UK BEAM Trial Team (2007) Do baseline characteristics predict response to treatment for low back pain? Secondary analysis of the UK BEAM dataset [ISRCTN32683578]. *Rheumatology*, 46, 1297-1302.

Urwin, M., Symmons, D., Allison, T., Brammah, T., Busby, H., Roxby, M., Simmons, A., Williams, G. (1998) Estimating the burden of musculoskeletal disorders in the community: the comparative prevalence of symptoms at different anatomical sites, and the relation to social deprivation *Annals of Rheumatic Diseases*, 57, 649-655.

Valat, J., Goupille, P., Rozenberg, S., Urbinelli, R., Allaert, F., Spine Group Of The Societe Francaise (2000) Acute low back pain: predictive index of chronicity from a cohort of 2487 subjects. *Joint, Bone Spine*, 67, 456-461.

Valet, F., Guinot, C., Mary, J. (2007) Log-linear non-uniform association models for agreement between two ratings on an ordinal scale. *Statistics in Medicine*, 26, 647-662.

Van Den Hoogen, H., Koes, B., Deville, W., Van Eijk, J., Bouter, L. (1996) The inter-observer reproducibility of Lasegue's sign in patients with low back pain in general practice. *British Journal of General Practice*, 46, 727-730.

Van Der Weide, W., Verbeek, J., Salle, H., Van Dijk, F. (1999) Prognostic factors for chronic disability from acute low-back pain in occupational health care. *Scandinavian Journal of Work Environment and Health*, 25, 50-56.

Van Der Windt, D., Kuijpers, T., Jellema, P., Van Der Heijden, G., Bouter, L. (2007) Do psychological factors predict outcome in both low-back pain and shoulder pain? *Annals of Rheumatic Diseases*, 66, 313-319.

Van Der Wurff, P., Hagmeijer, R., Meyne, W. (2000) Clinical tests of the sacroiliac joint. A systematic methodological review. Part 1: reliability. *Manual Therapy*, 5, 30-36.

Van Deursen, L., Patjijn, J., Ockhuysen, A., Vortman, B. (1990) The value of some clinical tests of the sacroiliac joint. *Journal of Manual Medicine*, 5, 96-99.

Van Dillen, L., Maluf, K., Sahrmann, S. (2009) Further examination of modifying patient-preferred movement and alignment strategies in patients with low back pain during symptomatic tests. *Manual Therapy*, 14, 52-60.

Van Dillen, L., Sahrmann, S., Norton, B., Caldwell, C., Flemming, D., McDonnell, M., Woolsey, N. (1998) Reliability of physical examination items used for classification of patients with low back pain. *Physical Therapy*, 78, 979-988.

Van Genderen, F., De Bie, R., Helders, P., Van Meeteren, N. (2003) Reliability research: towards a more clinically relevant approach. *Physical Therapy Reviews*, 8, 169-176.

Van Trijffel, E., Anderegg, Q., Bossuyt, P., Lucas, C. (2005) Inter-examiner reliability of passive assessment of intervertebral motion in the cervical and lumbar spine: a systematic review. *Manual Therapy*, 10, 256-269.

Van Trijffel, E., Oostendorp, R., Lindeboom, R., Bossuyt, P., Lucas, C. (2009) Perceptions and use of passive intervertebral motion assessment of the spine: a survey among physiotherapists specializing in manual therapy. *Manual Therapy*, 14, 243-251.

Van Tulder, M., Becker, A., Bekkering, T., Breen, A., Del Real, M., Hutchinson, A., Koes, B., Laerum, E., Malmivaara, A. (2006) Chapter 3: European guidelines for the management of acute nonspecific low back pain in primary care. *European Spine Journal*, 15, S169-191.

Van Tulder, M., Furlan, A., Bombardier, C., Bouter, L., The Editorial Board Of The Cochrane Collaboration Back Review Group (2003) Updated method guidelines for systematic reviews in the Cochrane Collaboration Back Review Group. *Spine*, 28, 1290-1299.

Varamini, A., Jam, B. (2006a) A need for a shared paradigm in the physical therapy classification and treatment of mechanical low back pain: (Part I): Limitations of the patho-anatomical approach. Available at:

http://www.aptei.com/articles/pdf/IRSAT_1.pdf [21st August 2008].

Varamini, A., Jam, B. (2006b) A need for a shared paradigm in the physical therapy classification and treatment of mechanical low back pain: (Part II): Practice Guideline versus classification systems.

Available at: http://www.aptei.com/articles/pdf/IRSAT_2.pdf [21st August 2008].

Vendrig, A. (1999) Prognostic factors and treatment-related changes associated with return to work in the multimodal treatment of chronic back pain. *Journal of Behavioral Medicine*, 22, 217-232.

Vernon, H., Humphreys, B. (2008) Chronic mechanical neck pain in adults treated by manual therapy: a systematic review of change scores in randomized controlled trials of a single session. *Journal of Manual and Manipulative Therapy*, 16, E42-52.

Vernon, H., Humphreys, B., Hagino, C. (2006) The outcome of control groups in clinical trials of conservative treatments for chronic mechanical neck pain: a systematic review. *BMC Musculoskeletal Disorders*, 7, 58.

Viera, A., Garrett, J. (2005) Understanding interobserver agreement: the kappa statistic. *Family Medicine*, 37, 360-363.

Viikari-Juntura, E. (1987) Interexaminer reliability of observations in physical examinations of the neck. *Physical Therapy*, 67, 1526-1532.

Viikari-Juntura, E. (1983) Neck and upper limb disorders among slaughterhouse workers: an epidemiologic and clinical study. *Scandinavian Journal of Work, Environment and Health*, 9, 283-290.

Viikari-Juntura, E., Riihimaki, H., Tola, S., Videman, T., Mutanen, P. (1994) Neck trouble in machine operating, dynamic physical work and sedentary work: a prospective study on occupational and individual risk factors. *Journal of Clinical Epidemiology*, 47, 1411-1422.

Viikari-Juntura, E., Takala, E., Riihimaki, H., Malmivaara, A., Martikainen, R., Jappinen, P. (1998) Standardized physical examination protocol for low back disorders: feasibility of use and validity of symptoms and signs. *Journal of Clinical Epidemiology*, 51, 245-255.

Viikari-Juntura, E., Takala, E., Riihimaki, H., Martikainen, R., Jappinen, P. (2000) Predictive validity of symptoms and signs in the neck and shoulders. *Journal of Clinical Epidemiology*, 53, 800-808.

Vingard, E., Mortimer, M., Wiktorin, C., Pernold, R., Fredriksson, K., Nemeth, G., Alfredsson, L., Musculoskeletal Intervention Center - Norrtalje Study Group (2002) Seeking care for low back pain in the general population: a two-year follow-up study: results from the MUSIC-Norrtalje Study. *Spine*, 27, 2159-2165.

Von Korff, M., Deyo, R, Cherkin, D., Barlow, W. (1993) Back pain in primary care: Outcomes at 1 year. *Spine*, 18, 855-862.

Vos, C., Verhagen, A., Passchier, J., Koes, B. (2007) Management of acute neck pain in general practice: a prospective study. *British Journal of General Practice*, 57, 23-28.

Vroomen, P., De Krom, M., Knottnerus, J. (2000) Consistency of history taking and physical examination in patients with suspected lumbar nerve root involvement. *Spine*, 25**,** 91-96.

Waddell, G. (1980) Nonorganic physical signs in low back pain. *Spine*, 5, 117-125.

Waddell, G., Bircher, M., Finlayson, D., Main, C. (1984) Symptoms and signs: Physical disease or illness behaviour? *British Medical Journal*, 289, 739-741.

Waddell, G., Klaber-Moffett, J., Burton, K. (2004) *The Neck Book*. London, TSO.

Waddell, G., Main, C., Morris, E., Venner, R., Rae, P., Sharmy, S., Galloway, H. (1982) Normality and reliability in the clinical assessment of backache. *British Medical Journal*, 284, 1519-1523.

Waddell, G., Mcculloch, J., Kummel, E., Venner, R. (1980) Nonorganic physical signs in low-back pain. *Spine*, 5, 117-125.

Waddell, G., Somerville, D., Henderson, I., Newton, M. (1992) Objective clinical evaluation of physical impairment in chronic low back pain. *Spine*, 17, 617-628.

Wainner, R. (2003) Reliability of the clinical examination: how close is close enough? *Journal of Orthopaedic and Sports Physical Therapy*, 33, 488-491.

Walker, A., Selfe, J. (1996) The Delphi method: a useful tool for the allied health researcher. *British Journal of Therapy and Rehabilitation*, 3, 677-681.

Wang, W., Olson, S., Campbell, A., Hanten, W., Gleeson, P. (2003) Effectiveness of physical therapy for patients with neck pain: an individualized approach using a clinical decision-making algorithm. *American Journal of Physical Medicine and Rehabilitation*, 82, 203-218.

Waris, P., Kuorinka, I., Kurppa, K. (1979) Epidemiological screening of occupational neck and upper limb disorders. Methods and criteria. *Scandinavian Journal of Work, Environment and Health*, 6, S25-38.

Warren, R., Warren, M. (2001) Whiplash injury sustained in motor vehicle accidents: factors influencing time off work. *Journal of Orthopaedic Medicine*, 23, 50-54.

Watson P., Booker, C. (2004) Returning the chronically unemployed with low back pain to employment. *European Journal of Pain*, 8, 359-369.

Watson, R., Singer, C., Sproule, J. (1996) Checking from behind in ice hockey: a study of injury and penalty data in the Ontario University Athletic Association Hockey League. *Clinical Journal of Sport Medicine*, 6, 108-111.

Webler, T., Levine, D., Rakel, H., Renn, O. (1991) A novel approach to reducing uncertainty: the group Delphi. *Technological Forecasting and Social Change*, 39, 253-263.

Weiner, B. (2008) Spine update: the biopsychosocial model and spine care. *Spine*, 33, 219-223.

Werneke, M. (2005) What if centralisers receive other treatment? *In: The 9th McKenzie Institute International Conference*, Hersonissos, June 3-5, 2005.

Werneke, M., Harris, D., Lichter, R. (1993) Clinical effectiveness of behavioral signs for screening chronic low-back pain patients in a work-oriented physical rehabilitation program. *Spine*, 18, 2412-2418.

Werneke, M., Hart, D. (2004) Categorizing patients with occupational low back pain by use of the Quebec Task Force Classification system versus pain pattern classification procedures: discriminant and predictive validity. *Physical Therapy*, 84, 243-254.

Werneke, M., Hart, D. (2005) Centralization: association between repeated end-range pain responses and behavioral signs in patients with acute non-specific low back pain. *Journal of Rehabilitation Medicine*, 37, 286-290.

Werneke, M., Hart, D. (2001) Centralization phenomenon as a prognostic factor for chronic low back pain and disability *Spine*, 26, 758-765.

Werneke, M., Hart, D. (2003) Discriminant validity and relative precision for classifying patients with nonspecific neck and back pain by anatomic pain patterns. *Spine*, 28, 161-166.

Werneke, M., Hart, D., Cook, D. (1999) A descriptive study on the centralization phenomenon. A prospective analysis. *Spine*, 24, 676-683.

Werneke, M., Hart, D., Resnik, L., Stratford, P. (2008) Centralization: prevalence and effect on treatment outcomes using a standardized operational definition and measurement method. *Journal of Orthopaedic and Sports Physical Therapy*, 38, 116-125.

Werneke, M., May, S. (2005) Letter to the editor. *Journal of Orthopaedic and.Sports Physical.Therapy*, 35, 843-844.

Wessels, T., Ewert, T., Limm, H., Rackwitz, B., Stucki, G. (2007) Change factors explaining reductions of "interference" in a multidisciplinary and an exercise prevention program for low back pain. *Clinical Journal of Pain*, 23, 629-634.

Wessels, T., Van Tulder, M., Sigl, T., Ewert, T., Limm, H., Stucki, G. (2006) What predicts outcome in non-operative treatments of chronic low back pain? A systematic review. *European Spine Journal*, 15, 1633-1644.

Wetzel, F., Donelson. R. (2003) The role of repeated end-range/ pain response assessment in the management of symptomatic lumbar discs. *Spine Journal*, 3, 146-154.

White, C., Lefrost, S., Amsel, R., Jeans, M. (1997) Predictors of the development of chronic pain. *Research in Nursing and Health*, 20, 309-318.

White, L., Thomas, J. (2002) The rater reliability of assessments of symptom provocation in patients with low back pain. *Journal of Back and Musculoskeletal Rehabilitation*, 16, 83-90.

Whiting, P., Rutjes, A., Reitsma, J., Bossuyt, P., Kleijnen, J. (2003) The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Medical Research Methodology*, 3, 25.

Whiting, P., Rutjes, A., Reitsma, J., Glas, A., Bossuyt, P., Kleijnen, J. (2004) Sources of variation and bias in studies of diagnostic accuracy. A systematic review. *Annals of Internal Medicine*, 140, 189-202.

Wilczynski, N., Haynes, B. (2005) Optimal search strategies for detecting clinically sound prognostic studies in EMBASE: An analytic survey. *Journal of the American Medical Informatics Association*, 12, 481-485.

Wilczynski, N., Haynes, B., Hedges Team (2004) Developing optimal search strategies for detecting clinically sound prognostic studies in MEDLINE: an analytic survey. *BMC Medicine*, 2, 23.

Wilczynski, N., Haynes, B., Hedges Team (2002) Robustness of empirical search strategies for clinical content in MEDLINE. *Proceedings of AMIA Symposium*, 904-908.

Williams, G. (1976) Comparing the joint agreement of several raters with another rater. *Biometrics*, 32, 619-627.

Williams, M., Williamson, E., Gates, S., Lamb, S., Cooke, M. (2007) A systematic literature review of physical prognostic factors for the development of Late Whiplash Syndrome. *Spine*, 32, E764-780.

Williamson, E., Williams, M., Gates, S., Lamb, S. (2007) A systematic literature review of psychological factors and the development of late whiplash syndrome. *Pain*, 135, 20-30.

Wilson, A., Averis, A., Walsh, K. (2003) The influences on and experiences of becoming nurse entrepreneurs: a Delphi study. *International Journal of Nursing Practice*, 9, 236-245.

Wilson, L., Hall, H., Mcintosh, G., Melles, T. (1999) Intertester reliability of a low back pain classification system. *Spine*, 24, 248-254.

World Health Organisation (WHO) (2001) *International Classification of Functioning, Disabilities and Health Problems.* Geneva, World Health Organisation.

World Health Organisation (WHO) (2004) *ICD-10: International statistical classification of diseases and related health problems: tenth revision. Volume II*. 2nd edition. Geneva, World Health Organisation.

World Medical Association (WMA) (2008) World Medical Association Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects. Available at: http://www.wma.net/e/policy/pdf/17c.pdf [November 2008].

Yin, W., Bogduk, N. (2008) The nature of neck pain in a private pain clinic in the United States. *Pain Medicine*, 9, 196-203.

Ylinen, J., Takala, E., Kautiainen, H., Nykanen, M., Hakkinen, A., Pohjolainen, T., Karppi, S., Airaksinen, O. (2005) Effect of long-term neck muscle training on pressure pain threshold: a randomized controlled trial. *European Journal of Pain*, 9, 673-681.

Young, S., Aprill, C., Laslett, M. (2003) Correlation of clinical examination characteristics with three sources of chronic low back pain. *Spine Journal*, 3, 460-465.

Zou, G., Klar, N. (2005) A non-iterative confidence interval estimating procedure for the intraclass kappa statistic with multinomial outcomes. *Biometrical Journal*, 5, 682-690.

Zusman, M. (2005) Letter to the editor: Clinical anatomy serving manual therapy. *Manual Therapy*, 10, 232-233.

APPENDICES

Appendix 1.1 Examples of physical examination procedures and their reported purpose.

PHYSICAL EXAMINATION ITEM	REPORTED PURPOSE
Observation	Helps determine the shape of the spinal curvatures, muscle bulk, the state of the soft tissues, or the presence of asymmetry (Magee, 2002; Petty, 2006).
Movement testing	
Gross movement testing i.e. active, passive, resisted, repeated or combined movements	Determination of the patient's range of motion and movement loss (Petty, 2006). A directional preference i.e. a direction of movement that elicits a favourable response may also be identified (McKenzie and May, 2006).
Segmental or spring testing	Evaluation of the quality and amount of segmental movement and dysfunction, assessment of segmental pain through provocation to identify the dysfunctional segments (Hollerwöger, 2006)
Static or sustained posture testing	Usually performed to identify a directional preference by provoking obscure symptoms or decreasing persistent symptoms not mechanically conclusive through movement testing.
Palpation	Determines impairment at the segmental level including location of a painful segment, quality of force-displacement relationship, the quantity of segmental motion in order to decide the direction and grade of mobilisation to be used in treatment (Jull et al., 1994; Abbott et al., 2009). May involve motion or static palpation, osseous or soft tissue pain (provocation of tenderness) (Stochkendahl et al., 2006).
(Myofascial) Trigger point assessment	Through the presence of a set of criteria (i.e. taut band, local tenderness, patient pain recognition, pain referral, local twitch response, jump sign), this type of assessment directs treatment to specific trigger points (Myburgh et al., 2008).
Neurological examination	May usually involve examination of sensation, muscle power, reflexes and nerve tension tests (Petty, 2006). Identifies the dysfunction of neural tissue, in the central or peripheral nervous system (Petty, 2006).
Special testing	
Sacro-iliac joint pain provocation testing	Aim to identify the sacro-iliac joint as the source of the patient's reported symptoms (Hancock et al., 2007).
Non-organic signs	Their presence (i.e. tenderness, simulation, distraction, regional, overreaction to examination) may suggest the presence of illness behaviour (Sobel et al., 2000).
Instability testing	May indicate the presence of segmental instability and direct treatment (Hicks et al., 2005).

Appendix 2.1 Search strategies of online databases for reliability studies.

Ovid-MEDLINE
1. exp Spinal Diseases/
2. ((spin$ or low back or lumbar or neck or cervical or thoracic) adj2 (pain or disorder$)).ab.
3. (measure$ adj pain).ab
4. ((pain or symptom$) adj (provocation or change$ or response$ or pattern$ or behavi$)).ab.
5. ((provocation or change$ or pattern$ or behavi$) adj10 (pain or symptom$ or sign$)).ab.
6. exp "Reproducibility of Results"/
7. reproducib$.mp
8. reliab$.mp.
9. agreement.mp
10. exp Observer Variation/
11. (inter-examiner or interexaminer or interobserver or inter-observer or interrater or inter-rater).mp.
12. 1 or 1
13. or/3-5
14. or/6-11
15. 12 and 13 and 14

Ovid-EMBASE
1. exp Spinal Disease/
2. ((spin$ or low back or lumbar or neck or cervical or thoracic) adj2 (pain or disorder$)).ab.
3. (measur$ adj pain).ab
4. ((pain or symptom$) adj (provocation or change$ or response$ or pattern$ or behavi$)).ab.
5. ((provocation or change$ or pattern$ or behave$) adj10 (pain or symptom$ or sign$)).ab.
6. exp RELIABILITY/
7. reliab$.mp
8. exp REPRODUCIBILITY/
9. reproducib$.mp.
10. agreement.mp
11. exp Observer Variation/
12. (inter-examiner or interexaminer or interobserver or inter-observer or interrater or inter-rater).mp.
13. 1 or 2
14. or/3-5
15. or/6-12
15. 13 and 14 and 15

Ovid – CINAHL
1. exp Spinal Diseases/
2. ((spin$ or low back or lumbar or neck or cervical or thoracic) adj2 (pain or disorder$ or disease$)).ab.
3. (measur$ adj pain).ab
4. ((pain or symptom$) adj (provocation or change$ or response$ or pattern$ or behavi$)).ab.
5. ((provocation or change$ or pattern$ or behave$) adj10 (pain or symptom$ or sign$)).ab.
6. exp RELIABILITY/
7. reliab$.mp
8. exp "Reproducibility of Results"/
9. reproducib$.mp.
10. agreement.mp
11. exp Interrater Reliability/
12. (interexaminer or inter-examiner or interobserver or inter-observer or interrater or inter-rater).mp.
13. 1 or 2
14. or/3-5
15. or/6-12
15. 13 and 14 and 15

Ovid-AMED
1. exp Spinal Disease/
2. ((spin$ or low back or lumbar or neck or cervical or thoracic) adj2 (pain or disorder$ or disease$)).ab.
3. (measur$ adj pain).ab
4. ((pain or symptom$) adj (provocation or change$ or response$ or pattern$ or behavi$)).ab.
5. ((provocation or change$ or pattern$ or behave$) adj10 (pain or symptom$ or sign$)).ab.
6. exp Observer variation/ or exp "Consistency and reliability"/ or exp "Reproducibility of results"/
7. reliab$.mp
8. reproducib$.mp
9. agreement.mp
10. (inter-examiner or interexaminer or interobserver or inter-observer or interrater or inter-rater).mp.
11. 1 or 2
12. or/3-5
13. or 6-10
14. 11 and 12 and 13

Appendix 2.2 Excluded reliability papers.

REASON FOR EXCLUSION	REFERENCES
No inter-rater reliability analysis	Donelson et al., 1990; Sweetman et al., 1992; Donelson et al., 1997; Kool et al., 2002
Inclusion of asymptomatic participants	Lindsay et al., 1995; Waddell et al., 1982; Strender et al., 1997a; Lundberg and Gerdle, 1999; Comeaux et al., 2001; Horneij et al., 2002
Participants suffering from other conditions	Jull et al., 1994; Gerwin et al., 1997; Christensen et al., 2003
Age less than 18 years old	Viikari-Juntura, 1987; Bertilson et al., 2006
Inadequate data presentation	Nelson et al., 1979; McConnell et al., 1980; Leboeuf, 1989; Leboeuf et al., 1989; Leboeuf, 1991
Irrelevant tests	Johnston et al., 1983; Lankhorst et al., 1982; Cibulka et al., 1988; van Deursen et al., 1990; Phillips and Twomey, 1996; Werneke et al., 1999
Lack of or no separate analysis for eligible groups	Beal, 1984; Delitto et al., 1992; Binkley et al., 1995; Donahue et al., 1996; McPartland and Goodridge, 1997; Wilson et al., 1999; French et al., 2000; Fritz and George, 2000

Appendix 2.3 Quality assessment results of the included inter-rater reliability studies.

Reference	Study population				Study conditions						Study results				Total score (%)
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	
Cleland (2006a)	3	0	0	0	5	5	10	5	5	10	5	5	10	10	73
Dionne (2006)	3	0	7	0	5	5	10	5	5	10	5	5	10	10	84
Fritz (2006)	3	4	0	6	3	5	10	5	0	10	5	5	5	10	62
Piva (2006)	3	0	7	10	4	5	10	10	5	0	0	5	10	10	74
Haswell (2004)	3	4	7	3	5	5	10	5	5	10	5	5	10	10	87
Heiss (2004)	3	4	0	3	1	0	10	5	5	10	5	0	10	0	56
Petersen (2004)	4	4	7	10	5	5	10	5	5	10	5	5	10	10	95
Pool (2004)	3	4	7	3	5	5	10	10	5	0	0	5	10	10	77
Fritz (2003)	1	0	0	0	5	5	0	10	0	0	0	5	5	10	46
Hicks (2003)	3	4	7	6	5	5	0	10	5	10	0	5	10	10	85
Kilpikoski (2002)	3	4	7	3	5	0	10	10	0	0	0	5	10	7	69
Seymour (2002)	2	0	0	0	0	0	10	10	5	10	5	5	10	7	64
White (2002)	4	0	0	3	5	5	10	5	5	10	5	5	10	10	77
Fritz (2000)	2	0	0	0	5	5	0	2	5	10	5	5	5	10	54
Hsieh (2000)	3	0	0	3	0	0	0	5	5	10	5	5	10	7	53
Razmjou (2000)	4	4	7	6	5	5	10	10	5	0	0	5	10	7	83
Sobel (2000)	2	4	7	3	4	5	0	0	5	10	5	5	5	7	62
Vroomen (2000)	4	0	7	10	4	5	0	5	5	10	5	5	10	10	80
Van Dillen (1998)	3	4	0	10	0	5	0	5	5	10	5	5	5	7	64
Strender (1997)	4	4	0	6	5	5	10	10	5	10	0	5	10	10	84
Van den Hoogen (1996)	3	0	7	3	0	0	0	10	5	10	0	5	10	7	60
Hubka (1994)	3	4	0	3	4	5	0	5	0	0	0	5	10	7	46
Laslett (1994)	2	0	0	3	4	5	0	10	5	10	0	0	10	7	56
Maher (1994)	3	4	0	10	4	0	10	10	0	10	0	5	5	10	66
Njoo (1994)	4	4	0	6	4	0	0	5	5	10	5	5	10	10	68
Boline (1993)	1	0	0	3	0	0	0	5	0	0	5	5	10	7	36
Nice (1992)	3	4	0	3	5	5	0	10	5	10	0	5	10	10	70
Waddell (1992)	2	0	0	6	0	0	0	10	0	0	0	0	5	7	30
Keating (1990)	2	0	0	3	4	0	0	5	0	10	5	5	10	7	51
Kilby (1990)	1	0	0	3	5	5	10	10	0	0	0	5	10	7	56
Spratt (1990)	3	4	0	3	5	0	0	0	0	0	5	5	5	7	37
McCombe (1989)	2	0	7	10	4	0	0	0	0	10	0	5	0	10	48
Boline (1988)	2	0	0	0	0	0	0	5	5	0	0	0	10	7	29
Korbon (1987)	1	0	0	3	0	0	0	10	0	0	0	0	5	0	19
Potter (1985)	1	0	0	0	4	5	10	5	0	0	0	5	5	0	35
Waddell (1980)	1	0	7	3	1	0	0	10	0	0	0	0	5	0	27

1:Adequate description of study population; 2: Representative of clinical practice; 3: Subjects selected randomly or consecutively; 4: Number of subjects; 5: Procedure clearly described and reproducible; 6: Procedure executed in a uniform manner; 7: Adequate measures to reduce bias; 8: Level of examiners; 9: Consensus/ training procedure prior to testing with pilot study; 10: More than one pair of examiners tested; 11: Multiple testing between examiners; 12: Standardised measure of test outcome; 13: Frequencies of outcome and agreement reported; 14: Appropriate statistics.

Appendix 2.4 Table of results of clinically induced symptom responses in neck pain (Font: green, ≥ 0.7, red, ≥ 0.4, black < 0.4; highlighted cells represent changing results or evidence).

SYMPTOM RESPONSE ITEM	PRIMARY AUTHOR	STATISTIC*	VARIANCE**	LOE (R≥0.7+Q≥0.6)	LOE (R≥0.7+Q≥0.5)	LOE (R≥0.4+Q≥0.6)	LOE (R≥0.4+Q≥0.5)
Movement testing – Gross movement							
Pain with single Cx ROM	Cleland (2006a)	K_W = 0 to **0.81**	-0.07 to 1.0	**Strong**	**Strong**	**Strong**	**Strong**
Centralisation/ Peripheralisation with single Cx ROM	Cleland (2006a)	K_W = -0.05 to **1.0**	-0.15 to 1.0				
Status change with single Cx AROM	Piva (2006)	K= 0.25 to **0.87**	0.12 to 0.94				
Directional preference in Cx testing	Dionne (2006)	**K = 0.46,** P <0.05	0.43 to 0.49	**Moderate**	**Moderate**	**Moderate**	**Moderate**
Pain provocation with Cx ROM	Pool (2004)	ICC = 0.36 to **0.71**	NR	**Moderate**	**Moderate**	**Moderate**	**Moderate**
Movement testing – Segmental movement							
Pain provocation on segmental Cx mobility testing (spring testing)	Cleland (2006a)	K_W = 0.12 to **0.90**	-0.04 to 1.0	**Conflicting**	**Conflicting**	**Strong**	**Strong**
Pain provocation on segmental Cx mobility testing	Cleland (2006a)	K_W = -0.16 to 0.15	-0.05 to 0.54				
Pain provocation on segmental Cx mobility testing	Piva (2006)	K = 0.29 to **0.76**	0.15 to 0.87				
Pain provocation on segmental Cx mobility testing	Pool (2004)	ICC = 0.22 to **0.80**	NR	**Moderate**	**Moderate**	**Moderate**	**Moderate**
Non-organic signs							
Superficial tenderness (Cx, upper Tx)	Sobel (2000)	K = 0.33	NR	**Moderate**	**Moderate**	**Moderate**	**Moderate**
Non-anatomic tenderness(Cx, Tx, Lx, branchial regions)	Sobel (2000)	K = 0.26	NR	**Moderate**	**Moderate**	**Moderate**	**Moderate**
Simulation (sitting)/(standing)	Sobel (2000)	K =0.16 to **0.46**	NR	**Moderate**	**Moderate**	**Moderate**	**Moderate**
Palpation							
Cx tenderness	Hubka (1994)	**K = 0.68**, p<0.001	NR	**Limited**	**Limited**	**Limited**	**Limited**

*** Values in the stastitic column represent either single reliability values or the range of multiple results on a physical examination procedure.**

**** Reflects the 95% confidence interval where given.**

Characteristics of individual studies:

Primary author	Patients (no. [M / F], mean age [range or SD]) / Examiners (no, occupation, practicing experience, experience with procedure and / or study training)
Cleland (2006a)	22 (4 / 18), 41 (12.9) yrs / 4 PTs 3-23 yrs, +/+
Dionne (2006)	20 (7 / 13), 43 (21-75) yrs / 54 NR 0-35 yrs +/+
Piva (2006)	30 (12 / 18), 41 (12) yrs / 2 NR 2 and 10 yrs, +/+
Pool (2004)	32 (12 / 20), 45.5 (9.2) yrs / 2 PTs NR, +/+
Sobel (2000)	26 (20 / 6), 42.08 (9.98) yrs / 3 (1 MD, 1 PT, 1 OT), NR, -/+
Hubka (1994)	30 (11 / 19), 39.4 (14.6) yrs / 2 Chir, 1 and 5 yrs, +/-

Abbreviations:

AROM, Active Range of Motion; Chir, Chiropractor; Cx, Cervical; ICC, Intraclass Correlation Coefficient; K, Kappa Coefficient; Kw, Weighted Kappa Coefficient; LOE, Levels of Evidence; Lx, Lumbar; MD, Medical Doctor; M / F, Male / Female; NR, Not Reported; OT, Occupational Therapist; PTs, Physiotherapists; Q, Quality; R, Reliability; ROM, Range of Motion; SD, Standard Deviation; Tx, Thoracic.

Appendix 2.5 Table of results of clinically induced symptom responses in back pain (Font: green, ≥ 0.7, red, ≥ 0.4, black < 0.4; highlighted cells represent changing results or evidence).

SYMPTOM RESPONSE ITEM	PRIMARY AUTHOR	STATISTIC*	VARIANCE**	LOE (R≥0.7+Q≥0.6)	LOE (R≥0.7+Q≥0.5)	LOE (R≥0.4+Q≥0.6)	LOE (R≥0.4+Q≥0.5)
Movement testing – Gross movement							
Status change with single Lx ROM	Fritz (2006)	K_W = 0.51 to 0.55	0.28 to 0.81	**Moderate**	**Moderate**	**Moderate**	**Moderate**
Pain on Lx movement	Hicks (2003)	K = 0.61 to 0.69	0.44 to 0.84	**Conflicting**	**Conflicting**	**Strong**	**Strong**
Pain on Lx movement	McCombe (1989)	K = 0.10 to 0.56 (S1) K = 0.42 to 0.58 (S2)	NR				
Pain on Lx movement	Van Dillen (1998)	K = 0.87 to 1.00	NR				
Pain on Lx movement	Strender (1997b)	K = 0.51 to 0.76	NR				
Pain aggravation or reproduction on repeated Lx AROM	Haswell (2004)	K = 0.17 to 0.60	-0.08 to 0.79	**Moderate**	**Moderate**	**Moderate**	**Moderate**
Classification based on pain during Lx movement	White (2002)	K = 0.02 to 0.62	-0.11 to 0.87	**Moderate**	**Moderate**	**Moderate**	**Moderate**
Status change with repeated Lx ROM	Fritz (2006)	K_W = 0.15 to 0.46	-0.06 to 0.69	**Conflicting**	**Conflicting**	**Strong**	**Strong**
Centralisation with Lx repeated movement testing	Kilpikoski (2002)	K= 0.7, p < 0.002	NR				
Changes in pain location/intensity by repeated movements	Kilby (1990)	K=0.51	NR				
Directional preference	Kilpikoski (2002)	K = 0.9, p < 0.000	NR				
Relevance of lateral shift	Kilpikoski (2002)	K = 0.7, p < 0.000	NR	**Conflicting**	**Conflicting**	**Strong**	**Strong**
Relevance of lateral shift	Seymour (2002)	K = 0.56	NR				
Relevance of lateral shift	Razmjou (2000)	K = 0.85	NR				
Relevance of lateral component	Kilpikoski (2002)	K = 0.4, p < 0.021	NR	**Conflicting**	**Conflicting**	**Strong**	**Strong**
Relevance of lateral component	Razmjou (2000)	K = 0.95	NR				
Movement testing – Segmental movement							
Pain provocation on segmental Lx mobility testing	Hicks (2003)	K = 0.25 to 0.55	0.11 to 0.67	**Strong**	**Strong**	**Strong**	**Strong**
Pain provocation on segmental Lx mobility testing (including spring testing)	Strender (1997b)	K = 0.38 to 0.56	NR				
Pain provocation on segmental Lx mobility testing	Boline (1988)	K= 0.00 to 0.65	NR				
Pain provocation on segmental Lx mobility testing	Maher (1994)	ICC = 0.67 to 0.73	0.55 to 0.81	**Moderate**	**Moderate**	**Moderate**	**Moderate**
Static tests							
Status change with sustained Lx extension	Fritz (2006)	K_W = 0.28	0.10 to 0.47	**Moderate**	**Moderate**	**Moderate**	**Moderate**
Palpation							
Spinal tenderness	Fritz (2003)	K = 0.35	-0.33 to 1.00	**Conflicting**	**Conflicting**	**Moderate**	**Moderate**
Spinal tenderness	Waddell (1992)	K = 0.60, p < 0.001	NR				
Midline tenderness	McCombe (1989)	K = 0.38 (S1) K = 0.47 (S2)	NR				
Paraspinal tenderness	Strender (1997b)	K=0.27	NR	**Conflicting**	**Conflicting**	**Conflicting**	**Conflicting**

Paraspinal tenderness	Waddell (1992)	K=0.80, p < 0.001	NR				
Paraspinal tenderness	McCombe (1989)	K = 0.11 (S1) K = 0.38 (S2)	NR				
Soft tissue pain	Boline (1993)	K = 0.40 to 0.78	NR	**Conflicting**	**Conflicting**	**Moderate**	**Moderate**
Soft tissue pain	Keating (1990)	K= 0.13 to 0.59	NR				
Buttock tenderness	Waddell (1992)	K = 0.59, p < 0.01	NR				
Buttock tenderness	McCombe (1989)	K = 0.31 (S1) K = 0.34 (S2)	NR				
Osseous pain	Keating (1990)	K = 0.19 to 0.66	NR	**Conflicting**	**Conflicting**	**Moderate**	**Moderate**
Osseous pain	Boline (1993)	K = 0.48 to 0.90	NR				
Sacroiliac tenderness	McCombe (1989)	K = 0.41 (S1) K = 0.28 (S2)	NR				
Iliac crest tenderness	McCombe (1989)	K = 0.50 (S1) K = 0.36 (S2)	NR				
Trigger point assessment							
Taut band	Petersen (2004)	K= 0.44	0.25 to 0.64	**Conflicting**	**Conflicting**	**Moderate**	**Moderate**
Taut band	Hsieh (2000)	K = 0.78 (with expert) K =0.008 (among examiners)	NR				
Referred pain pattern	Hsieh (2000)	K = 0.27 (with expert) K = 0.33 (among examiners)	NR	**Conflicting**	**Conflicting**	**Conflicting**	**Conflicting**
Referred pain pattern	Njoo (1994)	K= 0.36 to 0.46	-0.04 to 0.76				
Localised tenderness	Njoo (1994)	K = 0.58 to 0.73	0.43 to 0.85	**Moderate**	**Moderate**	**Moderate**	**Moderate**
Trigger point assessment	Nice (1992)	K = 0.29 to 0.38	NR	**Moderate**	**Moderate**	**Moderate**	**Moderate**
Neural tests							
Pain on neural tension testing	Petersen (2004)	K= 0.59	0.39 to 0.79	**Moderate**	**Moderate**	**Moderate**	**Moderate**
Pain on SLR	Vroomen (2000)	K = 0.36 to 0.68	NR	**Conflicting**	**Conflicting**	**Conflicting**	**Conflicting**
Pain on SLR	Strender (1997b)	K = 0.83	NR				
Pain on SLR	McCombe (1989)	K = 0.36 to 0.66 (S1) K = 0.44 to 0.81 (S2)	NR NR				
Pain on SLR	Van den Hoogen (1996)	K = 0.33	NR				
SLR crossed	McCombe (1989)	K = 0.74 (S1) K = -0.02 (S2)	NR NR	**Moderate**	**Moderate**	**Moderate**	**Moderate**
SLR crossed	Vroomen (2000)	K = 0.49 if SLR positive: K = 0.70	NR				
SLR sciatic stretch test	McCombe (1989)	K = 0.37 (S1) K = 0.62 (S2)	NR NR	**Limited**	**Limited**	**Limited**	**Limited**
Bragard sign	Vroomen (2000)	K = 0.66	NR	**Moderate**	**Moderate**	**Moderate**	**Moderate**
Valleix pressure points	Vroomen (2000)	K = 0.14	NR	**Moderate**	**Moderate**	**Moderate**	**Moderate**
Pain on Bowstring testing	McCombe (1989)	K = 0.26 to 0.49 (S1) K = 0.11 to 0.20 (S2)	NR	**Limited**	**Limited**	**Limited**	**Limited**
Pain on passive knee flexion	Waddell (1992)	K = 0.57, p < 0.001	NR	**Conflicting**	**Conflicting**	**Moderate**	**Moderate**
Femoral nerve stretch	McCombe (1989)	K = 0.23 to 0.37 (S1) K = 0.50 to 0.77 (S2)	NR				

SIJ & other pain provocation tests							
Compression	Strender (1997b)	K = 0.26	NR	**Conflicting**	**Conflicting**	**Conflicting**	**Conflicting**
Compression	Laslett (1994)	K = 0.73, p < 0.001	NR				
Compression	McCombe (1989)	K = 0.09 (S1) K = 0.16 (S2)	NR				
Distraction	Laslett (1994)	K = 0.69, p < 0.001	NR	**Moderate**	**Moderate**	**Conflicting**	**Conflicting**
Distraction	McCombe (1989)	K = 0.11 (S1) K = 0.36 (S2)	NR				
Posterior shear or thigh thrust test	Laslett (1994)	K = 0.88, p < 0.001	NR	**Limited**	**Moderate**	**Limited**	**Moderate**
Pelvic torsion	Laslett (1994)	K = 0.72-0.75, p < 0.001	NR	**Limited**	**Moderate**	**Limited**	**Moderate**
Sacral thrust	Laslett (1994)	K = 0.52, p< 0.001	NR	**Limited**	**Moderate**	**Limited**	**Moderate**
Cranial shear test	Laslett (1994)	K= 0.61, p < 0.001	NR	**Limited**	**Moderate**	**Limited**	**Moderate**
Maitland sacroiliac test	McCombe (1989)	K = 0.26 (S1) K = 0.38 (S2)	NR	**Limited**	**Limited**	**Limited**	**Limited**
Pain on hip flexion	McCombe (1989)	K = 0.40 (S1) K = 0.42 (S2)	NR	**Conflicting**	**Conflicting**	**Moderate**	**Moderate**
Pain on hip flexion	Waddell (1992)	K = 0.71, p < 0.001	NR				
Pain on hip abduction	Waddell (1992)	K = 0.56, p < 0.001	NR	**Limited**	**Limited**	**Limited**	**Limited**
Pain on resisted external hip rotation	McCombe (1989)	K = 0.63(S1) K = 0.38 (S2)	NR	**Limited**	**Limited**	**Limited**	**Limited**
Pain on resisted hip flexion	Waddell (1992)	K= 0.72, p < 0.001	NR	**Limited**	**Limited**	**Limited**	**Limited**
Pain on vertebral percussion	Vroomen (2000)	K = 0.32	NR	**Moderate**	**Moderate**	**Moderate**	**Moderate**
Combination of strategies							
Status change with repeated Lx movements, sustained positions and manual overpressure, mobilisation/ manipulation	Petersen (2004)	K = 0.64	0.48 to 0.80	**Moderate**	**Moderate**	**Moderate**	**Moderate**
Status change with Lx movement or sustained postures	Fritz (2000)	K (total) = 0.79 K (PT) = 0.82 K (PT students)=0.76 K (> 6 yrs experience) = 0.87 K (< 6yrs experience) = 0.82	0.77 to 0.81 0.81 to 0.84 0.76 to 0.77 0.86 to 0.90 0.81 to 0.83	**Limited**	**Moderate**	**Limited**	**Moderate**
Non-organic signs							
Superficial tenderness	McCombe (1989)	K = 0.29 (S1) K = 0.17 (S2)	NR NR	**Limited**	**Limited**	**Limited**	**Limited**
Simulation	McCombe (1989)	K = 0.25 (S1) K = 0.48 (S2)	NR NR	**Limited**	**Limited**	**Limited**	**Limited**
Instability tests							
Prone instability test	Fritz (2006)	K =0.52	0.29 to 0.75	**Strong**	**Strong**	**Strong**	**Strong**
Prone instability test	Hicks (2003)	K= 0.87	0.80 to 0.94				
Lx Posterior shear test	Hicks (2003)	K = 0.35	0.20 to 0.51	**Moderate**	**Moderate**	**Moderate**	**Moderate**

*** Values in the stastitic column represent either single reliability values or the range of multiple results on a physical examination procedure.**

**** Reflects the 95% confidence interval where given.**

Studies not included in the levels of evidence analysis:

Primary author	Procedures and results
Heiss (2004)	**Movement testing - Gross movement**
	Centralisation with Lx ROM: 44% to 75% agreement
Spratt (1990)	**Palpation**
	Back tenderness: 100% adjusted agreement (same session), 54% (different sessions)
	Movement testing - Gross movement
	Pain on single Lx movement testing: 100% adjusted agreement (same session), 46 - 51% (different sessions)
	Pain on repeated Lx movement testing: 100% adjusted agreement (same session), 54 - 59% (different sessions)
	Neural tests
	Seated SLR: 100% adjusted agreement (same session), 73% (different sessions)
	Supine SLR: 100% adjusted agreement (same session), 48 - 67% (different sessions)
	Bilateral active SLR: 100% adjusted agreement (same session), 44% (different sessions)
	Femoral stretch: 100% adjusted agreement (same session), 54 - 56% (different sessions)
	Other tests
	Single Williams knee pull: 100% adjusted agreement (same session), 29% (different sessions)
	Repeated Williams Knee pull: 100% adjusted agreement (same session), 35% (different sessions)
	Single partial push up: 100% adjusted agreement (same session), 58% (different sessions)
	Repeated partial push-up: 100% adjusted agreement (same session), 53% (different sessions)
	Instability test: 100% adjusted agreement (same session), 51% (different sessions)
	Non-organic signs
	Trunk twist: 100% adjusted agreement (same session), 76% (different sessions)
	Head compression: 100% adjusted agreement (same session), 89% (different sessions)
	Skin tenderness: 100% adjusted agreement (same session), 73% (different sessions)
	Distraction (flip): 97-100% adjusted agreement (same session), 56% (different sessions)
Korbon (1987)	**Non-organic signs**
	Superficial or non-anatomic tenderness: 59 – 82% agreement
	Axial loading: 87% agreement, 0.69 correlation
	Rotation: 64% agreement, 0.57 correlation
Potter (1985)	**SIJ tests**
	Supine iliac gapping test: 94.12% agreement
	Side-lying iliac compression test: 76.47% agreement
Waddell (1980)	**Non-organic signs**
	Superficial tenderness: 80% agreement
	Non-anatomic tenderness: 80% agreement
	Simulation
	Axial loading: 78% agreement
	Rotation: 78% agreement
	Distraction (flip): 86% agreement (SLR)

Characteristics of individual studies:

Primary author	Patients (no. [M / F], mean age [range or SD]) / Examiners (no, occupation, practicing experience, experience with procedure and / or study training)
Fritz (2006)	60 (31/29), 36.6 (10.5) yrs/ 30 PTs, 10 experts (+/-), 10 5 yrs (-/-), 10 <5yrs (-/-)
Haswell (2004)	35 (16/19), 41.6 (13.2) yrs/ 4 PTs, >7 years (mean 12.5), NR /+
Heiss (2004)	45 (15/30), 41.3 (13.3) yrs/ 3 PTs, 18-29 yrs, -/+
Petersen (2004)	90 (36/54), 38 (11.6) yrs/ 4 PTs, 7-27 yrs, +/+
Fritz (2003)	20 (NR), NR/ NR, PTs, NR, NR /+
Hicks (2003)	63 (25/38), 36.0 (10.3) yrs/ 4 PTs, 2-8 yrs, +/+
Kilpikoski (2002)	39 (24/15), 40 (24-55) yrs/ 2 PTs, 5 years on average, +/-
Seymour (2002)	15 (NR); NR/ 6 PTs, 1-24 years (mean 9.3 years), +/+
White (2002)	37; (10/27), 37.2 (13.6) yrs/ 4 NR, 5 yrs minimum, NR/ +
Fritz (2000)	12 (7/5), NR/ 80 (40 PTs/ 40 PT students), 7.5 yrs (PTs), -/+
Hsieh (2000)	26 (14/12), 47.9 (13.6) yrs/ 8 Chir/MD, 3-6 years, 1 (+/+), 7 (-/+)
Razmjou (2000)	45 (20/25), 47 (14) yrs/ 2 PTs, 12 & 24 years, +/-
Vroomen (2000)	91 (48/ 43), 46 (11.2) yrs/ 3 MD, NR, +/NR
van Dillen (1998)	95 (41/54), 44.07 (13.29) yrs/ 5 NR, NR, -/+
Strender (1997)	71 (28/43), 37.7 ± 11.7 (PT group) 41.2 ± 15.7 (MD group)/ 4 (2Pts/ 2MD) , NR, +/+
Van den Hoogen (1996)	S1: 50 (25/25), 46 (NR), S2: 48 (25/23), 40 (NR)/ NR MD, NR, NR
Laslett (1994)	51 (NR)/ 6 NR, (NR/+)
Maher (1994)	90 (31/59), 45.37 (14.16) yrs/ 6 PTs, >5yrs (8-21), +/-
Njoo (1994)	61 (34/27), 36.2 (9.8) yrs/ 5 (1 MD/ 4 med students), NR, NR/+
Boline (1993)	28 NR, NR/ 3 Chir, NR, +/NR
Nice (1992)	50 (19/31), 39 (13.4) yrs/ 12 PTs, 3-17 yrs, 7 (+/+) 5 (-/+)
Waddell (1992)	60 (NR), NR/ 2 NR, NR, NR/NR
Keating (1990)	21 (3/ 18), NR (23-60) yrs/ 3 Chir, 2.5-10 yrs, NR/ +
Kilby (1990)	41 (18/23), 42 (18-68) yrs/ 2 PTs, NR, +/-
Spratt (1990)	42 (19/23), 38.9 (20.6-59) yrs/ 3 MD, NR, 2 +/+, 1-/+
McCombe (1989)	S1: 50 (26/24), 44.3 (12.2) yrs/ 2MD; S2: 33 (26/7),46.1 (14.6) yrs/ 2 (1 MD/ 1 PTs)
Boline (1988)	23 (NR), NR/ 2 Chir, NR, +/+
Korbon (1987)	39 (NR) , NR yrs/ 2 MD, NR, NR/NR
Potter (1985)	17 (10/7), 39 (24-58) yrs/ 8 PTs, NR , +/+
Waddell (1980)	50 (NR), NR/ 2 MD, NR, NR/NR

Abbreviations:

AROM, Active Range of Motion; Chir, Chiropractor; ICC, Intraclass Correlation Coefficient; K, Kappa Coefficient; Kw, Weighted Kappa Coefficient; Lx, Lumbar; MD, Medical Doctor; M / F, Male / Female; NR, Not Reported; PTs, Physiotherapists; ROM, Range of Motion; SD, Standard Deviation; SIJ, Sacroiliac Joint; SLR, Straight Leg Raise; S1, Sample 1; S2, Sample 2.

Appendix 2.6 Table of results of studies for types / level of training or experience of raters (Font: green, ≥ 0.7, red, ≥ 0.4, black < 0.4; highlighted cells represent changing results or evidence).

SYMPTOM RESPONSE ITEM	PRIMARY AUTHOR	STATISTIC*	VARIANCE**	LOE (R≥0.7+Q≥0.6)	LOE (R≥0.7+Q≥0.5)	LOE (R≥0.4+Q≥0.6)	LOE (R≥0.4+Q≥0.5)
Movement testing – Gross movement							
Pain on Lx movement	McCombe (1989)	K = 0.10 to 0.56 (S1)	NR	Limited	Limited	Limited	Limited
		K = 0.42 to 0.58 (S2)		Limited	Limited	Limited	Limited
Palpation							
Midline tenderness	McCombe (1989)	K = 0.38 (S1)	NR	Limited	Limited	Limited	Limited
		K = 0.47 (S2)		Limited	Limited	Limited	Limited
Paraspinal tenderness	McCombe (1989)	K = 0.11 (S1)	NR	Limited	Limited	Limited	Limited
		K = 0.38 (S2)		Limited	Limited	Limited	Limited
Buttock tenderness	McCombe (1989)	K = 0.31 (S1)	NR	Limited	Limited	Limited	Limited
		K = 0.34 (S2)		Limited	Limited	Limited	Limited
Sacroiliac tenderness	McCombe (1989)	K = 0.41 (S1)	NR	Limited	Limited	Limited	Limited
		K = 0.28 (S2)		Limited	Limited	Limited	Limited
Iliac crest tenderness	McCombe (1989)	K = 0.50 (S1)	NR	Limited	Limited	Limited	Limited
		K = 0.36 (S2)		Limited	Limited	Limited	Limited
Trigger point assessment							
Taut band	Hsieh (2000)	K = 0.78 (with expert)	NR	Limited	Moderate	Limited	Moderate
		K = 0.008 (among examiners)		Limited	Moderate	Limited	Moderate
Referred pain pattern	Hsieh (2000)	K = 0.27 (with expert)	NR	Limited	Moderate	Limited	Moderate
		K = 0.33 (among examiners)		Limited	Moderate	Limited	Moderate
Neural tests							
Pain on SLR	McCombe (1989)	K = 0.36 to 0.66 (S1)	NR	Limited	Limited	Limited	Limited
		K = 0.44 to 0.81 (S2)	NR	Limited	Limited	Limited	Limited
SLR crossed	McCombe (1989)	K = 0.74 (S1)	NR	Limited	Limited	Limited	Limited
		K = -0.02 (S2)	NR	Limited	Limited	Limited	Limited
SLR sciatic stretch test	McCombe (1989)	K = 0.37 (S1)	NR	Limited	Limited	Limited	Limited
		K = 0.62 (S2)	NR	Limited	Limited	Limited	Limited
Pain on Bowstring testing	McCombe (1989)	K = 0.26 to 0.49 (S1)	NR	Limited	Limited	Limited	Limited
		K = 0.11 to 0.20 (S2)		Limited	Limited	Limited	Limited
Femoral nerve stretch	McCombe (1989)	K = 0.23 to 0.37 (S1)	NR	Limited	Limited	Limited	Limited
		K = 0.50 to 0.77 (S2)		Limited	Limited	Limited	Limited

SIJ & other pain provocation tests							
Compression	McCombe (1989)	K = 0.09 (S1)	NR	**Limited**	**Limited**	**Limited**	**Limited**
		K = 0.16 (S2)		**Limited**	**Limited**	**Limited**	**Limited**
Distraction	McCombe (1989)	K = 0.11 (S1)	NR	**Limited**	**Limited**	**Limited**	**Limited**
		K = 0.36 (S2)		**Limited**	**Limited**	**Limited**	**Limited**
Maitland sacroiliac test	McCombe (1989)	K = 0.26 (S1)	NR	**Limited**	**Limited**	**Limited**	**Limited**
		K = 0.38 (S2)		**Limited**	**Limited**	**Limited**	**Limited**
Pain on hip flexion	McCombe (1989)	K = 0.40 (S1)	NR	**Limited**	**Limited**	**Limited**	**Limited**
		K = 0.42 (S2)		**Limited**	**Limited**	**Limited**	**Limited**
Pain on resisted external hip rotation	McCombe (1989)	K = 0.63(S1)	NR	**Limited**	**Limited**	**Limited**	**Limited**
		K = 0.38 (S2)		**Limited**	**Limited**	**Limited**	**Limited**
Combination of strategies							
Status change with Lx movement or sustained postures	Fritz (2000)	K (total) = 0.79	0.77 to 0.81	**Limited**	**Moderate**	**Limited**	**Moderate**
		K (PT) = 0.82	0.81 to 0.84	**Limited**	**Moderate**	**Limited**	**Moderate**
		K (PT students) = 0.76	0.76 to 0.77	**Limited**	**Moderate**	**Limited**	**Moderate**
		K (> 6 yrs experience) = 0.87	0.86 to 0.90	**Limited**	**Moderate**	**Limited**	**Moderate**
		K (< 6yrs experience) = 0.82	0.81 to 0.83	**Limited**	**Moderate**	**Limited**	**Moderate**
Non-organic signs							
Superficial tenderness	McCombe (1989)	K = 0.29 (S1)	NR	**Limited**	**Limited**	**Limited**	**Limited**
		K = 0.17 (S2)	NR	**Limited**	**Limited**	**Limited**	**Limited**
Simulation	McCombe (1989)	K = 0.25 (S1)	NR	**Limited**	**Limited**	**Limited**	**Limited**
		K = 0.48 (S2)	NR	**Limited**	**Limited**	**Limited**	**Limited**

*** Values in the stastitic column represent either single reliability values or the range of multiple results on a physical examination procedure.**

**** Reflects the 95% confidence interval where given.**

Abbreviations:

K, Kappa Coefficient; LOE, Levels of Evidence; Lx, Lumbar; NR, Not Reported; PT, Physiotherapist; Q, Quality; R, Reliability; SIJ, Sacroiliac Joint; SLR, Straight Leg Raise; S1, Sample 1; S2, Sample 2.

Appendix 3.1 Search strategies of online databases for prognostic studies investigating symptom response in spinal pain.

OVID-MEDLINE
1. exp Spinal Diseases/
2. ((spin$ or low back or back or lumbar or neck or cervical or thoracic) adj2 (pain or disorder$)).ab.
3. exp Prognosis/
4. prognos$.mp.
5. predict$.mp.
6. exp "Predictive Value of Tests"/
7. (measur$ adj pain).ab.
8. ((pain or symptom$) adj (provocation or change$ or response$ or pattern$ or behavi$)).ab.
9. (provocation or change$ or pattern$ or behavi$) adj10 (pain or symptom$ or sign$).ab.
10. 1 OR 2
11. OR/3-6
12. OR/7-9
13. 10 AND 11 AND 12

OVID-EMBASE
1. ((spin$ or low back or back or lumbar or neck or cervical or thoracic) adj2 (pain or disorder$)).ab.
2. exp Spine Disease/
3. exp PROGNOSIS/
4. prognos$.mp.
5. predict$.mp.
6. exp prediction/
7. (measur$ adj pain).ab.
8. ((pain or symptom$) adj (provocation or change$ or response$ or pattern$ or behavi$)).ab.
9. ((provocation or change$ or pattern$ or behavi$) adj10 (pain or symptom$ or sign$)).ab.
10. 1 or 2
11. or/3-6
12. or/7-9
13. 10 and 11 and 12

OVID CINAHL
1. exp Spinal Diseases/
2. ((spin$ or low back or back or lumbar or neck or cervical or thoracic) adj2 (pain or disorder$ or disease$)).ab.
3. exp PROGNOSIS/
4. prognos$.mp.
5. exp PREDICTIVE VALIDITY/ or exp "PREDICTIVE VALUE OF TESTS"/
6. predict$.mp.
7. (measur$ adj pain).ab.
8. ((pain or symptom$) adj (provocation or change$ or response$ or pattern$ or behavi$)).ab.
9. ((provocation or change$ or pattern$ or behavi$) adj10 (pain or symptom$ or sign$)).ab.
10. 1 or 2
11. or/3-6
12. or/7-9
13. 10 and 11 and 12

OVID AMED
1 exp Spinal disease/
2 ((spin$ or low back or back or lumbar or neck or cervical or thoracic) adj2 (pain or disorder$)).ab.
3 exp Prognosis/
4 prognos$.mp.
5 predict$.mp.
6 exp "Predictive value of tests"/
7 (measur$ adj pain).ab.
8 ((pain or symptom$) adj (provocation or change$ or response$ or pattern$ or behavi$)).ab.
9 ((provocation or change$ or pattern$ or behavi$) adj10 (pain or symptom$ or sign$)).ab.
10 1 or 2
11 or/3-6
12 or/7-9
13 10 and 11
14 12 and 13

Appendix 3.2 List of prognostic citations excluded after the full-text screening and justification for their exclusion.

REASON FOR EXCLUSION	REFERENCES
Inadequate data	Troup et al., 1981a; Lloyd and Troup, 1983; Korbon et al., 1987; Donelson et al., 1990; Pennie and Agambar, 1991; Lancourt and Kettelhut, 1992; Hopwood and Abram, 1993 (full-text not available); Werneke et al., 1993; Radanov et al., 1994; Jordan, 1996 (full-text not available); McIntosh et al., 2000; Richter et al., 2004.
Irrelevant tests	Troup et al., 1981b; Mendelson et al., 1983; Murphy and Cornish, 1984; McNeil et al., 1986; Sandstrom, 1986; Gore et al., 1987; Bradish et al., 1988; Deyo and Diehl, 1988; Doxey et al., 1988; Lanier and Stockton, 1988; Gallagher et al., 1989; Hurri, 1989; Polatin et al., 1989; Lacroix et al., 1990; Radanov et al., 1991; Lehmann et al., 1993; Von Korff et al., 1993; Coste et al., 1994; Haazen et al., 1994; Lindstrom et al., 1995; Main and Watson, 1995; Radanov et al., 1995; Cherkin et al., 1996; Ohlund et al., 1996; Radanov and Sturzenegger, 1996b; Dionne et al., 1997; Infante - Rivard and Lortie, 1997; Ingemarsson et al., 1997; Nordin et al., 1997; Bendix et al., 1998; Haldorsen et al., 1998; Heikkila et al., 1998; Skargren and Oberg, 1998; Macfarlane et al., 1999; Muller et al., 1999; Schiottz-Christensen and Nielsen, 1999; Soderlund and Linberg, 1999; Thomas et al., 1999; Van der Weide et al., 1999; Vendrig, 1999; Carey et al., 2000; Nyiendo et al., 2000; Potter et al., 2000; Seferlis et al., 2000; Soderlund et al., 2000; Valat et al., 2000; Cutler et al., 2001; Warren and Warren, 2001; Alexandre, 2002; Axen et al., 2002; Chiradejnant et al., 2002; Damush et al., 2002; Goldstein et al., 2002; Hunt et al., 2002; Tubach et al., 2002; Cassidy et al., 2003; Schectman et al., 2003; IJzelenberg and Burdorf, 2004; Michaelson et al., 2004; Watson and Booker, 2004; Axen et al., 2005b; Bekkering et al., 2005; Dionne et al., 2005; George and Hirsh, 2005; Gun et al., 2005; Hagen et al., 2005; Ylinen et al., 2005; Boersma and Linton, 2006; Gorbach et al., 2006; Jellema et al., 2006; Heneweer et al., 2007.
Lack of or no separate analysis for eligible groups	Anonymous, 1966; Pedersen, 1981; Norris and Watt, 1983; Roland and Morris, 1983; Dworkin et al., 1986; Burton and Tillotson, 1991; Klenerman et al., 1995; Radanov and Sturzenegger, 1996a; Polatin et al., 1997; White et al., 1997; Fritz et al., 2000b; Taylor et al., 2001; Fritz and George, 2002; Kool et al., 2002; Vingard et al., 2002; Jakobsson et al., 2003; Koopman et al., 2004; Axen et al., 2005a; Grotle et al., 2005.
Inappropriate design (cross – sectional comparisons)	Waddell, 1980; Waddell et al., 1984; Parascandola, 1993; Smythe, 1994; Dionne et al., 1999 ; Haas et al., 2002; Lyle et al., 2005

Appendix 3.3 Characteristics of included prognostic cohorts.

Primary author	QA score	N	Participant characteristics and treatment	Follow-up	Symptom response variables	Other potentially prognostic variables
Neck pain						
Cleland (2007)	16	80	Pts with / without unilateral extremity pain referred for PT at a hospital site Tx manipulation; CROM exercises and advice to remain active	After treatment, mean time 2.3 days to 6.3 days	CP / P with cervical motion testing, Cx compression, Cx distraction test, Upper limb test	Body diagram, pain intensity NRS, NDI, FABQ, mode of onset, nature and location of symptoms, aggravating and relieving factors, prior history of neck pain, neurological assessment findings, postural assessment findings CROM, length and strength of upper quarter, endurance of deep neck flexors, Cx and Tx segmental mobility, special tests
Tseng (2006)	16	100	Pts referred for PT at the outpatient department of two hospital sites Cx manipulation	After treatment (1 session)	Cx compression, Cx distraction	Demographic information, diagnoses, area of complaints, onset patterns, stages of symptoms, unilateral or bilateral pattern, aggravating and relieving activities or movements, self report measures i.e. current and worse pain intensity NRS, NDI, psychological wellbeing (CHQ12), CROM, Side gliding mobility
Tuttle (2005, 2006)	11	29	Pts attending a private PT clinic Manual therapy, advice	Mean time 6.1 days (2 - 14 days) and at discharge	Within session changes in pain location , within session changes in pain intensity	Within and between session changes in total and limited ROM, GPES
Back pain						
George (2005)	17	28	Pts classified for specific exercise intervention; referred to four PT clinics Specific exercise	6 months	Centralisation with lumbar testing	History of LBP, duration of LBP, leg pain during present episode, ODQ, FAB about physical activity and FAB about work, present pain intensity NRS
Hicks (2005)	17	54	Pts referred to three PT outpatient clinics Stabilization exercise programme	8 weeks	Lx posterior shear test, , Lx prone instability test (+) / (-)	Demographic information, mode of onset, duration of symptoms, number of previous episodes, response to previous treatments, distribution of symptoms for current episode, ranking of activities, pain intensity, FABQ, disability (ODQ), ROM, aberrant motions, segmental mobility, ligamentous laxity, muscle endurance and strength, active SLR
Skytte (2005)	12	60	Pts referred from primary care to the Rheumatology Department of a University Hospital Standardised treatment pathway including medication, advice and exercises	1, 2, 3, 6 and 12 months	CP	N/A
Niemisto (2004)	15	204	Employed subjects suffering from chronic complaints	12 months	CP, Lx neural tension test	Sociodemographic variables, LBP characteristics, disability (ODQ), HRQOL, work ability, psychological variables, physical activity;

			Randomised into combined manipulation, exercise and physician consultation or consultation alone			treatment group, mobility
Hahne (2004)	15	53	Pts presenting to six private PT clinics Passive joint mobilization, education / advice, exercises, McKenzie movement therapy, electrotherapy, soft tissue massage, traction, manipulation, muscle stretching, neural stretching	Mean time 4.8 days (2 - 11 days)	Within session changes in pain intensity	Between session changes in pain intensity and AROM
Werneke (1999, 2001, 2004)	18	223	Pts with acute LBP with or without referred symptoms referred by a physician to PT Exercise, education, therapeutic modalities	12 months	CP, non-CP	Age, gender, multiple sites of pain, leg pain at intake, pain intensity at intake, duration of symptoms, prior spinal pain, prior work loss, prior Worker's Compensation Benefits, Payer, Job physical demands, work status, work loss, work satisfaction, non-organic signs, overt pain behaviours, depressive symptoms, FAB, rehabilitation program factors,
Enthoven (2003)	16	44	Pts visiting two Primary Health Care Centres (GP or PT as first contact) GP consultation: medication and additional diagnostic investigations if necessary	4 weeks, 12 months	Increase in pain intensity after physical examination	Pain, disability (ODQ), somatic or depressive distress, general health, mobility, endurance
Flynn (2002)	16	71	Pts referred to outpatients PT clinics Manipulation	After treatment	Lx, SI pain provocation tests, CP / P with single motion	Demographic information, pain intensity, pain location, disability (ODQ), FABQ. duration of symptoms, mode of onset, prior history of LBP, episodes becoming more frequent,, best position, non-organic signs, mobility tests, mobility discrepancy/ symmetry tests, SLR, presence of lateral shift, hypomobility with spring testing
Viikari-Juntura (1998)	6	242	Workers presenting to OHS centres Absence from work (5 days max), physiotherapy and ergonomic advice	60 days	Pain in Lx or leg movement	Demographics, job category, sick leave during 60 days prior to the examination. LBP symptoms location, pattern, easing factors, weakness of foot, incontinence symptoms, locking, neurological assessment variables, SLR
Karas (1997)	11	171	Pts with or without referred leg pain presenting to five clinics of the CBI CBI protocol of active exercise	6 months	CP	Waddell scores
van den Hoogen (1997)	14	443	Pts from 11 general practice clinics GP consultation including PT	12 months	SLR: Pain in the low back or buttock, thigh or leg / foot	Demographics, duration of LBP, sciatica, mode of onset, history of preceding episodes, history of surgery because of LBP, pain severity, disability, occupational back load, time of commuting by car, NHP, GP judgement, treatment, Limited SLR, maximal lumbar flexion, pelvic tilt and scoliosis
Burton (1995, 2004)	15	252	Pts with a new episode seen in a group practice of osteopaths Manipulation, advice, exercise	12 months , average of 4 years	Lx neural tension test (SLUMP)	Demographic information, details of history of back trouble, treatment, symptomatic details, physical examination (including nerve root tension tests), psychosocial variables
Long	11	223	Pts with or without referred leg symptoms attending	9 months,	CP	N/A

(1995)			a private interdisciplinary rehabilitation centre programme Work hardening programme (i.e. physiotherapy, exercise conditioning, work simulation, education, psychological intervention)	24 months.		
Neck and back pain						
Werneke (1999, 2003)	16	289	Pts with acute neck and back pain presenting to PT clinics Exercise, education, therapeutic modalities	At discharge, 12 months	CP, non-CP	N/A
Hellsing (1994)	12	120	Pts presenting to community primary care / outpatients Bed rest up to 5 days or severe cases. Active care (e.g. education, biomechanical counselling, physical activity, workplace intervention).	12 months	Number of positive pain tests on movement	Medical history, clinical examination, function al ability(ADL), pain intensity (VAS), number of pain free days during past week, disability (100 mm VAS), sick leave 2 years prior to study

Abbreviations:

ADL, Activities of Daily Living; CBI, Canadian Back Institute; CHQ12, 12-item Chinese Health Questionnaire; CP, Centralisation Phenomenon; CROM, Cervical Range of Motion; Cx, Cervical; FAB, Fear Avoidance Beliefs; FABQ, Fear Avoidance Beliefs Questionnaire; GPES, Global Perceived Effect Scale; HLQoL, Health-Related Quality of Life; LBP, Low Back Pain; N, number of participants; N / A, Not Applicable; NDI, Neck Disability Index; NRS, Numerical Rating Scale; ODQ, Oswestry Disability Questionnaire; P, Peripheralisation; Pts, Patients; PT, Physiotherapy; ROM, Range of Motion; SI, Sacro-iliac; SLR, Straight Leg Raise; Tx, Thoracic; VAS, Visual Analogue Scale.

Appendix 3.4 Quality assessment results for included studies of prognostic factors. Grey areas represent items where assessors disagreed.

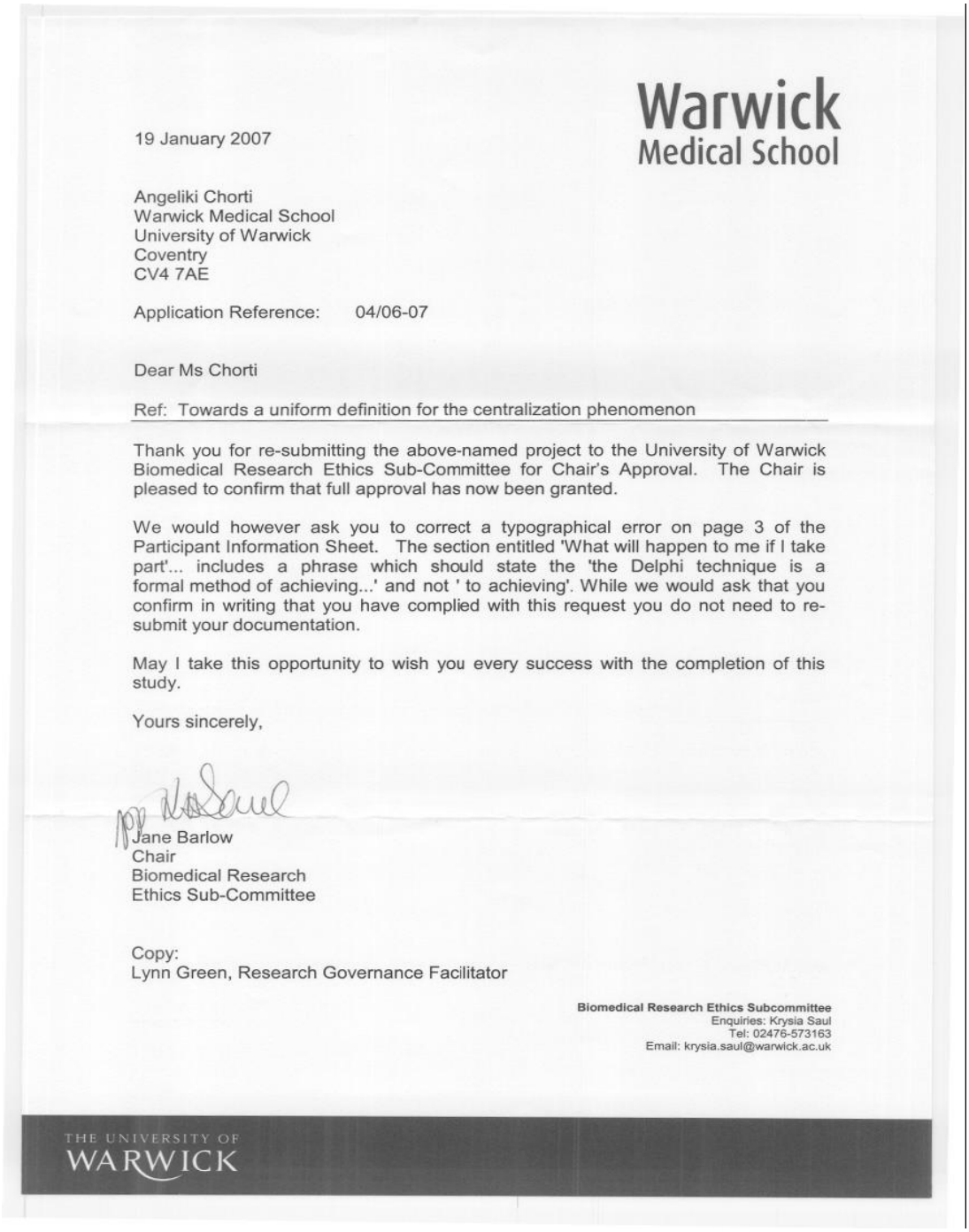
Primary author	A	B	C	D	E	F	G	H	I	J	Total score	Other comments
Cleland (2007)	2	1	2	0	2	2	1/**2**	1	2	2	**16**	Development of clinical prediction rule based on small numbers and short-term data; transferability in question, arbitrary cut-off points for the dichotomisation of outcome into success and failure groups
Tseng (2006)	2	1	2	0	2	2	0/**2**	1	2	2	**16**	Development of clinical prediction rule based on small numbers and short-term data; reliability of potential factors not established; outcome assessor not blinded; arbitrary definition and cut-off points of treatment successes, 50% of unexplained variance, transferability in question
George (2005)	2	1	1/**2**	1/**2**	2	2	2/**0**	1/**2**	1/**2**	2	**17**	Results based on a secondary analysis of data on acute low back pain patients; reliability of potential factors not established; small sample size; 51% and 71% of unexplained variance for disability and pain intensity, transferability in question
Skytte (2005)	2	1	1/**0**	0	2	2	1/**3**	1/**2**	1	0	**13**	Study restricting to participants with referred symptoms and sciatica/ severe presentation; small and unjustified sample size
Tuttle (2005, 2006)	2	0/**1**	1	0	1	1	2	1	2	2/**0**	**11**	Small and unjustified sample size; reliability of potential factors established in asymptomatic volunteers; results based on short-term follow up data
Niemisto (2004)	2	0	**1**/0	0	**2**/1	2	3	1	2	2	**15**	Study sample restricted to chronic low back pain participants; recruitment via advertisement

Hahne (2004)	2	1	2	2/**0**	2	1	2	1/**2**	2	2/**1**	**15**	Small and unjustified sample size, especially in subgroup analyses resulting in lack of precision; reliability of potential prognostic variables established in asymptomatic subjects; Findings generalisable to physiotherapists working in private practice
Enthoven (2003)	0/**2**	1	2	2/**0**	2	0/**1**	2/**3**	1	1/**2**	1	**15**	Unjustified sample size; reliability not established for potential prognostic variables; arbitrary threshold for the division of patients reporting pain after the baseline physical examination
Hicks (2003, 2005)	2	1	2	0	2	2	2	2	2	2	**17**	Development of clinical prediction rule based on small and unjustified numbers; short-term follow up; arbitrary definition and cut-off point for treatment success and failure groups
Flynn (2002)	2	1	2	0	2	2	2	**1**/2	2	2	**16**	Development of clinical prediction rule based on small and unjustified numbers; transferability questioned, arbitrary cut-off points for outcomes; results generalisable to outpatient physiotherapy clinics
Werneke (1999, 2003)	2	1	1	**1**/0	2	2	2	2	2	1	**16**	Temporal difference in the measurement of potential predictors (multiple-visit definition for centralisation versus baseline definition for other potential prognostic variables)
Werneke (1999, 2001, 2004)	2	1	1	**1**/0	2	2	3	2	2	2	**18**	Temporal difference in the measurement of potential predictors (multiple-visit definition for centralisation versus baseline definition for other potential prognostic variables)
Viikari-Juntura (1998)	0	1	1	0	1	0	**0**/2	1	1/2	1/2	**6**	More severe spectrum of participants; restricted outcome selection; no justification for sample

												size; results generalisable in an occupational health setting
Karas (1997)	2	1	1	0	1	2	2	1	1	0/1	**12**	No sample size justification; temporal difference in the measurement of potential predictors; restricted outcome selection
van den Hoogen (1997)	2	1	2	0	2	1	1	1	2	2	**14**	No justification for sample size; no standardisation in the intervention provided; possible overrepresentation of more severe cases at follow-up
Burton (1995, 2004)	2	1	2	2/**1**	2	1	1	1	2	2	**15**	No sample size justification; subgroup analyses based on small numbers; variability in the provided treatment; results generalisable in osteopathic practices
Long (1995)	2	1	1	0	2	1	3/**1**	1	2	0	**11**	No sample size justification; reliability of potential prognostic indicators not established; low rates for long-term follow-up; results generalisable to chronic low back pain
Hellsing (1994)	1	1	2	0/**2**	2/**1**	1	**2**/3	1	1	0	**12**	No sample size justification; study potentially influencing referrals; treatment not standardised; results generalisable to primary care

Quality categories: A: Case Definition; B: Source population; C: Representativeness; D: Patient Selection; E: Participants; F: Treatment; G: Follow-up; H: Outcome; I: Prognostic factors; J: Analysis.

Appendix 4.1 Biomedical Research Ethics Sub-Committee full approval letter and final confirmation email.



Warwick
Medical School

19 January 2007

Angeliki Chorti
Warwick Medical School
University of Warwick
Coventry
CV4 7AE

Application Reference: 04/06-07

Dear Ms Chorti

Ref: Towards a uniform definition for the centralization phenomenon

Thank you for re-submitting the above-named project to the University of Warwick Biomedical Research Ethics Sub-Committee for Chair's Approval. The Chair is pleased to confirm that full approval has now been granted.

We would however ask you to correct a typographical error on page 3 of the Participant Information Sheet. The section entitled 'What will happen to me if I take part'... includes a phrase which should state the 'the Delphi technique is a formal method of achieving...' and not ' to achieving'. While we would ask that you confirm in writing that you have complied with this request you do not need to re-submit your documentation.

May I take this opportunity to wish you every success with the completion of this study.

Yours sincerely,

pp Jane Barlow
Chair
Biomedical Research
Ethics Sub-Committee

Copy:
Lynn Green, Research Governance Facilitator

Biomedical Research Ethics Subcommittee
Enquiries: Krysia Saul
Tel: 02476-573163
Email: krysia.saul@warwick.ac.uk

THE UNIVERSITY OF
WARWICK

Chorti, Angeliki

From: Barlow, Jane **Sent:** Wed 24/01/2007 18:05
To: Chorti, Angeliki
Cc: McCarthy, Christopher; Saul, Krysia; Lamb, Sarah
Subject: RE:
Attachments:

Dear Angeliki
Thank you for your email confirming that you have made the requested changes.

I would also like to take this opportunity to confirm that I have spoken with Professor Lamb this afternoon, and am happy for you to make the proposed amendment in which you have added the word 'endeavour' to the information sheet.

It is not always possible to recruit the sample proposed, but this need not concern the ethics committee.
With all best wishes
Jane

Dr Jane Barlow
Reader in Public Health
Warwick Medical School
University of Warwick
Coventry
CV4 7AL

Tel: 02476 574884
Fax: 02476 574879

Website: www.warwick.ac.uk/fac/med/staff/barlow/
Directions to WMS: http://www.warwick.ac.uk/fac/med/newsfront/archive/pre0607/directions/

From: Chorti, Angeliki
Sent: Wed 24/01/2007 16:21
To: Barlow, Jane
Cc: McCarthy, Christopher; Saul, Krysia; Cromwell, Elizabeth
Subject:

Application reference: 04/06-07

Ref: Towards a uniform definition for the centralization phenomenon

Dear Ms Barlow,

I am writing to confirm that I have corrected the typographical error on page 3 of the Participant Information Sheet as requested.

Yours sincerely,

Angeliki Chorti
Warwick Postgraduate Research Fellow/ PhD student
Warwick Medical School
University of Warwick
Coventry CV4 7AL
UK
Tel: 024765 74653
E-mail: **A.Chorti@warwick.ac.uk**

Appendix 4.2 Delphi study pack: invitation letter.

[Enter name & address]

[Enter date]

Dear [name],

Toward a uniform definition for the centralization phenomenon: a Delphi study

You are invited to participate in a study which involves a 3-round Delphi technique aiming to achieve consensus on the operational criteria and related definition of the centralization phenomenon and identify further clinical or research-related issues that need to be addressed in this study or in future research into centralization. This study forms part of a PhD degree at the Medical School of the University of Warwick in the United Kingdom and its findings will be used to inform subsequent reliability and prognostic studies.

Before you decide whether you want to take part in the study, please read the enclosed information in order to understand why this study is conducted and what it involves. You do not have to immediately decide whether you want to participate in the study; you can first discuss it with others if you wish to. However, we would be grateful if we could have a response from you (we pay the postage) by **[date]**. If you would like further information, please contact me (my contact details are given below).

Your involvement is important to the success of this project. We would therefore be grateful if you could support this study through your participation and commitment.

Kind regards,

Angeliki Chorti
Chief Investigator and Study Co-ordinator
Warwick Medical School
University of Warwick, Coventry CV4 7AL
United Kingdom
Tel: 0044(0) 2476574653
E-mail: A.Chorti@warwick.ac.uk

Appendix 4.3 Delphi study pack: participant information sheet.

PARTICIPANT INFORMATION SHEET

Toward a uniform definition for the centralization phenomenon: A Delphi study

You are invited to participate in a study which is undertaken as part of a research degree project at the Medical School of the University of Warwick in the United Kingdom. Before you decide whether you want to take part in the study, it is important for you to understand why this study is conducted and what it involves. Please take time to read the following information carefully. Part 1 describes the purpose of the study and what you will be required to do if you wish to take part. Part 2 provides more detailed information about the conduct of the study. You do not have to immediately decide whether you want to participate in the study; you can (if you want to) discuss it first with others and then make up your mind upon participation. However, we would be grateful if we knew your decision by [date]. If you have any further questions, please do ask and we will be happy to provide you with more information.

PART 1

What is the purpose of this study?

Which are the core criteria for the operational definition of the centralization phenomenon in the spine? This is a question often being raised but rarely being answered in the same way. The purpose of this study is therefore to achieve agreement on the definition of the centralization phenomenon in the clinical practice and research of the spine. In particular, emphasis will be placed on the establishment of uniform criteria for testing and classification procedures based on the physical examination sign centralization. A second objective will be to identify further clinical or research-related issues that need to be raised, resolved by consensus or explored in future research into the centralization phenomenon. This objective may be as important as achieving agreement and if consensus can not be achieved, this study will also give insights into necessary areas for future research.

What is centralization?

Centralization was originally described by McKenzie as a clinical phenomenon occurring when the patient reports that the pain moves from a distal area to a location more central or near midline position in the spine during spinal movement testing. However, since the original description, several refinements have been made resulting in various definitions in the literature. Differences in the type of the loading strategy (e.g. single versus repeated movements, repeated movements only versus mobilization and manipulation strategies), the direction of the loading strategy (e.g. one versus multiple directions), criteria for a positive CP (e.g. decrease in intensity versus change in the anatomical location of symptoms), time frame for the judgment of CP (e.g. one versus multiple visits) and classification groups e.g. centralization/ non-centralization, centralization/peripheralization/no symptom change, centralization, partial reduction, non-centralization are some examples illustrating this variability.

Why is this study necessary?

It has been suggested that the observed variation in definitions of centralization may have serious implications for research and clinical practice. For example, since centralization is commonly used in the diagnosis, prognosis and treatment of patients with spinal problems, the use of different definitions may result in inconsistent reports of prevalence, diagnostic categories, prognostic indicators and therapy. It has also been suggested that although the use of various definitions of centralization has been supported in the literature, studies have not consistently favoured one approach over another resulting in highlighting the problem rather than solving it. In the light of the above, collaborative efforts among clinicians and researchers are required to standardise the criteria for centralization in clinical practice and research.

Why have I been chosen?

You have been chosen because your colleagues nominated you and/or you have had research or teaching responsibilities and clinical experience of using the centralization phenomenon and the system involving it.

Do I have to take part?

No. It is entirely up to you to decide whether you want to participate or not. You will be given the study pack which includes this information sheet and if you decide to participate, you will be asked to sign two copies of a consent form. A copy of this information sheet together with one signed copy of the consent form will be given for you to keep. You are still free to withdraw at any time and without giving any reason and this having no implications on your rights and benefits.

What will happen to me if I take part?

This study endeavours the participation of two equal in number groups, one representing the McKenzie and one the Delitto classification system, in a Delphi survey technique. The Delphi technique is a formal approach to achieving a group decision from a panel of informed individuals through a series of questionnaires sent in consecutive rounds.

Self-administered questionnaires will be sent to you and to the other members of the Delphi group four times within a period of approximately 4 months. Further information about what each questionnaire includes and what is required of you is provided below.

What do I have to do?

First, you will be sent the study pack. You will be asked to sign two copies of the study's consent form and fill in a questionnaire about you also including questions about your suitability for the study.

Round 1

In round one, you will be asked to vote for your preferred definition of centralization and on further issues that need to be addressed. The opinions of the individuals who responded in the questionnaire will be grouped into common areas and a quantitative frequency analysis will be performed by two independent members of the research team. Results will then be sent back to you in the form of a new questionnaire in the second round.

Round 2

In the second round, you will be asked to rank your level of agreement with opinions from round one on a five-point scale (1 totally disagree – 5 totally agree) and provide written comments where appropriate. The decision for the inclusion of an item will be made if 80% of the Delphi members agree that it should be included. A margin of variability of 5% will also be permitted and items reaching 75% agreement will be sent to the third round for reconsideration in the light of other participants' opinions. Finally, items falling below 70% will be excluded.

Round 3

This is the final round of the Delphi study. In this round, you will be presented with the summary scores and comments for inclusion of each item and will be asked to decide with a 'yes/no' option whether remaining items should be included or excluded.

Completing and returning the questionnaires

Completing the questionnaires in each round should not take more than a few minutes. Sometimes it may take longer than this, depending on the suggestions and comments you make. Please take your time to read the instructions and questions carefully and make sure that you have answered all relevant sections. Once finished, you must return your responses to the addresses provided in the questionnaires. Please note that each time you have 3 weeks to respond; if no response has been received before the deadline, reminders will be sent out to you.

What will happen after the study?

You will be sent a report with the results of the final round. You will also be given the opportunity to make your final comments.

What are the possible benefits and risks of taking part?

We can not promise you any direct benefit from taking part in the study. Completing the questionnaires may take some of your time. However, the contribution you make may facilitate communication and promote collaborative efforts among clinicians and researchers and lead to improving the care of patients. Please note that all the information about your participation in this study will be kept confidential. There is a possibility of your colleagues knowing of your participation in this study, but even if your participation is known to them,

your opinions and comments will remain strictly anonymous. Further details are included in Part 2.

This is the end of Part 1 of the Information Sheet. If the information in Part 1 has interested you and you are considering participation, please continue to read the additional information in Part 2 before making any decision.

PART 2

What if relevant new information becomes available?

If new information that answers all the research questions on the definition of centralization becomes available during the course of the Delphi study, you will be informed about it and discuss whether you wish or should continue in the study. If you decide to continue, you will be asked to sign an updated consent form. If the study stops for any other reason, you will be informed about the reasons for this action.

What will happen if I don't want to carry on with the study?

It is up to you to decide whether you wish to continue or withdraw from the study. However, information collected up to your withdrawal will still be used.

Will my taking part in this study be kept confidential?

All your details will remain confidential. No information that identifies you will be shared by anyone other than the research team and access to view identifiable data will be permitted only to authorized individuals. Even if your participation is known by your colleagues, your judgments and opinions will still remain strictly anonymous. We will make sure that the information you provide is presented in the form of statistical summaries and your comments are anonymous avoiding any direct quotation of what you have said.

What if there is a problem?

Any research where participant data is involved may carry the risk of negligent harm by breach of confidence. Every effort will be made to avoid this possibility. If you are harmed and this is due to someone's negligence, you may have grounds for a legal action for compensation against the University of Warwick; however, you may still have to pay your legal costs.

What use will be made with the information provided?

The results of the Delphi study will be used to inform a reliability and prognostic study investigating the centralization phenomenon in neck pain. Findings from the Delphi study may be presented to medical and health professionals and submitted to scientific conferences and peer-reviewed/ professional journals. Because this study is part of the requirements for a PhD degree, internal publication will be sought through the PhD thesis.

Who has reviewed the study?

This study has been reviewed and received a favourable ethical decision from the Biomedical Research Ethics Sub-committee of the University of Warwick.

Contact details

If you wish to receive further information about the study, please contact:

Angeliki Chorti
Chief Investigator and Study Co-ordinator
Warwick Medical School
University of Warwick
Coventry CV4 7AL
Tel: 0044 (0) 24765 74653
Email: A.Chorti@warwick.ac.uk

If you wish to make a complaint about the study, please contact:

Professor Sarah Elizabeth Lamb
Director of Clinical Trials Unit
Warwick Medical School
University of Warwick
Coventry CV4 7AL
Tel: 0044 (0) 24765 75855
E-mail: S.Lamb@warwick.ac.uk

OR

Dr. Chris McCarthy
Assistant Professor in Rehabilitation
Warwick Medical School
University of Warwick
Coventry CV4 7AL
Tel: 0044 (0) 24765 75856
E-mail: C.J.McCarthy@warwick.ac.uk

Thank you for taking time to read this sheet.

Appendix 4.4 Delphi study pack: participant eligibility questionnaire.

PARTICIPANT INFORMATION AND ELIGIBILITY QUESTIONNAIRE

Toward a uniform definition for the centralization phenomenon: A Delphi study

We would like to know more about you. Some of the information you give us will be used to determine your suitability for this study and some other to obtain a profile of the Delphi group.

Please answer the following questions in the spaces provided. If there is not enough space, please continue on a separate sheet of paper and attach this to your form indicating the question that it relates to.

Your answers will be kept strictly confidential and the information that you provide here will not be used by any other than the research team at the University of Warwick.

If you have any difficulties or questions relating to this questionnaire, please do not hesitate to contact Miss Angeliki Chorti (Tel: 0044(0)2476574653, E-mail: A.Chorti@warwick.ac.uk) who will be happy to provide you with further assistance.

When the questionnaire is completed, please return it ***by post together with the two copies of the consent form (using the pre-paid envelope) by – [date] –*** *(details given in the last page). Please use a BLACK or BLUE pen rather than a pencil.*

Thank you very much for your time

Section 1

We would like to know more about you. Some of the information you give us will be used to contact you and some other will be used to obtain a profile of the Delphi group.

Please answer the following questions in the spaces provided. If there is not enough space, please continue on a separate sheet of paper and attach this to your form indicating the question that it relates to.

1. Surname……..

2. Name..……..

3. Date of birth ……/……/…… (dd/mm/yy)

4. Gender Male/ Female
(Please delete as appropriate)

5. Preferred contact address …….....

6. Town/ City..... ……..

7. County/ State ……... ……

8. Postcode/ Zip code……

9. Country …..…….……

10. Telephone (including STD code)

...……..……..….. (Home)

……..……..…… (Other)

11. E-mail ……..……

12. Basic Professional Qualification(s)……..…….

13. Other professional or academic qualifications and memberships (e.g. this may include specialty certification, postgraduate academic degrees etc.)

……

14. Current professional activities

□ Clinical

□ Teaching

□ Research

Thank you for taking the time to complete this section. Please proceed to Section 2 in the following pages.

Section 2

This section includes questions that form part of the eligibility criteria for this study. Please answer all questions in this section. For questions 15, 17, 19, 20 and 21, please select one answer only.

15. Have you used centralization in your clinical practice?

Yes/No

(Please delete as appropriate)

16. How long have you been using centralization in your clinical practice?

Years ….. Months …….

17. Please state which system you predominantly use to elicit centralization

McKenzie system/ Delitto system/ None of the above

(Please delete as appropriate)

18. If answered 'none of the above' in question 17, please state details

……..

19. Are you currently involved in ongoing research concerning the centralization phenomenon and/or the McKenzie/Delitto system?

Yes/No

(Please delete as appropriate)

20. Have you published or co-authored research publications in the peer-reviewed literature concerning the centralization phenomenon and/or the McKenzie/Delitto system?

Yes/No

(Please delete as appropriate)

21. Do you have any teaching responsibilities involving the centralization phenomenon and/or the McKenzie/Delitto system?

Yes/No

(Please delete as appropriate)

22. If answered 'Yes' in question 21, please state details

……

This is the end of this questionnaire.

Please return your completed form
(together with the two copies of the consent form)
by using the pre-paid envelope no later than

[date]

Angeliki Chorti
Warwick Medical School
University of Warwick
Coventry CV4 7AL
United Kingdom

Thank you for your time

Appendix 4.5 Delphi study pack: consent form.

CONSENT FORM

Toward a uniform definition for the centralization phenomenon: A Delphi study.

INSTRUCTIONS

*In order to have a written record of your agreement to participate in this study, you need to complete and sign **two** copies of the consent form given below. Before doing so, please make sure that you read and understand the information provided about this study.*

*Please note that in order to consent to take part in this study, you must agree **to each of the statements provided by ticking each box in the form and add your name, signature and date at the bottom of the page**. If you leave a blank box or do not complete this form, we will consider that you are not willing to take part in this study.*

*Please **post the two copies of** this form by **[date]** to:*

Angeliki Chorti
Warwick Medical School
University of Warwick
Coventry CV4 7AL
United Kingdom

For any queries:
A.Chorti@warwick.ac.uk
Tel: 0044(0) 24765 74653

CONSENT FORM

Toward a uniform definition for the centralization phenomenon: A Delphi study.

Please initial box

1. I confirm that I have read and understand the Participant Information sheet dated 21/12/2007 for the above study. I have had the opportunity to consider the information, ask questions and have had these answered satisfactorily. ☐

2. I understand that my participation is voluntary and that I am free to withdraw at any time, without giving any reason, without my legal rights being affected. ☐

3. I understand that the research team will be informed of my participation in the study, and authorized individuals from the Medical School of the University of Warwick may have access to my anonymised data. I give permission to these individuals where it is relevant to have access to the data I have provided. ☐

4. I understand that the information I provide for the study will remain confidential and that I will be given anonymity in any publications or reports that arise from this research ☐

5. I agree to take part in the above study. ☐

_____ _____ _____

Name of Participant (BLOCK CAPITALS) Date Signature

_____ _____ _____

Name of Researcher Date Signature

N.B. When completed, 1 for participant; 1 for researcher

Study Number:
Participant Identification Number for this study:

FIRST ROUND QUESTIONNAIRE

Toward a uniform definition for the centralization phenomenon: A Delphi study

This is the first round questionnaire. This form comprises questions about the centralization phenomenon and may take a few minutes to complete. Please read each question carefully and make sure that you have answered ***all*** *the questions in the spaces provided. If there is not enough space, please continue on a separate sheet of paper and attach this to your form indicating the question that it relates to.*

Your answers will be kept strictly confidential and the information that you provide here will not be used by any other than the research team or authorised individuals at the University of Warwick. In order to ensure that your opinions remain anonymous, we would be grateful if your answers do not bear any personal identifiers and you do not talk about your answers with anyone.

If you have any difficulties or questions relating to this questionnaire, please do not hesitate to contact Miss Angeliki Chorti (Tel: 0044(0)2476574653, E-mail: A.Chorti@warwick.ac.uk) who will be happy to provide you with further assistance.

When the questionnaire is completed, please return it electronically or on paper by – date – (details given in the last page). If you decide to complete a paper version of this questionnaire, please use a BLACK or BLUE pen rather than a pencil.

Thank you very much for your time

1. Please list the criteria that should be used for your preferred operational definition of centralization as a physical sign and related symptom response groups.

Testing procedure:

Symptom response groups based on the centralization phenomenon:

2. Please write any further issues that do not fit into the context of the above question but are important when considering centralization in spinal clinical practice and related research.

3. Please list any important questions around centralization that you would like to be addressed in future research.

Please continue on a separate sheet of paper if needed and attach this to your form indicating the question that it relates to.

This is the end of the first-round questionnaire.

Please return your completed form no later than

Electronically by e-mail to:

A.Chorti@warwick.ac.uk

OR

By post using the enclosed pre-paid envelope to:

Angeliki Chorti
Warwick Medical School
University of Warwick
Coventry CV4 7AL
United Kingdom

We would like to thank you again for your time and considered opinions

Appendix 4.7 Delphi second-round questionnaire.

Toward a uniform definition for the centralization phenomenon: A Delphi study

Instructions

This is the second round questionnaire of the Delphi study. In this questionnaire, you are asked to rank your level of agreement with opinions from round one on a five-point scale (1 strongly agree – 5 strongly disagree) and provide written comments where appropriate. This form can be completed electronically or in paper format. To complete it electronically, please ***double click*** on the highlighted fields to mark or to type your answer for each statement. If completing the paper version, please ***put a cross*** in the box that is closest to how you feel. If you would like to complete the questionnaire in paper format, please contact the researcher using the contact details provided at the end of this questionnaire.

This questionnaire has derived from the content analysis and the collation by our research team of your responses in the first round of the Delphi study. Every effort has been made to be as inclusive as possible, however, if you feel that we have missed something, please add this in the section 'Please feel free to provide reasons for your choices and/ or add any important comments/ questions here:'. If you wish so, you can also provide reasons for your choices in the same section.

Your answers will remain strictly confidential. In order to ensure that your opinions remain anonymous, we would be grateful if your comments do not bear any personal identifiers and you do not talk about your answers with anyone.

Please return the completed form by: **[date]** If you have any difficulties or questions about this questionnaire, please feel free to contact Angeliki Chorti (**A.Chorti@warwick.ac.uk**, 0044(0)2476150405).

Thank you for your time and considered opinions.

Section 1: General definition

*Please **mark one box** for each statement that is closest to how you feel.*

(1)= Strongly Agree (2) = Agree (3) = Neutral (4) = Disagree (5) = Strongly Disagree

Centralization should generally be defined as:

	(1)	(2)	(3)	(4)	(5)
1. A lasting improvement in patient status (abolishment of distal symptoms or improvement of signs) in response to a defined movement which can vary in terms of direction, degree, duration, repetition and position used	☐	☐	☐	☐	☐
2. A lasting abolishment or decrease in intensity of the most distal radicular symptoms and signs in response to repeated movements or static positioning, traction or a combination (traction and repeated movements).	☐	☐	☐	☐	☐
3. An abolishment of the most distal radiating symptoms in response to repeated movement testing.	☐	☐	☐	☐	☐
4. An improvement in location, intensity or frequency of symptoms in response to single or repeated movement testing or sustained postures	☐	☐	☐	☐	☐
5. An abolishment of peripheral symptoms in response to repeated movement testing and overpressure	☐	☐	☐	☐	☐
6. A reduction or abolishment of peripheral symptoms in response to repeated movement testing and patient or therapist overpressure	☐	☐	☐	☐	☐
7. The movement of symptoms in a proximal direction in response to repeated end-range movement testing only	☐	☐	☐	☐	☐
8. Distal symptoms moving proximally in response to repeated movement testing and/or sustained positions or therapist mobilization.	☐	☐	☐	☐	☐
9. Distal symptoms moving and remaining proximally in response to repeated movement testing and/or sustained positions or therapist mobilization. In patients with axial symptoms only, the decrease in such symptoms is defined as 'centralization' if such decrease is (a) substantial (>50% reduction) and (b) sustained	☐	☐	☐	☐	☐
10. Distal symptoms (pain, numbness or tingling) traveling proximally towards the central spine in response to therapeutic loading strategies	☐	☐	☐	☐	☐

	(1)	(2)	(3)	(4)	(5)
11. An improvement of the most distal symptom regardless of the testing methods used (i.e. during movement testing, treatment, over time). In order of priority, the type (i.e. pain, paresthesia or anesthesia), location and intensity of symptom are considered in the hierarchy of improvement in the definition.	☐	☐	☐	☐	☐
12. Movement of pain only to a proximal location in response to movement testing	☐	☐	☐	☐	☐
13. The most distal pain disappearing and moving proximally in response to repeated end-range movements or static loading	☐	☐	☐	☐	☐
14. The progressive and stable reduction of the most distal pain towards the spinal midline in response to standardized repeated end-range movement or sustained loading testing procedures. Testing may involve multiple directions and various starting positions, progression of forces or alternative forces (e.g. clinician assistance) as well as more than one visit	☐	☐	☐	☐	☐

Please feel free to provide reasons for your choices and/ or add any important comments/ questions here:

Section 2: Operational criteria

*Please **mark one box** for each statement that is closest to how you feel.*

(1)= Strongly Agree (2) = Agree (3) = Neutral (4) = Disagree (5) = Strongly Disagree

A. Population for whom the test is intended

	(1)	(2)	(3)	(4)	(5)
15. Centralization can be best appreciated only in patients who demonstrate referred or radiating symptoms originating from the spine	☐	☐	☐	☐	☐
16. Centralization can be best appreciated only in patients with low back pain (i.e. and not patients with symptoms originating from other spinal areas)	☐	☐	☐	☐	☐
17. Before testing, patients should be screened for the presence of red flags (i.e. serious pathology indicators)	☐	☐	☐	☐	☐
18. Before testing, patients should be screened for the presence of yellow flags (e.g. Waddell signs)	☐	☐	☐	☐	☐
19. Patients with distal symptoms above the knee that meet other parts of a clinical prediction rule for success with spinal manipulation are most likely to benefit from manipulation regardless of their response to repeated movement testing	☐	☐	☐	☐	☐
20. The centralization phenomenon becomes complex in chronic cases with an adherent nerve root involved. In such cases, the close monitoring of the symptoms and potential postural deformities is required and patients are treated on an individual basis	☐	☐	☐	☐	☐

Please feel free to provide reasons for your choices and/ or add any important comments/ questions here:

B. Potential test users (examiners)

	(1)	(2)	(3)	(4)	(5)
21. Centralization should be recognized by the average clinician rather than requiring years of training	☐	☐	☐	☐	☐
22. Training and experience of examiners is essential in order to have consistent results among clinicians	☐	☐	☐	☐	☐

Please feel free to provide reasons for your choices and/ or add any important comments/ questions here:

C. Tools used for documentation

	(1)	(2)	(3)	(4)	(5)
23.The pattern of questioning and documenting patient responses i.e. pain intensity and location should be highly standardised	☐	☐	☐	☐	☐
24. For changes in pain location, a clear overlay numeric template should be used	☐	☐	☐	☐	☐
25. The area of the most distal symptoms should be marked and reassessed using a measurement from a bony landmark. The determination of the most distal pain should also be confirmed through palpation	☐	☐	☐	☐	☐

Please feel free to provide reasons for your choices and/ or add any important comments/ questions here:

D. Loading strategy – Type

Centralization can be elicited:

	(1)	(2)	(3)	(4)	(5)
26. By repeated end-range movement testing only	☐	☐	☐	☐	☐
27. By repeated and/ or sustained end-range movement testing	☐	☐	☐	☐	☐
28. By single and repeated movements. Sustained movements are used with caution to prevent confusion with instability	☐	☐	☐	☐	☐
29. By repeated and/or sustained end-range movement testing, or manual techniques executed by therapist if appropriate	☐	☐	☐	☐	☐
30. By repeated and/or sustained end-range movement testing, overpressure by patient or therapist or manual techniques executed by therapist if appropriate	☐	☐	☐	☐	☐
31. Regardless of testing methods used. This can involve test movements or after immediate treatment application, or post treatment over time	☐	☐	☐	☐	☐

Please feel free to provide reasons for your choices and/ or add any important comments/ questions here:

E. Loading strategy - Planes and directions

Testing for centralization should:

	(1)	(2)	(3)	(4)	(5)
32. Involve only the sagittal plane	☐	☐	☐	☐	☐
33. Involve the standard planes of movement available to the spinal region involved	☐	☐	☐	☐	☐
34. Involve the standard planes of movement available to the spinal region and/or a combination of movements if appropriate	☐	☐	☐	☐	☐

Please feel free to provide reasons for your choices and/ or add any important comments/ questions here:

F. Loading strategy - Movement testing

Movement testing for centralization should:

	(1)	(2)	(3)	(4)	(5)
35. Not be based solely on single movements. Test movements must be performed repeatedly, to the fullest of the patient's available end-range	☐	☐	☐	☐	☐
36. Not include more than 10 repetitions for each test movement	☐	☐	☐	☐	☐
37. Include repetitions that are continued long enough to ensure that the status of the symptoms has changed and/or is clear.	☐	☐	☐	☐	☐

	(1)	(2)	(3)	(4)	(5)
38. Not necessarily be performed in a specific order. The sequence of movement testing does not affect the outcome or the ability to detect centralization	☐	☐	☐	☐	☐

Please feel free to provide reasons for your choices and/ or add any important comments/ questions here:

G. Criteria for positive test

When defining the presence of centralization:

	(1)	(2)	(3)	(4)	(5)
39. Changes in the intensity of symptoms should not be considered	☐	☐	☐	☐	☐
40. Changes in the neurological status (symptoms or signs) should not be considered	☐	☐	☐	☐	☐
41. Nonspinal conditions or underlying disease states potentially causing or influencing symptoms should be excluded or taken into consideration	☐	☐	☐	☐	☐
42. The reduction of peripheral symptoms towards the spinal midline should be progressive and stable. Distal pain which simply abolishes rather than progressively receding toward the spine should not be considered in the definition	☐	☐	☐	☐	☐
43. Observed changes should be retained over time (lasting change)	☐	☐	☐	☐	☐

Please feel free to provide reasons for your choices and/ or add any important comments/ questions here:

H. Testing – Timeframe

When testing for centralization:

	(1)	(2)	(3)	(4)	(5)
44. The response to testing may be obvious during the first examination or may require evaluation over a period of time to confirm the phenomenon	☐	☐	☐	☐	☐
45. Symptoms may be sequentially and lastingly abolished at each session (pure response) or gradually abolished in a progressive manner but not sequentially (partial response)	☐	☐	☐	☐	☐

Please feel free to provide reasons for your choices and/ or add any important comments/ questions here:

I. Safety issues

	(1)	(2)	(3)	(4)	(5)
46. The patient's status (improving or worsening) should be considered beyond symptom relief i.e. neurological picture, range of motion. For example, if the patient's symptoms are improving but his neurological status is compromised, centralization should be considered an undesirable outcome for this patient.	☐	☐	☐	☐	☐
47. Symptoms getting worse proximally but improving distally should be considered a positive sign	☐	☐	☐	☐	☐
48. Education of patients is essential following the use of these techniques so that movement is normalized as symptoms settle and patients do not develop fear of movement	☐	☐	☐	☐	☐

Please feel free to provide reasons for your choices and/ or add any important comments/ questions here:

Section 3: Other issues related to centralization

*Please **mark one box** for each statement that is closest to how you feel.*

(1)= Strongly Agree (2) = Agree (3) = Neutral (4) = Disagree (5) = Strongly Disagree

	(1)	(2)	(3)	(4)	(5)
49. Centralizers have a good prognosis	☐	☐	☐	☐	☐
50. Centralization may be a stronger prognostic factor than psychosocial variables	☐	☐	☐	☐	☐
51. Centralizers have an internal disc disruption	☐	☐	☐	☐	☐

Please feel free to provide reasons for your choices and/ or add any important comments/ questions here:

Section 4: Future research

*Please **mark one box** for each future area and question that is closest to how you feel.*

(1)= Strongly Agree (2) = Agree (3) = Neutral (4) = Disagree (5) = Strongly Disagree

Future research should look at:

	(1)	(2)	(3)	(4)	(5)
52. Operational definition for centralization:	☐	☐	☐	☐	☐
53. Head to head comparisons of different operational definitions	☐	☐	☐	☐	☐
54. Clarification of the term centralization i.e. number of repetitions, type of change, magnitude of change (e.g. belt line pain), duration and timeframe of required change, necessity of provocative testing for central symptoms	☐	☐	☐	☐	☐
55. Conceptual model for centralization:	☐	☐	☐	☐	☐
56. Is centralization an anatomical phenomenon? Centralization in relation to tissue response	☐	☐	☐	☐	☐
57. Mechanism causing the centralization phenomenon (including physiological mechanism)	☐	☐	☐	☐	☐
58. Diagnostic accuracy of centralization:	☐	☐	☐	☐	☐
59. Criterion validity of centralization (e.g. using provocation discography as a standard; centralization as a tool for assessing the severity of a disc lesion; centralization as a tool for ruling out pathologies other than discogenic pain)	☐	☐	☐	☐	☐
60. Cross validation of diagnostic accuracy in different patient populations and examiners	☐	☐	☐	☐	☐
61. Potential irreducible derangement i.e. pain moving centrally but neurological symptoms moving distally: at what point should patients be referred for further investigation?	☐	☐	☐	☐	☐
62. Centralization and prevalence:	☐	☐	☐	☐	☐
63. Expected rates of centralization in clinical settings for acute and chronic patients with LBP	☐	☐	☐	☐	☐
64. Effect of training and procedures in the prevalence and outcomes of centralizers	☐	☐	☐	☐	☐
65. Centralization and course/ prognosis:	☐	☐	☐	☐	☐
66. Do centralizers have a favourable course when left untreated (natural history)?	☐	☐)	☐	☐	☐
67. Clinical response versus natural course of centralizers	☐	☐	☐	☐	☐

	(1)	(2)	(3)	(4)	(5)
68. Prescriptive validity of centralization:	☐	☐	☐	☐	☐
69. Do centralizers benefit from interventions other than directional movement exercise and which intervention provides the greatest benefit?	☐	☐	☐	☐	☐
70. Comparison of outcomes of non-responders to centralization for various interventions	☐	☐	☐	☐	☐
71. Centralization and outcomes (i.e. ability to return to work, psychosocial outcomes, economic outcomes, health care utilization, QoL, recurrences and ability to self-manage ; Cost effectiveness of the McKenzie assessment prior to disc surgery, injections, imaging or any other treatments)	☐	☐	☐	☐	☐
72. Centralization and subgroups:	☐	☐	☐	☐	☐
73. Clinical predictors (CPR) of patients responding (or not) to directional preference exercises (including: Effect of patient compliance and attitudes of centralizers on outcomes; Effect of severe disability and psychosocial distress on the diagnostic, prognostic and treatment characteristics of centralizers; Centralization versus other prognostic factors)	☐	☐	☐	☐	☐
74. Are centralizers and other clinical subgroups (e.g. patients satisfying CPR for manipulation or trunk stabilization treatment or candidates for disc surgery) made up of the same patients? How often do centralization findings co-exist with other findings? (i.e. relationship of centralization with other variables e.g. psychosocial, clinical findings e.g. segmental provocation signs, sciatica and EMG findings). How soon will positive EMGs become normal after centralization has been achieved and maintained?	☐	☐	☐	☐	☐
75. Centralization and the cervical spine:	☐	☐	☐	☐	☐
76. Centralization and contained cervical disc pathology	☐	☐	☐	☐	☐
77. Reliability of detecting, prevalence and outcomes in neck pain	☐	☐	☐	☐	☐
78. Predictive validity of centralization in the management of neck pain	☐	☐	☐	☐	☐
79. Reproducibility:	☐	☐	☐	☐	☐
80. Stability/ reversibility of the centralization phenomenon	☐	☐	☐	☐	☐
81. Effect of the clinician, patient, standardization and knowledge of the test on the reliability of centralization	☐	☐	☐	☐	☐
82. Role of history and examiner's training in predicting the presence of centralization	☐	☐	☐	☐	☐
83. Case studies:	☐	☐	☐	☐	☐
84. The rare patient whose pain centralizes up to the lumbar spine but then remains unchanged and may worsen with exercise. Generally improves over time and is usually initiated by long sitting time	☐	☐	☐	☐	☐
85. Education and training:	☐	☐	☐	☐	☐
86. How do we best educate healthcare professionals that abolishment of leg or arm symptoms in a patient with central or foraminal stenosis is not a "centralizer", but a separate subgroup of their own?	☐	☐	☐	☐	☐

Please feel free to provide reasons for your choices and/ or add any important comments/ questions here:

This is the end of the second round Delphi questionnaire.

Please make sure you have selected only one answer for each question.

Thank-you for taking the time to complete this questionnaire

Please return the completed questionnaire to: Angeliki Chorti, Warwick Medical School, University of Warwick, Gibbet Hill Road, Coventry, CV4 7AL

Electronically to: A.Chorti@warwick.ac.uk

By: [date]

If you wish to complete the questionnaire in paper format, please contact **Ms Angeliki Chorti** using the contact details provided above. A hard version will be sent to you, along with a return envelope. Alternatively, please feel free to print and send back. Again, a return envelope will happily be provided.

Emailed questionnaires will be treated anonymously.

Study Number:
Participant Identification Number for this study:
Feedback report

Appendix 4.8 Delphi feedback report from round 2.

Toward a uniform definition for the centralization phenomenon: A Delphi study

This is the feedback report from the second round of the Delphi study. This report is divided into two parts. Part 1 refers to the general statistics of your ratings and includes your comments and Part 2 presents a graphical display of your responses. For any questions regarding this form, please feel free to contact Ms Angeliki Chorti (Email: **A.Chorti@warwick.ac.uk**, Tel: 0044(0)2476150405) who will be happy to provide you with further information.

Thank you for reading this.

PART 1: GENERAL STATISTICS AND COMMENTS

Please note that the smaller the value of the median, the greater the degree of agreement with each statement (due to scoring system in the second round: 1-strongly agree to 5- strongly disagree). N = number of valid responses; % agree = percentage of participants that agreed with statement (collapsed categories: strongly agree + agree) green = ≥ 80% agreement; yellow = 70-79% agreement.

Section 1: General definition

Centralization should generally be defined as:

	N	% agree	Median (Quartiles)	Your rating in round 2
1. A lasting improvement in patient status (abolishment of distal symptoms or improvement of signs) in response to a defined movement which can vary in terms of direction, degree, duration, repetition and position used	**29**	**72.4**	**2.00 (1.00-3.00)**	
2. A lasting abolishment or decrease in intensity of the most distal radicular symptoms and signs in response to repeated movements or static positioning, traction or a combination (traction and repeated movements).	**29**	**55.2**	**2.00 (2.00-4.00)**	
3. An abolishment of the most distal radiating symptoms in response to repeated movement testing.	**29**	**48.3**	**3.00 (2.00-3.00)**	
4. An improvement in location, intensity or frequency of symptoms in response to single or repeated movement testing or sustained postures	**29**	**55.2**	**3.00 (3.00-4.00)**	
5. An abolishment of peripheral symptoms in response to repeated movement testing and overpressure	**29**	**44.8**	**3.00 (2.00-3.00)**	
6. A reduction or abolishment of peripheral symptoms in response to repeated movement testing and patient or therapist overpressure	**29**	**34.5**	**3.00 (2.00-3.00)**	
7. The movement of symptoms in a proximal direction in response to repeated end-range movement testing only	**29**	**27.6**	**3.00 (2.00-4.00)**	
8. Distal symptoms moving proximally in response to repeated movement testing and/or sustained positions or therapist mobilization.	**29**	**65.5**	**2.00 (1.50-3.00)**	
9. Distal symptoms moving and remaining proximally in response to repeated movement testing and/or sustained positions or therapist mobilization. In patients with axial symptoms only, the decrease in such symptoms is defined as 'centralization' if such decrease is (a) substantial (>50% reduction) and (b) sustained	**29**	**44.8**	**3.00 (1.50-4.00)**	
10. Distal symptoms (pain, numbness or tingling) traveling proximally towards the central spine in response to therapeutic loading strategies	**29**	**58.6**	**2.00 (2.00-3.00)**	
11. An improvement of the most distal symptom regardless of the testing methods used (i.e. during movement testing, treatment, over time). In order of priority, the type (i.e. pain, paresthesia or anesthesia), location and intensity of symptom are considered in the hierarchy of improvement in the definition.	**29**	**37.9**	**3.00 (2.00-4.00)**	
12. Movement of pain only to a proximal location in response to movement testing	**29**	**20.7**	**3.00 (3.00-4.00)**	
13. The most distal pain disappearing and moving proximally in response to repeated end-range movements or static loading	**29**	**44.8**	**3.00 (2.00-3.50)**	
14. The progressive and stable reduction of the most distal pain towards the spinal midline in response to standardized repeated end-range movement or sustained loading testing procedures. Testing may involve multiple directions and various starting positions, progression of forces or alternative forces (e.g. clinician assistance) as well as more than one visit	**29**	**79.3**	**2.00 (1.00-2.00)**	

Comments:

- I consider this phenomenon only as regards to pain; Pain most usually centralizes rather than other types of symptoms.
- Centralization should focus on end-range patient movements and their effect on symptoms, including pain. However, it should not be limited to pain.
- Some definitions demonstrate a failure to differentiate between signs and symptoms.
- This sign should appear only in response to repeated movements or sustained positioning at end range
- Definitions of centralization should not involve static postures or therapist overpressure because this confounds with other competing diagnoses.
- The definition needs to state that the symptoms (usually pain) recedes proximally toward the midline of the spine
- The words 'movement' and 'moving' should be used with caution because to those unfamiliar with the concept it may be misunderstood as the creation of new symptoms. Clarification that the more proximal pain is present from the start should be made.
- There is both a process of centralization and an end-point defined as ''centralized symptoms''; Lasting/persisting improvement is yet another characteristic of the underlying pain generator that also affects patients' prognosis but is not required for pain to centralize. Lasting improvement requires a stability of the centralizing change and should not be required in the definition of centralization
- Centralization is a dynamic phenomenon: symptoms are sequentially going to the spine, abolishing from their more peripheral position they are usually not produced in every part of the limb (for example a patient who has leg pain at the beginning and have back pain only at the end... and the pain has not gone to the thigh).
- Cardinal points are: it applies to symptoms, it is the effect of repeated movements or static loading (both at end range), it remains after, is a phenomenon that appears usually quickly
- Definition in statement 2 specifies radicular pain thereby excluding somatic referred pain; Depends on definition of traction. Manual traction can be part of the therapist intervention, but motorized traction should not be included.
- Definitions in statements 3, 8, 13 do not address lasting change
- Definition in statement 4: Single may not be sufficient to evaluate
- Definition in statements 5, 6: Does this mean that overpressure is required?

- Definition in statement 7: This leaves out sustained positioning
- Definition in statements 9, 10: Centralization does not occur with central symptoms only
- Definition in statement 12: Is pain the only symptom?

Section 2: Operational criteria

A. Population for whom the test is intended

	N	% agree	Median (Quartiles)	Your rating in round 2
15. Centralization can be best appreciated only in patients who demonstrate referred or radiating symptoms originating from the spine	**29**	**79.3**	**1.00 (1.00-2.00)**	
16. Centralization can be best appreciated only in patients with low back pain i.e. and not patients with symptoms originating from other spinal areas	**29**	**17.2**	**5.00 (4.00-5.00)**	
17. Before testing, patients should be screened for the presence of red flags (i.e. serious pathology indicators)	**28**	**85.7**	**1.00 (1.00-1.75)**	
18. Before testing, patients should be screened for the presence of yellow flags (e.g. Waddell signs)	**29**	**31.0**	**3.00 (2.00-4.00)**	
19. Patients with distal symptoms above the knee that meet other parts of a clinical prediction rule for success with spinal manipulation are most likely to benefit from manipulation regardless of their response to repeated movement testing	**29**	**13.8**	**4.00 (3.00-4.50)**	
20. The centralization phenomenon becomes complex in chronic cases with an adherent nerve root involved. In such cases, the close monitoring of the symptoms and potential postural deformities is required and patients are treated on an individual basis	**29**	**48.3**	**3.00 (2.00-4.50)**	

Comments:

- Statement 15: Radicular pain does not centralize- somatic referred pain does.
- Statement 16: Though anecdotally reported, not aware of any peer review studies suggesting centralization in other areas of the spine
- Statement 17: Screening for red flags is mandatory to all patients prior to the initiation of any testing or treatment; Screening for red flags should be done on all patients, but this does not preclude some provisional testing for centralization, though a referral is always warranted; Absence of centralization supports the need for further investigation.
- Statement 18: Yellow flags should be considered on all patients, but this does not affect the desirability of their assessment for centralization; Waddell's testing should be performed only after a thorough and unbiased clinical exam including testing for centralization. Waddell's signs may improve when the patient can effectively control pain and gain a sense of hope over their condition.
- Statement 19: This is unknown; The clinical prediction rule was developed in comparison to stabilization exercise; A description of the clinical prediction rule for success is not provided; Most will respond to manipulation (Childs et al.) but unknown how patients' response to repeated movement testing will affect this.
- Statement 20: This applies to all patients, not just those tested for centralization; Chronic cases are usually complex regardless of the assessment done; Complexity is reduced if CP is found; Little utility in evidence based clinical decision making since there is not a method that has been validated in the medical literature to reliably identify patients with an adherent nerve root

B. Potential test users (examiners)

	N	% agree	Median (Quartiles)	Your rating in round 2
21. Centralization should be recognized by the average clinician rather than requiring years of training	**27**	**81.5**	**1.00 (1.00-2.00)**	
22. Training and experience of examiners is essential in order to have consistent results among clinicians	**27**	**63.0**	**2.00 (1.00-3.00)**	

Comments:

- Statement 21: The literature supports this statement; Training is essential to improve inter-examiner reliability; The average clinician should recognize centralization when it occurs but will not have the skills to elicit centralization in many cases; Should not require extensive training to recognize once a suitable definition is determined; The average clinician should be trained in identifying CP at the undergraduate level, just as assessment for evidence of a neurologic deficit should be part of basic education; The average clinician should have training and experience. This is not specialized knowledge or only gained through courses in one method; Can not answer question 21 unless "average" and "years" are defined; Some training and clinical practice is required as with any of our skills, this is best available in continuing education courses; What is essential is the time "practicing" the skill in the clinic. Good instruction only starts the process
- Statement 22: Statement is too open ended; Training should be separated from experience; Training is necessary for consistency but experience is not necessary; There is evidence of no significant difference for years and between practitioners and students; Training and experience is less important than clear and agreed upon definitions and reduction of complexity for consistent results among clinicians; The literature supports these choices ; See Kilby et al 1990

C. Tools used for documentation

	N	% agree	Median (Quartiles)	Your rating in round 2
23.The pattern of questioning and documenting patient responses i.e. pain intensity and location should be highly standardised	**29**	**82.8**	**2.00 (1.00-2.00)**	
24. For changes in pain location, a clear overlay numeric template should be used	**29**	**44.8**	**3.00 (2.00-3.00)**	
25. The area of the most distal symptoms should be marked and reassessed using a measurement from a bony landmark. The determination of the most distal pain should also be confirmed through palpation	**29**	**13.8**	**4.00 (4.00-5.00)**	

Comments:

- Statement 23: Standardization is essential to ensure reliability, but statements 24 and 25 are unnecessary; This may apply to research protocols but not clinical practice; Not sure what you mean by 'highly', but standardization helps
- Statement 24: Template for research purposes but not required for clinical assessment; For changes in pain location, a categorical response option should be given (e.g. foot, below knee/not foot, etc.); Perhaps in research protocols but not routine clinical practice
- Statement 25: Precise measurement not required in clinical exam, only in research; Palpation cannot determine the location of the most distal pain; Palpation is unreliable and invalid; This would be time consuming; Unclear; Is it spinal or peripheral palpation?; The underlying structures, not being involved, shouldn't be painful; Palpation has never been documented as part of the centralization testing

D. Loading strategy – Type

Centralization can be elicited:

	N	% agree	Median (Quartiles)	Your rating in round 2
26. By repeated end-range movement testing only	**29**	**13.8**	**4.00 (3.00-5.00)**	
27. By repeated and/ or sustained end-range movement testing	**28**	**78.6**	**2.00 (1.00-2.00)**	
28. By single and repeated movements. Sustained movements are used with caution to prevent confusion with instability	**29**	**27.6**	**4.00 (3.00-5.00)**	
29. By repeated and/or sustained end-range movement testing, or manual techniques executed by therapist if appropriate	**29**	**89.7**	**2.00 (1.00-2.00)**	
30. By repeated and/or sustained end-range movement testing, overpressure by patient or therapist or manual techniques executed by therapist if appropriate	**29**	**82.8**	**1.00 (1.00-2.00)**	
31. Regardless of testing methods used. This can involve test movements or after immediate treatment application, or post treatment over time	**28**	**28.6**	**3.00 (2.00-4.00)**	

Comments

- Important not only what improves status but also the opposite should worsen status. One must differentiate between decrease in symptoms v. improvement of status e.g. stenosis where flexion may decrease symptoms and extension worsen status but on resumption of ambulation the symptoms recur thus not improved status
- Centralization should be mutually exclusive from sustained and therapist overpressure
- Statement 26: Leaves out other interventions
- Statement 27: Single movements don't tell much and I don't understand what is meant by instability
- Statement 29: Disagree with manipulation, but agree with overpressure or manual traction with the cervical spine
- Statement 30: Manual techniques i.e. manipulation/mobilization are not included in the definition
- Statement 31: This statement is unclear; Depends on purpose of the testing i.e. initial decision making or determine response to treatment. E.g. I use repeated movements during initial examination to test, but would look upon a centralization response after treatment as favourable, regardless of the treatment used.

E. Loading strategy - Planes and directions

Testing for centralization should:

	N	% agree	Median (Quartiles)	Your rating in round 2
32. Involve only the sagittal plane	**29**	**13.8**	**5.00 (4.00-5.00)**	
33. Involve the standard planes of movement available to the spinal region involved	**29**	**55.2**	**2.00 (2.00-3.50)**	
34. Involve the standard planes of movement available to the spinal region and/or a combination of movements if appropriate	**29**	**79.3**	**1.00 (1.00-2.00)**	

Comments:

- Statement 32: If not elicited with sagittal plane motions the patient may be served better with a different approach; Non-sagittal testing is not required if centralization is elicited with sagittal plane testing; Sagittal is primary, but may need to pivot pelvis asymmetrically
- Statements 33, 34: Confusing statement
- Statement 34: A full examination often only requires standard planes of motion, but combinations of planes are required for maybe 30% of cases in order to identify the CP

F. Loading strategy - Movement testing

Movement testing for centralization should:

	N	% agree	Median (Quartiles)	Your rating in round 2
35. Not be based solely on single movements. Test movements must be performed repeatedly, to the fullest of the patient's available end-range	**29**	**75.9**	**2.00 (1.00-2.50)**	
36. Not include more than 10 repetitions for each test movement	**29**	**31.0**	**4.00 (2.00-5.00)**	
37. Include repetitions that are continued long enough to ensure that the status of the symptoms has changed and/or is clear.	**29**	**75.9**	**2.00 (1.00-2.50)**	
38. Not necessarily be performed in a specific order. The sequence of movement testing does not affect the outcome or the ability to detect centralization	**29**	**37.9**	**3.00 (2.00-4.00)**	

Comments:

- Statement 35: True, but ignores static positioning; Repeated movements may not be necessary, but I can't think of an example
- Statement 36: May require several sets of exercises and more than one treatment session; Some cases with relatively minor pain will require many repetitions 30-50 even. If clear evidence of CP is found with 5-10 movements and it is a reversible phenomenon, then further testing is not required. The number of repetitions or the duration of sustained loading is chosen on a case by case basis depending on the pain response, severity of pain, age and general condition of the patient, patient willingness etc
- Statement 38: It might, if one has aggravated the symptoms; It is not necessary to perform tests in a specific order to assess centralization. The order of testing can, however influence the results; Changing the order of testing can be done for a number of reasons based on a clinical reasoning process. Using clues from the history may minimize the testing needed. If symptoms are vague or minimal provocative testing may be needed so that deduction/centralization can be better observed. Provocative testing can be skipped when the history is very clear or if the patient is very acute or pain is high

G. Criteria for positive test

When defining the presence of centralization:

	N	% agree	Median (Quartiles)	Your rating in round 2
39. Changes in the intensity of symptoms should not be considered	**29**	**24.1**	**4.00 (3.50-5.00)**	
40. Changes in the neurological status (symptoms or signs) should not be considered	**29**	**13.8**	**5.00 (3.50-5.00)**	
41. Nonspinal conditions or underlying disease states potentially causing or influencing symptoms should be excluded or taken into consideration	**28**	**82.1**	**2.00 (1.00-2.00)**	
42. The reduction of peripheral symptoms towards the spinal midline should be progressive and stable. Distal pain which simply abolishes rather than progressively receding toward the spine should not be considered in the definition	**29**	**34.5**	**3.00 (2.00-4.00)**	
43. Observed changes should be retained over time (lasting change)	**29**	**75.9**	**1.00 (1.00-2.50)**	

Comments:

- Statement 39: Intensity of pain is an important component, not just pain location; Changes in intensity are the starting point indicating when to continue testing. It might be too early to use the term "centralization" YET, but keep testing because a decrease in intensity is commonly seen before the symptom actually abolishes or moves to a more proximal location.
- Statement 40: Neurologic status should always be monitored. Occasionally a reflex will return or disappear during testing and this is important
- Statement 41:: Unclear context; One should always monitor and consider co-morbidities; Centralization is an indication related to mechanical spine issues, not disease; More than one factors in statement; Underlying disease states should be excluded, and non-spinal conditions should be taken into account
- Statement 42: Basically correct but still debatable to some extent; Stability is required only for a good prognosis but should not be required in determining the presence of centralization; Instability may be underlying but does not necessarily contraindicate in the acute stage; That's what McKenzie says; Ideal if changes are preserved over time, but looking for longer periods of centralisation and stabilisation of the condition are also important; Reference supporting this statement is Young S, et al. Correlation of physical examination characteristics with three sources of chronic low back pain. The Spine Journal 2003; 3: 460-465
- Statement 43: CP is primarily a spinal (disc) phenomenon (Laslett M et al Eur Spine J 2006); Even transient evidence of centralization is useful in the determination of motion preference; Retained assuming that the patient has been compliant and avoid the provocative posture or movements

H. Testing – Timeframe

When testing for centralization:

	N	% agree	Median (Quartiles)	Your rating in round 2
44. The response to testing may be obvious during the first examination or may require evaluation over a period of time to confirm the phenomenon	**28**	**82.1**	**1.50 (1.00-2.00)**	
45. Symptoms may be sequentially and lastingly abolished at each session (pure response) or gradually abolished in a progressive manner but not sequentially (partial response)	**27**	**70.4**	**2.00 (2.00-3.00)**	

Comments:

- Statement 44: Should occur during initial exam but not over time; Need to define "time" 3 days of testing OK, not three weeks.
- Statement 45: Unclear statement; The meaning of sequentially is not clear. Is this temporal or spatial or type of symptoms or distalness? "Lasting" abolition is not a requirement to identify centralization.

I. Safety issues

	N	% agree	Median (Quartiles)	Your rating in round 2
46. The patient's status (improving or worsening) should be considered beyond symptom relief i.e. neurological picture, range of motion. For example, if the patient's symptoms are improving but his neurological status is compromised, centralization should be considered an undesirable outcome for this patient.	**28**	**64.3**	**2.00 (1.00-3.00)**	
47. Symptoms getting worse proximally but improving distally should be considered a positive sign	**28**	**89.3**	**2.00 (1.00-2.00)**	
48. Education of patients is essential following the use of these techniques so that movement is normalized as symptoms settle and patients do not develop fear of movement	**28**	**85.7**	**1.00 (1.00-2.00)**	

Comments:

- Statement 46: Confusing statement; Centralization is never undesirable, however worsening of neurological status would be considered failure to centralize. Neurological signs should be given priority over pain response; Centralization is usually accompanied by improving neurological status. It has been shown that ROM improves with centralization not the opposite; Some patients have neurological signs that do not improve as pain centralizes but as long as the neurological status is not worse, centralization is desirable. I have never seen a patient who centralized and remained better whose neurological status deteriorated; If neurological status is compromised at the beginning of the exam, then improving symptoms i.e. centralization should be pursued as a favourable sign; There is no magic in centralization. It is best not considered as a separate entity, but is simply a means of assessing change in symptoms which is appropriately included as part of an overall framework or hierarchy of information used in assessment of symptoms
- Statement 47: This is usually correct, but only if the peripheral pain rapidly abolishes and the increased proximal pain is not too severe or prolonged more than a few days; There is usually an increase in proximal pain during centralization. However, patients who have chemically sensitized discs can have partial centralization with tremendous increases in proximal pain. Their pain does not remain centralized when the spine is loaded
- Statement 48: This is a true statement, but what does it have to do with centralization; This is a question likely to elicit false positive responses as who could possibly disagree with the need for patient education. To say that it is essential in this situation more than others or for all patients, however I don't think is warranted or necessary

Section 3: Other issues related to centralization

	N	% agree	Median (Quartiles)	Your rating in round 2
49. Centralizers have a good prognosis	**29**	**100**	**1.00 (1.00-1.50)**	
50. Centralization may be a stronger prognostic factor than psychosocial variables	**29**	**72.4**	**2.00 (1.00-3.00)**	
51. Centralizers have an internal disc disruption	**29**	**41.4**	**3.00 (2.00-4.00)**	

Comments:

- Statement 49: Research is clear on this statement
- Statement 50: Unknown at this time; I am not sure that it has been studied to the degree necessary to be able to make such a statement; Research still in dispute for this statement
- Statement 51: At present, it appears to be so, but more evidence is needed; This is correct in patients who are not severely disabled (Roland Morris category) or distressed (Distress Risk Assessment Method) see Laslett M et al Eur Spine J 2006. All undistressed and minimally disabled patients who centralize satisfy ISIS criteria for internal disc disruption; Grade 1-3 internal disc disruption (using the Dallas Disco gram scale) will centralize. Grade 4 will not; Many centralizers have a herniated disc and sciatica. Many others likely have an internal disc problem.

Section 4: Future research

Future research should look at:

	N	% agree	Median (Quartiles)	Your rating in round 2
52. Operational definition for centralization:	**25**	**72.0**	**2.00 (1.00-3.00)**	
53. Head to head comparisons of different operational definitions	**29**	**58.6**	**2.00 (2.00-3.00)**	
54. Clarification of the term centralization i.e. number of repetitions, type of change, magnitude of change (e.g. belt line pain), duration and timeframe of required change, necessity of provocative testing for central symptoms	**29**	**55.2**	**2.00 (1.50-3.00)**	
55. Conceptual model for centralization:	**20**	**65.0**	**2.00 (1.00-3.00)**	
56. Is centralization an anatomical phenomenon? Centralization in relation to tissue response	**29**	**55.2**	**2.00 (2.00-4.00)**	
57. Mechanism causing the centralization phenomenon (including physiological mechanism)	**28**	**75.0**	**2.00 (1.25-2.75)**	
58. Diagnostic accuracy of centralization:	**20**	**85.0**	**2.00 (1.00-2.00)**	
59. Criterion validity of centralization (e.g. using provocation discography as a standard; centralization as a tool for assessing the severity of a disc lesion; centralization as a tool for ruling out pathologies other than discogenic pain)	**29**	**69.0**	**2.00 (1.00-3.00)**	
60. Cross validation of diagnostic accuracy in different patient populations and examiners	**29**	**75.9**	**2.00 (1.00-2.50)**	
61. Potential irreducible derangement i.e. pain moving centrally but neurological symptoms moving distally: at what point should patients be referred for further investigation?	**29**	**69.0**	**2.00 (1.00-3.00)**	
	N	**% agree**	**Median (Quartiles)**	**Your rating in round 2**

	N	% agree	Median (Quartiles)	Your rating in round 2
62. Centralization and prevalence:	**24**	**75.0**	**2.00 (1.00-2.75)**	
63. Expected rates of centralization in clinical settings for acute and chronic patients with LBP	**29**	**79.3**	**2.00 (1.00-2.00)**	
64. Effect of training and procedures in the prevalence and outcomes of centralizers	**29**	**89.7**	**2.00 (1.00-2.00)**	
65. Centralization and course/ prognosis:	**21**	**90.5**	**2.00 (1.00-2.00)**	
66. Do centralizers have a favourable course when left untreated (natural history)?	**29**	**89.7**	**2.00 (1.00-2.00)**	
67. Clinical response versus natural course of centralizers	**28**	**92.9**	**2.00 (1.00-2.00)**	
68. Prescriptive validity of centralization:	**19**	**89.5**	**1.00 (1.00-2.00)**	
69. Do centralizers benefit from interventions other than directional movement exercise and which intervention provides the greatest benefit?	**29**	**86.2**	**1.00 (1.00-2.00)**	
70. Comparison of outcomes of non-responders to centralization for various interventions	**29**	**93.1**	**1.00 (1.00-2.00)**	
71. Centralization and outcomes (i.e. ability to return to work, psychosocial outcomes, economic outcomes, health care utilization, QoL, recurrences and ability to self-manage ; Cost effectiveness of the McKenzie assessment prior to disc surgery, injections, imaging or any other treatments)	**29**	**86.2**	**1.00 (1.00-2.00)**	
72. Centralization and subgroups:	**19**	**89.5**	**1.00 (1.00-2.00)**	
73. Clinical predictors (CPR) of patients responding (or not) to directional preference exercises (including: Effect of patient compliance and attitudes of centralizers on outcomes; Effect of severe disability and psychosocial distress on the diagnostic, prognostic and treatment characteristics of centralizers; Centralization versus other prognostic factors)	**29**	**79.3**	**1.00 (1.00-2.00)**	
74. Are centralizers and other clinical subgroups (e.g. patients satisfying CPR for manipulation or trunk stabilization treatment or candidates for disc surgery) made up of the same patients? How often do centralization findings co-exist with other findings? (i.e. relationship of centralization with other variables e.g. psychosocial, clinical findings e.g. segmental provocation signs, sciatica and EMG findings). How soon will positive EMGs become normal after centralization has been achieved and maintained?	**29**	**82.8**	**2.00 (1.00-2.00)**	
75. Centralization and the cervical spine:	**21**	**90.5**	**1.00 (1.00-2.00)**	
76. Centralization and contained cervical disc pathology	**29**	**55.2**	**2.00 (1.50-3.50)**	
77. Reliability of detecting, prevalence and outcomes in neck pain	**29**	**93.1**	**1.00 (1.00-2.00)**	
78. Predictive validity of centralization in the management of neck pain	**29**	**93.1**	**1.00 (1.00-2.00)**	
79. Reproducibility:	**20**	**85.0**	**2.00 (1.00-2.00)**	
80. Stability/ reversibility of the centralization phenomenon	**29**	**72.4**	**2..00 (1.00-3.00)**	
81. Effect of the clinician, patient, standardization and knowledge of the test on the reliability of centralization	**29**	**65.5**	**2.00 (1.00-3.00)**	
82. Role of history and examiner's training in predicting the presence of centralization	**29**	**65.5**	**2.00 (2.00-3.00)**	
83. Case studies:	**18**	**55.6**	**2.00 (1.00-3.25)**	
84. The rare patient whose pain centralizes up to the lumbar spine but then remains unchanged and may worsen with exercise. Generally improves over time and is usually initiated by long sitting time	**28**	**35.7**	**3.00 (2.00-3.00)**	
85. Education and training:	**19**	**78.9**	**2.00 (1.00-2.00)**	
86. How do we best educate healthcare professionals that abolishment of leg or arm symptoms in a patient with central or foraminal stenosis is not a "centralizer", but a separate subgroup of their own?	**29**	**48.3**	**3.00 (1.00-3.00)**	

Comments:

- Statement 52: A new definition is unnecessary. McKenzie has provided an operational definition; Coming to some sort of agreement by this method will most likely produce a product on which none of us can agree whole heartedly; A good operational definition is not as critical to clinically managing patients as it is to performing research
- Statement 53: What would be the purpose?
- Statement 54: McKenzie has provided clarification
- Statement 60: Validation of the Laslett M et 2006 results is urgently required
- Statement 61: This is unlikely to occur. Patients whose pain will only peripheralise and not centralize have other things going on instead of primary discogenic pain (examples taken from my unpublished research data includes: gas filled disc, swollen dorsal root ganglion, end plate avulsion fractures)
- Statement 63: The CP and directional preference are very common in LBP cases (acute or chronic) and research is already available
- Statement 66, 67: How will you leave them untreated when they find out that they can positively affect their symptoms with simple movements?
- Statement 72: Too many variables
- Statement 77: Some data exist
- Statement 81: Not sure what your statement means
- Statement 82: Don't know what you mean. Predicting prior to examination?

- Statement 84: Is number 84 a question; This is not a rare patient in my experience – it's common for a centralizing patient to not improve until they control their posture sufficiently to avoid worsening between exercise sessions.
- Statement 86: This is a question which cannot be answered by the choices available; Centralization as the result of a single direction of repeated test movements (lumbar or cervical) is still centralization, regardless of the imaging findings of central or lateral stenosis. There are patients with a pseudoclaudication history whose pain centralizes with flexion or extension and they find their walking tolerance improves as a result. No one has data on this but it is a good area of research. Until that data is collected and analyzed, we can only provide education based on our anecdotal experience, but that does not include teaching that stenotics cannot centralize their pain

---

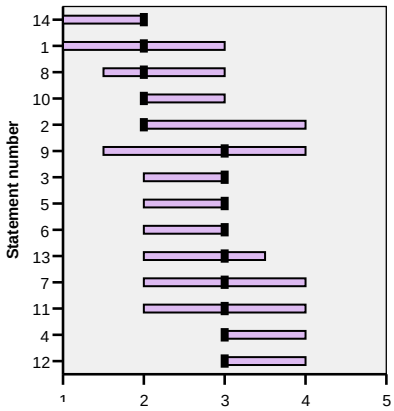
General comments:

- Some items contain multiple roots but only one statement of agreement or not. These should be carefully reviewed especially when the roots are conflicting
- The overall survey was long and somewhat burdensome

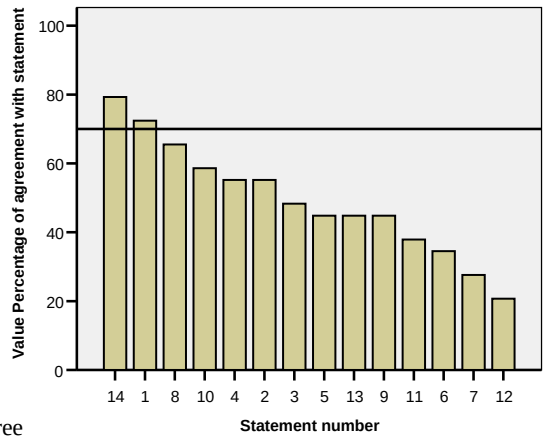
PART 2: GRAPHS

This part presents a graphical display of your responses. The statement numbers are presented in the graphs in order of importance (i.e. best to worse rating in round 2). The graphs on the left side (a) present the medians and interquartile ranges of agreement in each statement (1, strongly agree to 5, strongly disagree), whereas the graphs on the right hand side (b) refer to the percentage of agreement with each statement. The vertical line in graphs (b) represents the cut-off point of 70%.

Section 1: General definition



Strongly agree Strongly disagree



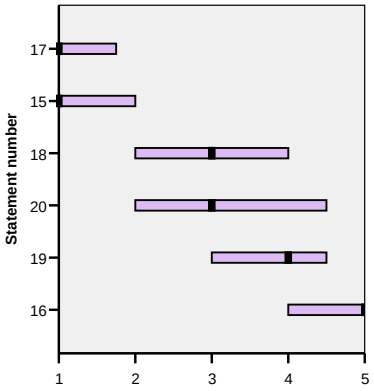
(b)

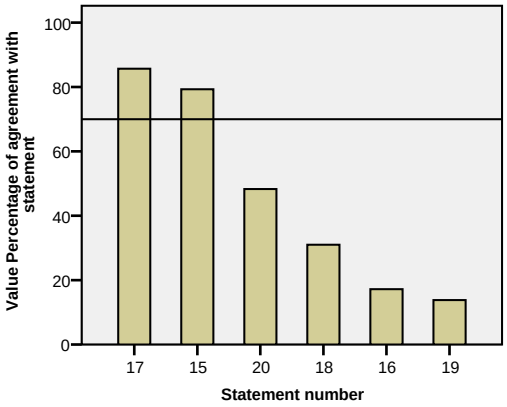
1. A lasting improvement in patient status (abolishment of distal symptoms or improvement of signs) in response to a defined movement which can vary in terms of direction, degree, duration, repetition and position used; 2. A lasting abolishment or decrease in intensity of the most distal radicular symptoms and signs in response to repeated movements or static positioning, traction or a combination (traction and repeated movements); 3. An abolishment of the most distal radiating symptoms in response to repeated movement testing; 4. An improvement in location, intensity or frequency of symptoms in response to single or repeated movement testing or sustained postures; 5. An abolishment of peripheral symptoms in response to repeated movement testing and overpressure; 6. A reduction or abolishment of peripheral symptoms in response to repeated movement testing and patient or therapist overpressure; 7. The movement of symptoms in a proximal direction in response to repeated end-range movement testing only; 8. Distal symptoms moving proximally in response to repeated movement testing and/or sustained positions or therapist mobilization; 9. Distal symptoms moving and remaining proximally in response to repeated movement testing and/or sustained positions or therapist mobilization. In patients with axial symptoms only, the decrease in such symptoms is defined as 'centralization' if such decrease is (a) substantial (>50% reduction) and (b) sustained; 10. Distal symptoms (pain, numbness or tingling) traveling proximally towards the central spine in response to therapeutic loading strategies; 11. An improvement of the most distal symptom regardless of the testing methods used (i.e. during movement testing, treatment, over time). In order of priority, the type (i.e. pain, paresthesia or anesthesia), location and intensity of symptom are considered in the hierarchy of improvement in the definition; 12. Movement of pain only to a proximal location in response to movement testing; 13. The most distal pain disappearing and moving proximally in response to repeated end-range movements or static loading; 14. The progressive and stable reduction of the most distal pain towards the spinal midline in response to standardized repeated end-range movement or sustained loading testing procedures. Testing may involve multiple directions and various starting positions, progression of forces or alternative forces (e.g. clinician assistance) as well as more than one visit

---

Section 2: Operational criteria

A. Population for whom the test is intended



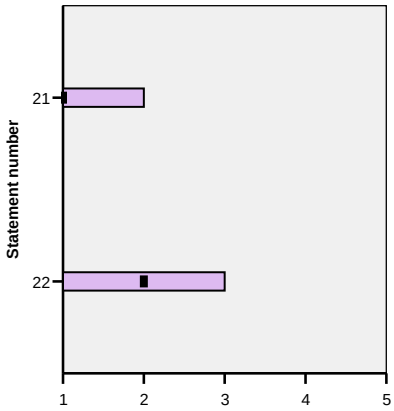


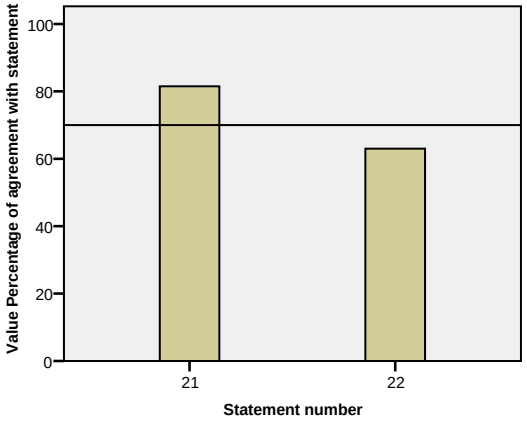
Strongly agree Strongly disagree

(b)

15. Centralization can be best appreciated only in patients who demonstrate referred or radiating symptoms originating from the spine; 16. Centralization can be best appreciated only in patients with low back pain i.e. and not patients with symptoms originating from other spinal areas
17. Before testing, patients should be screened for the presence of red flags (i.e. serious pathology indicators); 18. Before testing, patients should be screened for the presence of yellow flags (e.g. Waddell signs); 19. Patients with distal symptoms above the knee that meet other parts of a clinical prediction rule for success with spinal manipulation are most likely to benefit from manipulation regardless of their response to repeated movement testing; 20. The centralization phenomenon becomes complex in chronic cases with an adherent nerve root involved. In such cases, the close monitoring of the symptoms and potential postural deformities is required and patients are treated on an individual basis

B. Potential test users (examiners)



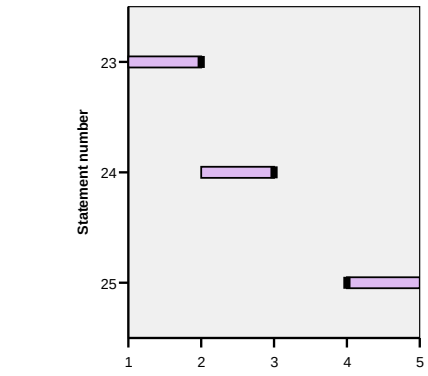


Strongly agree Strongly disagree

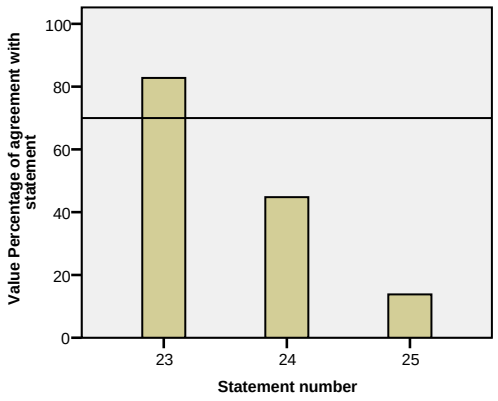
(b)

21. Centralization should be recognized by the average clinician rather than requiring years of training; 22. Training and experience of examiners is essential in order to have consistent results among clinicians

C. Tools used for documentation



Strongly agree



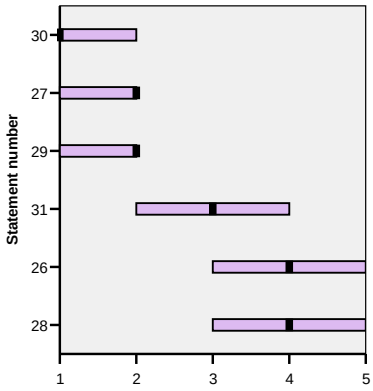
Strongly disagree

(b)

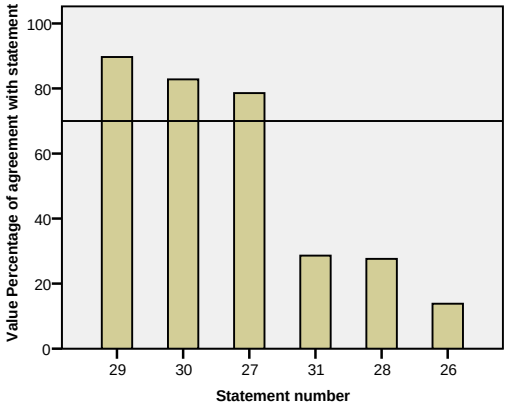
23. The pattern of questioning and documenting patient responses i.e. pain intensity and location should be highly standardised; 24. For changes in pain location, a clear overlay numeric template should be used; 25. The area of the most distal symptoms should be marked and reassessed using a measurement from a bony landmark. The determination of the most distal pain should also be confirmed through palpation

D. Loading strategy – Type

Centralization can be elicited:



Strongly agree



Strongly disagree

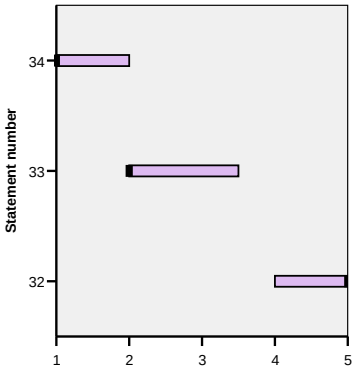
(a)

(b)

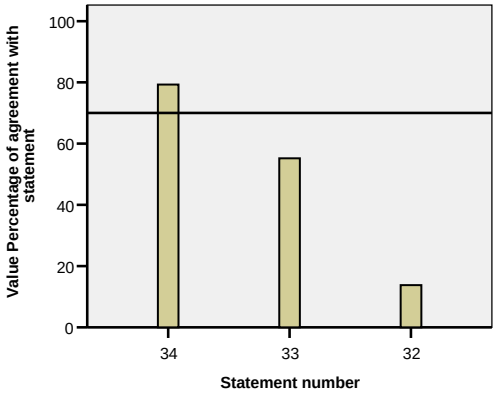
26. By repeated end-range movement testing only; 27. By repeated and/ or sustained end-range movement testing; 28. By single and repeated movements. Sustained movements are used with caution to prevent confusion with instability; 29. By repeated and/or sustained end-range movement testing, or manual techniques executed by therapist if appropriate; 30. By repeated and/or sustained end-range movement testing, overpressure by patient or therapist or manual techniques executed by therapist if appropriate; 31. Regardless of testing methods used. This can involve test movements or after immediate treatment application, or post treatment over time

E. Loading strategy - Planes and directions

Testing for centralization should:



Strongly agree



Strongly disagree

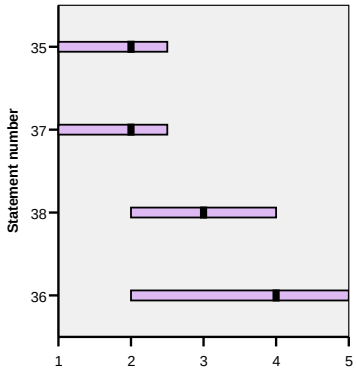
(a)

(b)

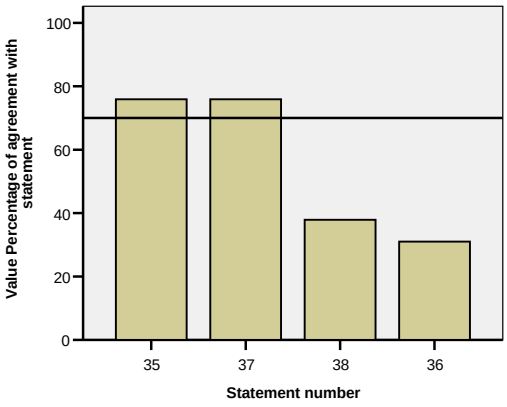
32. Involve only the sagittal plane; 33. Involve the standard planes of movement available to the spinal region involved; 34. Involve the standard planes of movement available to the spinal region and/or a combination of movements if appropriate

F. Loading strategy - Movement testing

Movement testing for centralization should:



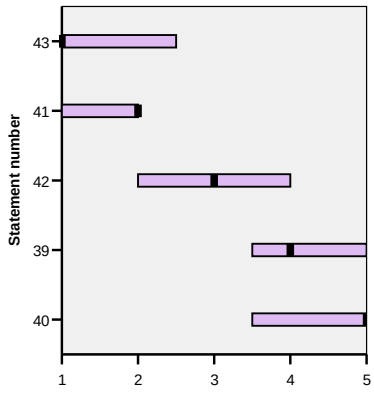
(a)



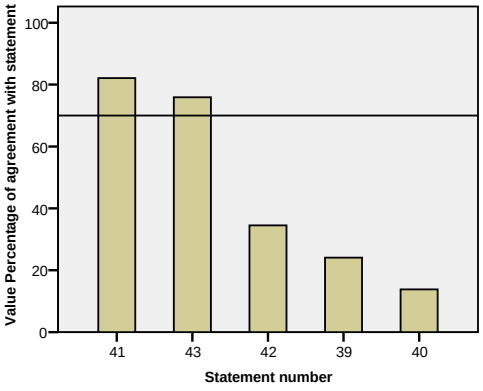
(b)

35. Not be based solely on single movements. Test movements must be performed repeatedly, to the fullest of the patient's available end-range; 36. Not include more than 10 repetitions for each test movement; 37. Include repetitions that are continued long enough to ensure that the status of the symptoms has changed and/or is clear; 38. Not necessarily be performed in a specific order. The sequence of movement testing does not affect the outcome or the ability to detect centralization

G. Criteria for positive test



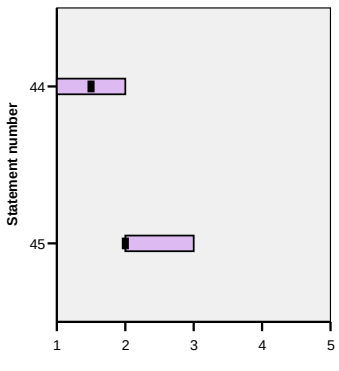
(a)



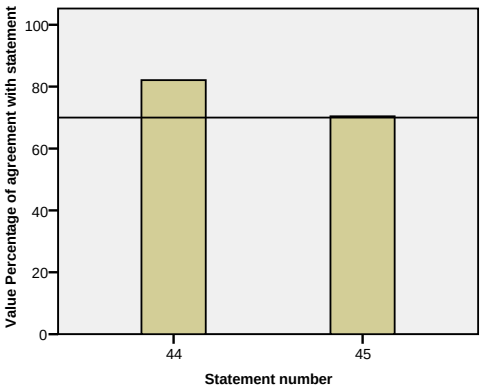
(b)

39. Changes in the intensity of symptoms should not be considered; 40. Changes in the neurological status (symptoms or signs) should not be considered; 41. Nonspinal conditions or underlying disease states potentially causing or influencing symptoms should be excluded or taken into consideration; 42. The reduction of peripheral symptoms towards the spinal midline should be progressive and stable. Distal pain which simply abolishes rather than progressively receding toward the spine should not be considered in the definition; 43. Observed changes should be retained over time (lasting change)

H. Testing – Timeframe



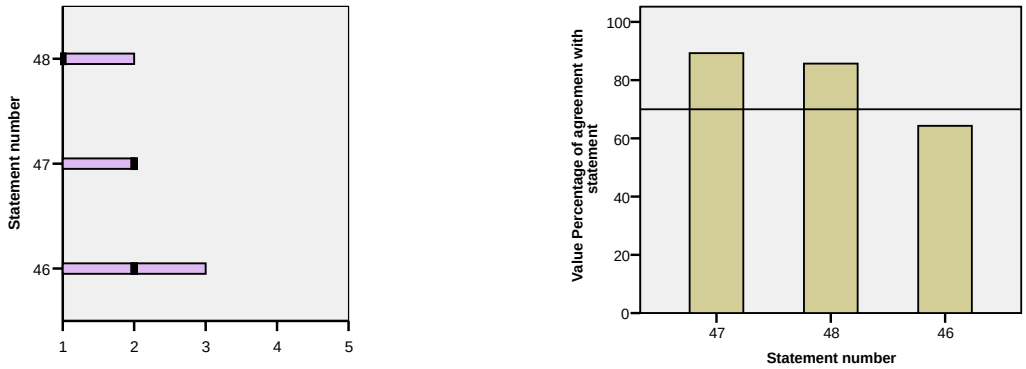
(a)



(b)

44. The response to testing may be obvious during the first examination or may require evaluation over a period of time to confirm the phenomenon; 45. Symptoms may be sequentially and lastingly abolished at each session (pure response) or gradually abolished in a progressive manner but not sequentially (partial response)

I. Safety issues

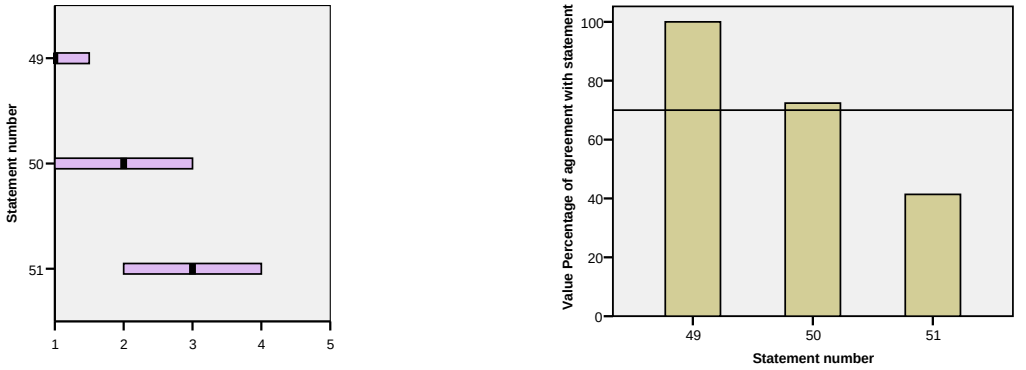


Strongly agree Strongly disagree (a) (b)

46. The patient's status (improving or worsening) should be considered beyond symptom relief i.e. neurological picture, range of motion. For example, if the patient's symptoms are improving but his neurological status is compromised, centralization should be considered an undesirable outcome for this patient; 47. Symptoms getting worse proximally but improving distally should be considered a positive sign; 48. Education of patients is essential following the use of these techniques so that movement is normalized as symptoms settle and patients do not develop fear of movement

---

Section 3: Other issues related to centralization



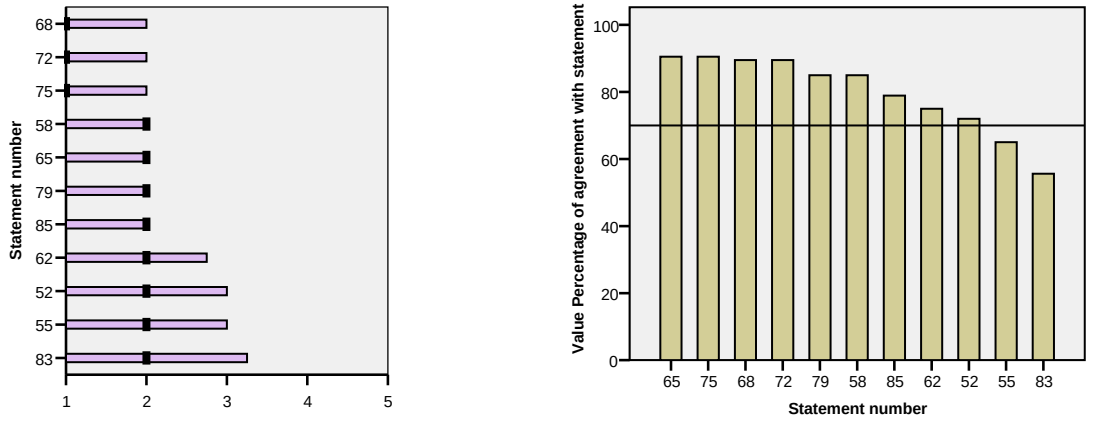
Strongly agree Strongly disagree (a) (b)

49. Centralizers have a good prognosis; 50. Centralization may be a stronger prognostic factor than psychosocial variables; 51. Centralizers have an internal disc disruption

---

Section 4: Future research

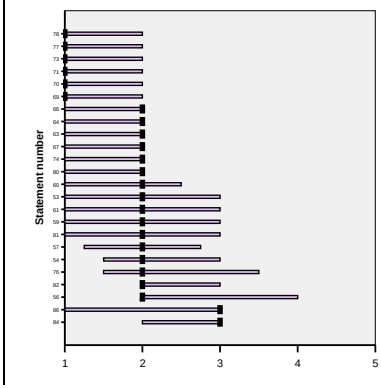
Future research areas:



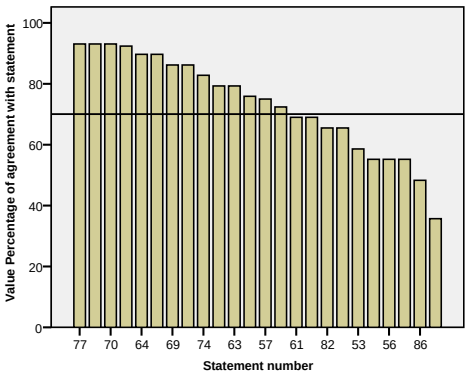
Strongly agree Strongly disagree (a) (b)

52. Operational definition for centralization; 55. Conceptual model for centralization; 58. Diagnostic accuracy of centralization; 62. Centralization and prevalence; 65. Centralization and course/ prognosis; 68. Prescriptive validity of centralization; 72. Centralization and subgroups; 75. Centralization and the cervical spine; 79. Reproducibility; 83. Case studies; 85. Education and training

Research topics:



Strongly agree Strongly disagree (a)



(b)

53. Head to head comparisons of different operational definitions; 54. Clarification of the term centralization i.e. number of repetitions, type of change, magnitude of change (e.g. belt line pain), duration and timeframe of required change, necessity of provocative testing for central symptoms; 56. Is centralization an anatomical phenomenon? Centralization in relation to tissue response; 57. Mechanism causing the centralization phenomenon (including physiological mechanism); 59. Criterion validity of centralization (e.g. using provocation discography as a standard; centralization as a tool for assessing the severity of a disc lesion; centralization as a tool for ruling out pathologies other than discogenic pain); 60. Cross validation of diagnostic accuracy in different patient populations and examiners; 61. Potential irreducible derangement i.e. pain moving centrally but neurological symptoms moving distally: at what point should patients be referred for further investigation? 63. Expected rates of centralization in clinical settings for acute and chronic patients with LBP; 64. Effect of training and procedures in the prevalence and outcomes of centralizers; 66. Do centralizers have a favourable course when left untreated (natural history)?; 67. Clinical response versus natural course of centralizers; 69. Do centralizers benefit from interventions other than directional movement exercise and which intervention provides the greatest benefit?; 70. Comparison of outcomes of non-responders to centralization for various interventions; 71. Centralization and outcomes (i.e. ability to return to work, psychosocial outcomes, economic outcomes, health care utilization, QoL, recurrences and ability to self-manage ; Cost effectiveness of the McKenzie assessment prior to disc surgery, injections, imaging or any other treatments ;73. Clinical predictors (CPR) of patients responding (or not) to directional preference exercises (including: Effect of patient compliance and attitudes of centralizers on outcomes; Effect of severe disability and psychosocial distress on the diagnostic, prognostic and treatment characteristics of centralizers; Centralization versus other prognostic factors); 74. Are centralizers and other clinical subgroups (e.g. patients satisfying CPR for manipulation or trunk stabilization treatment or candidates for disc surgery) made up of the same patients? How often do centralization findings co-exist with other findings? (i.e. relationship of centralization with other variables e.g. psychosocial, clinical findings e.g. segmental provocation signs, sciatica and EMG findings). How soon will positive EMGs become normal after centralization has been achieved and maintained?; 76. Centralization and contained cervical disc pathology; 77. Reliability of detecting, prevalence and outcomes in neck pain; 78. Predictive validity of centralization in the management of neck pain; 80. Stability/ reversibility of the centralization phenomenon; 81. Effect of the clinician, patient, standardization and knowledge of the test on the reliability of centralization; 82. Role of history and examiner's training in predicting the presence of centralization; 84. The rare patient whose pain centralizes up to the lumbar spine but then remains unchanged and may worsen with exercise. Generally improves over time and is usually initiated by long sitting time; 86. How do we best educate healthcare professionals that abolishment of leg or arm symptoms in a patient with central or foraminal stenosis is not a "centralizer", but a separate subgroup of their own?

This is the end of the feedback report.

Please do not forget to complete the third round questionnaire and send it back to us!

Thank you for your time and patience

Toward a uniform definition for the centralization phenomenon: A Delphi study

Instructions

This is the last questionnaire of the Delphi study. In this questionnaire, you are asked to reconsider the following statements and answer with 'yes or no' in the light of other participants' responses and feedback. If completing electronically, please ***click*** on the highlighted fields to mark or to type your answer for each statement. If completing the paper version, please ***put a cross*** in the box that is closest to how you feel. Before returning the completed form, please make sure you have answered **all** questions.

Please return the completed form by: **[date].** If you have any difficulties or questions about this questionnaire, please feel free to contact Angeliki Chorti (**A.Chorti@warwick.ac.uk**, TEL: 0044(0)2476150405).

Thank you again for your time and patience with this process.

Section 1: General definition

*Please **mark one box** for each statement that is closest to how you feel.*

Centralization should generally be defined as:

	Yes	No	% agree in Round 2
1. A lasting improvement in patient status (abolishment of distal symptoms or improvement of signs) in response to a defined movement which can vary in terms of direction, degree, duration, repetition and position used	☐	☐	**72.4%**
14. The progressive and stable reduction of the most distal pain towards the spinal midline in response to standardized repeated end-range movement or sustained loading testing procedures. Testing may involve multiple directions and various starting positions, progression of forces or alternative forces (e.g. clinician assistance) as well as more than one visit	☐	☐	**79.3%**
Please feel free to provide reasons for your choices and/ or add any important comments/ questions here:			

Section 2: Operational criteria

*Please **mark one box** for each statement that is closest to how you feel.*

A. Population for whom the test is intended

	Yes	No	% agree in Round 2
15. Centralization can be best appreciated only in patients who demonstrate referred or radiating symptoms originating from the spine	☐	☐	**79.3%**
16. Centralization can be best appreciated only in patients with low back pain (i.e. and not patients with symptoms originating from other spinal areas)	☐	☐	**17.2%**
17. Before testing, patients should be screened for the presence of red flags (i.e. serious pathology indicators)	☐	☐	**85.7%**
18. Before testing, patients should be screened for the presence of yellow flags (e.g. Waddell signs)	☐	☐	**31.0%**
19. Patients with distal symptoms above the knee that meet other parts of a clinical prediction rule for success with spinal manipulation are most likely to benefit from manipulation regardless of their response to repeated movement testing	☐	☐	**13.8%**
20. The centralization phenomenon becomes complex in chronic cases with an adherent nerve root involved. In such cases, the close monitoring of the symptoms and potential postural deformities is required and patients are treated on an individual basis	☐	☐	**48.3%**
Please feel free to provide reasons for your choices and/ or add any important comments/ questions here:			

B. Potential test users (examiners)

	Yes	No	% agree in Round 2
21. Centralization should be recognized by the average clinician rather than requiring years of training	☐	☐	**81.5%**
22. Training and experience of examiners is essential in order to have consistent results among clinicians	☐	☐	**63.0%**
22.1. Training of examiners is essential in order to have consistent results among clinicians	☐	☐	**New item**
22.2. The experience of examiners is essential in order to have consistent results among clinicians	☐	☐	**New item**
Please feel free to provide reasons for your choices and/ or add any important comments/ questions here:			

C. Tools used for documentation

	Yes	No	% agree in Round 2
23.The pattern of questioning and documenting patient responses i.e. pain intensity and location should be highly standardised	☐	☐	**82.8%**
23.1.The pattern of questioning and documenting patient responses i.e. pain intensity and location should be highly standardised in clinical practice and research	☐	☐	**New item**
23.2.The pattern of questioning and documenting patient responses i.e. pain intensity and location should be highly standardised in clinical practice only	☐	☐	**New item**
23.3.The pattern of questioning and documenting patient responses i.e. pain intensity and location should be highly standardised in research only	☐	☐	**New item**
24. For changes in pain location, a clear overlay numeric template should be used	☐	☐	**44.8%**
25. The area of the most distal symptoms should be marked and reassessed using a measurement from a bony landmark. The determination of the most distal pain should also be confirmed through palpation	☐	☐	**13.8%**
Please feel free to provide reasons for your choices and/ or add any important comments/ questions here:			

D. Loading strategy – Type N.B: please choose (with a 'yes' answer) **one** statement that is most representative of how you feel instead of multiple answers

Centralization can be elicited:

	Yes	No	% agree in Round 2
26. By repeated end-range movement testing only	☐	☐	**13.8%**
27. By repeated and/ or sustained end-range movement testing	☐	☐	**78.6%**
27a. By repeated and/ or sustained end-range movement testing **only**	☐	☐	**New item**
28. By single and repeated movements. Sustained movements are used with caution to prevent confusion with instability	☐	☐	**27.6%**
29. By repeated and/or sustained end-range movement testing, or manual techniques executed by therapist if appropriate	☐	☐	**89.7%**
30. By repeated and/or sustained end-range movement testing, overpressure by patient or therapist or manual techniques executed by therapist if appropriate	☐	☐	**82.8%**
31. Regardless of testing methods used. This can involve test movements or after immediate treatment application, or post treatment over time	☐	☐	**28.6%**
Please feel free to provide reasons for your choices and/ or add any important comments/ questions here:			

E. Loading strategy - Planes and directions N.B: please choose (with a 'yes' answer) **one** statement that is most representative of how you feel instead of multiple answers

Testing for centralization should:

	Yes	No	% agree in Round 2
32. Involve only the sagittal plane	☐	☐	**13.8%**
33. Involve the standard planes of movement available to the spinal region involved	☐	☐	**55.2%**
34. Involve the standard planes of movement available to the spinal region and/or a combination of movements (e.g. flexion with rotation) if appropriate	☐	☐	**79.3%**
Please feel free to provide reasons for your choices and/ or add any important comments/ questions here:			

F. Loading strategy - Movement testing

Movement testing for centralization should:

	Yes	No	% agree in Round 2
35. Not be based solely on single movements. Test movements must be performed repeatedly, to the fullest of the patient's available end-range	☐	☐	**75.9%**
36. Not include more than 10 repetitions for each test movement	☐	☐	**31.0%**
37. Include repetitions that are continued long enough to ensure that the status of the symptoms has changed and/or is clear.	☐	☐	**75.9%**
38. Not necessarily be performed in a specific order. The sequence of movement testing does not affect the outcome or the ability to detect centralization	☐	☐	**37.9%**
Please feel free to provide reasons for your choices and/ or add any important comments/ questions here:			

G. Criteria for positive test

When defining the presence of centralization:

	Yes	No	% agree in Round 2
39. Changes in the intensity of symptoms should not be considered	☐	☐	**24.1%**
40. Changes in the neurological status (symptoms or signs) should not be considered	☐	☐	**13.8%**
41. Nonspinal conditions or underlying disease states potentially causing or influencing symptoms should be excluded or taken into consideration	☐	☐	**82.1%**
41a. Underlying disease states potentially causing or influencing symptoms should be excluded	☐	☐	**New item**
41b. Nonspinal conditions potentially causing or influencing symptoms should be taken into consideration	☐	☐	**New item**
42. The reduction of peripheral symptoms towards the spinal midline should be progressive and stable. Distal pain which simply abolishes rather than progressively receding toward the spine should not be considered in the definition	☐	☐	**34.5%**
43. Observed changes should be retained over time (lasting change after testing)	☐	☐	**75.9%**
Please feel free to provide reasons for your choices and/ or add any important comments/ questions here:			

H. Testing – Timeframe

When testing for centralization:

	Yes	No	% agree in Round 2
44. The response to testing may be obvious during the first examination or may require evaluation over a period of time to confirm the phenomenon	☐	☐	**82.1%**
45. Symptoms may be sequentially and lastingly abolished at each session (pure response) or gradually abolished in a progressive manner but not sequentially (partial response)	☐	☐	**70.4%**
Please feel free to provide reasons for your choices and/ or add any important comments/ questions here:			

I. Safety issues

	Yes	No	% agree in Round 2
46. The patient's status (improving or worsening) should be considered beyond symptom relief i.e. neurological picture, range of motion. For example, if the patient's symptoms are improving but his neurological status is compromised, centralization should be considered an undesirable outcome for this patient.	☐	☐	**64.3%**
47. Symptoms getting worse proximally but improving distally should be considered a positive sign	☐	☐	**89.3%**
48. Education of patients is essential following the use of these techniques so that movement is normalized as symptoms settle and patients do not develop fear of movement	☐	☐	**85.7%**
Please feel free to provide reasons for your choices and/ or add any important comments/ questions here:			

Section 3: Other issues related to centralization

*Please **mark one box** for each statement that is closest to how you feel.*

	Yes	No	% agree in Round 2
49. Centralizers have a good prognosis	☐	☐	**100%**
50. Centralization may be a stronger prognostic factor than psychosocial variables	☐	☐	**72.4%**
51. Centralizers have an internal disc disruption	☐	☐	**41.4%**
Please feel free to provide reasons for your choices and/ or add any important comments/ questions here:			

Section 4: Future research

*Please **mark one box** for each future area and question that is closest to how you feel.*

Future research should look at:

	Yes	No	% agree in Round 2
52. Operational definition for centralization:	☐	☐	**72.0%**
57. Mechanism causing the centralization phenomenon (including physiological mechanism)	☐	☐	**75.0%**
58. Diagnostic accuracy of centralization:	☐	☐	**85.0%**
60. Cross validation of diagnostic accuracy in different patient populations and examiners	☐	☐	**75.9%**
62. Centralization and prevalence:	☐	☐	**75.0%**
63. Expected rates of centralization in clinical settings for acute and chronic patients with LBP	☐	☐	**79.3%**
64. Effect of training and procedures in the prevalence and outcomes of centralizers	☐	☐	**89.7%**
65. Centralization and course/ prognosis:	☐	☐	**90.5%**
66. Do centralizers have a favourable course when left untreated (natural history)?	☐	☐	**89.7%**
67. Clinical response versus natural course of centralizers	☐	☐	**92.9%**
68. Prescriptive validity of centralization:	☐	☐	**89.5%**
69. Do centralizers benefit from interventions other than directional movement exercise and which intervention provides the greatest benefit?	☐	☐	**86.2%**
70. Comparison of outcomes of non-responders to centralization for various interventions	☐	☐	**93.1%**
71. Centralization and outcomes (i.e. ability to return to work, psychosocial outcomes, economic outcomes, health care utilization, QoL, recurrences and ability to self-manage ; Cost effectiveness of the McKenzie assessment prior to disc surgery, injections, imaging or any other treatments)	☐	☐	**86.2%**
72. Centralization and subgroups:	☐	☐	**89.5%**
73. Clinical predictors (CPR) of patients responding (or not) to directional preference exercises (including: Effect of patient compliance and attitudes of centralizers on outcomes; Effect of severe disability and psychosocial distress on the diagnostic, prognostic and treatment characteristics of centralizers; Centralization versus other prognostic factors)	☐	☐	**79.3%**
74. Are centralizers and other clinical subgroups (e.g. patients satisfying CPR for manipulation or trunk stabilization treatment or candidates for disc surgery) made up of the same patients? How often do centralization findings co-exist with other findings? (i.e. relationship of centralization with other variables e.g. psychosocial, clinical findings e.g. segmental provocation signs, sciatica and EMG findings). How soon will positive EMGs become normal after centralization has been achieved and maintained?	☐	☐	**82.8%**

	Yes	No	% agree in Round 2
74a. Are centralizers and other clinical subgroups (e.g. patients satisfying CPR for manipulation or trunk stabilization treatment or candidates for disc surgery) made up of the same patients?	☐	☐	**New item**
74b. How often do centralization findings co-exist with other findings? (i.e. relationship of centralization with other variables e.g. psychosocial, clinical findings e.g. segmental provocation signs, sciatica and EMG findings).	☐	☐	**New item**
74c. How soon will positive findings become normal after centralization has been achieved and maintained?	☐	☐	**New item**
75. Centralization and the cervical spine:	☐	☐	**90.5%**
77. Reliability of detecting, prevalence and outcomes in neck pain	☐	☐	**93.1%**
77a. Reliability of detection of centralization in neck pain	☐	☐	**New item**
77b. Prevalence of centralization in neck pain	☐	☐	**New item**
77c. Outcomes of centralizers in neck pain	☐	☐	**New item**
78. Predictive validity of centralization in the management of neck pain	☐	☐	**93.1%**
79. Reproducibility:	☐	☐	**85.0%**
80. Stability/ reversibility of the centralization phenomenon	☐	☐	**72.4%**
81a. Effect of the clinician characteristics on the reliability of centralization	☐	☐	**New item**
81b. Effect of the patient characteristics on the reliability of centralization	☐	☐	**New item**
81c. Effect of standardization of the test on the reliability of centralization	☐	☐	**New item**
81d. Effect of knowledge of the test on the reliability of centralization	☐	☐	**New item**
82a. Role of history in predicting the presence of centralization (prior to examination)	☐	☐	**New item**
82b. Role of examiner's training in predicting the presence of centralization	☐	☐	**New item**
85. Education and training:	☐	☐	**78.9%**
Please feel free to provide reasons for your choices and/ or add any important comments/ questions here:			

This is the end of the third round Delphi questionnaire.

Please make sure you have answered all the questions and selected only one answer for each question.

Thank-you for taking the time to complete this questionnaire

Please return the completed questionnaire to: Angeliki Chorti, Warwick Medical School, University of Warwick, Gibbet Hill Road, Coventry, CV4 7AL

Electronically to: A.Chorti@warwick.ac.uk

By: [date]

If you wish to complete the questionnaire in paper format, please contact **Ms Angeliki Chorti** using the contact details provided above. A hard version will be sent to you, along with a return envelope. Alternatively, please feel free to print and send back. Again, a return envelope will happily be provided.

Emailed questionnaires will be treated anonymously.

Appendix 4.10 Delphi feedback report from round 3.

Toward a uniform definition for the centralization phenomenon: A Delphi study

This is the feedback report for the third round of the Delphi study. This report is divided into two parts. Part 1 refers to the general statistics of your ratings and includes your comments and Part 2 presents a graphical display of your responses.

For any questions regarding this form, feel free to contact Ms Angeliki Chorti (Email: **A.Chorti@warwick.ac.uk**, Tel: 0044(0)2476150405) who will be happy to provide you with further information.

Thank you for reading this.

PART 1: GENERAL STATISTICS AND COMMENTS

N = number of valid responses; % agree = percentage of participants that agreed with statement, green ≥ 80% agreement; 70% ≤yellow ≤ 79% agreement.

Section 1: General definition

Centralization should generally be defined as:

	N	% agree
1. A lasting improvement in patient status (abolition of distal symptoms or improvement of signs) in response to a defined movement which can vary in terms of direction, degree, duration, repetition and position used	**30**	**63.3**
14. The progressive and stable reduction of the most distal pain towards the spinal midline in response to standardized repeated end-range movement or sustained loading testing procedures. Testing may involve multiple directions and various starting positions, progression of forces or alternative forces (e.g. clinician assistance) as well as more than one visit	**30**	**80.0**

COMMENTS
Statement 1: Does not include other loading strategies; The general definition should include the movement of symptoms toward the midline; Signs should not be considered in the definition
Statement 14: Too focused; More comprehensive than other definitions; Too flexible; Does not reflect standardized examination; Problematic when referring to central pain only; Potential confusion with natural course of symptoms or non-specific improvement due to time element; Stability and lasting change should be defined better.

Section 2: Operational criteria

A. Population for whom the test is intended

	N	% agree
15. Centralization can be best appreciated only in patients who demonstrate referred or radiating symptoms originating from the spine	**30**	**86.7**
16. Centralization can be best appreciated only in patients with low back pain i.e. and not patients with symptoms originating from other spinal areas	**29**	**0.00**
17. Before testing, patients should be screened for the presence of red flags (i.e. serious pathology indicators)	**30**	**96.7**
18. Before testing, patients should be screened for the presence of yellow flags (e.g. Waddell signs)	**31**	**32.3**
19. Patients with distal symptoms above the knee that meet other parts of a clinical prediction rule for success with spinal manipulation are most likely to benefit from manipulation regardless of their response to repeated movement testing	**31**	**22.6**
20. The centralization phenomenon becomes complex in chronic cases with an adherent nerve root involved. In such cases, the close monitoring of the symptoms and potential postural deformities is required and patients are treated on an individual basis	**30**	**36.7**

COMMENTS
Statement 15: Best and only contradict each other
Statement 16: Leaves out other spinal areas
Statement 17: The dangers associated with the testing procedure are minimal; Patients with red flags should not be excluded from the testing unless they can not perform the movements; If a patient is not appropriate for mechanical therapy, he/she will not centralise.
Statement 18: Psychosocial factors can have an impact on musculoskeletal pain and may confound the clinical interpretation of physical findings; Prefer FABQ to Waddell signs; Research has shown that physical therapists do not fully understand how to integrate and evaluate the psychosocial domains with physical domain during the clinical examination of patients with low back pain
Statement 19: Childs et al. Ann Int. Med, 2004 have provided evidence for this statement and also good evidence that patients with symptoms distal to the knee are not as likely to respond to manipulation; In a study by Browder et al. PT 2007, the treatment effect was dramatic in comparison to an extension-oriented treatment approach (EOTA); The manipulation CPR has not been thoroughly tested and we know of one published example that shows

manipulation to not be the optimal treatment for a patient who fits the manipulation CPR characteristics; Any benefit from manipulation would be overshadowed by the benefit of teaching self-care to centralize and abolish symptoms as well as empowerment for prevention of recurrences using directional exercises and posture modifications; This statement is more relevant to the McKenzie method
Statement 20: Close monitoring and individualized treatment would apply to all patients anyway; No reliable/valid method for 'diagnosing' an adherent nerve root; If the patient has an adherent nerve root, there is no centralization

B. Potential test users (examiners)

	N	% agree
21. Centralization should be recognized by the average clinician rather than requiring years of training	**30**	**80.0**
22. Training and experience of examiners is essential in order to have consistent results among clinicians	**30**	**56.7**
22.1. Training of examiners is essential in order to have consistent results among clinicians	**30**	**86.7**
22.2. The experience of examiners is essential in order to have consistent results among clinicians	**29**	**44.8**

COMMENTS
Statement 21: To be useful and widely used it should be recognized by average and entry-level clinicians; We do not have a definition of what "average" is. More research is needed to clarify the issues around training and experience; This statement depends on whether clinicians have sufficient education in entry level physiotherapy programs; Post-graduate training may be required
Statements 22, 22.1, 22.2: Standardization (e.g. specificity and clarity of definitions) is required to achieve the same shared baseline knowledge of procedures; Training is important for consistency, experience is not; Good education and learning experiences are more important than years of experience; Experience is desirable and improves matters, but we all start out as novices; Prior research has demonstrated that appropriate training is needed in order to have consistency and reliability; Consistency of results might depend on the quality, level and amount of training and experience; Current research does support training but there is no research on the level of experience; Training is required, and should be at undergraduate level. All physiotherapists should know how to do the test procedures and interpret the patient's responses.

C. Tools used for documentation

	N	% agree
23.The pattern of questioning and documenting patient responses i.e. pain intensity and location should be highly standardised	**30**	**73.3**
23.1. The pattern of questioning and documenting patient responses i.e. pain intensity and location should be highly standardised in clinical practice and research	**30**	**76.7**
23.2. The pattern of questioning and documenting patient responses i.e. pain intensity and location should be highly standardised in clinical practice only	**30**	**3.3**
23.3. The pattern of questioning and documenting patient responses i.e. pain intensity and location should be highly standardised in research only	**30**	**20.0**
24. For changes in pain location, a clear overlay numeric template should be used	**31**	**16.1**
25. The area of the most distal symptoms should be marked and reassessed using a measurement from a bony landmark. The determination of the most distal pain should also be confirmed through palpation	**31**	**0.00**

COMMENTS
Statements 23, 23.1, 23.2, 23.3: The word "highly" should be defined or excluded; Standardisation may help, but we do not know. It has not been studied; Standardization does not necessarily have to be prescriptive. It may just imply consistency and clarity of reporting method; Failure to recognise the pain distribution and intensity prior to test manoeuvres is common among clinicians, therefore standardisation of assessment is important; Research may require standardisation but with different standards depending on the research question; Standardisation is essential for some research studies in order to be replicated; Routine clinical work may require less standardisation; In research, a researcher may choose highly standardized procedure different from McKenzie's recommendations for the purposes of determining if another method is superior.
Statements 24: Too prescriptive for clinical practice, and unnecessary; For research purposes only; The evidence and a recent systematic review on centralization support a measurement tool to document centralization. Perhaps one of several reasons for the large variance in the prevalence rates of centralization across studies is the lack of a standardized documentation process; This is an objective measurement tool and physical therapy guidelines encourage therapists to use objective measurement procedures; Pain overlay template is one possibility however, there are other ways especially if using computer-based assessment.
Statement 25: Too prescriptive for clinical practice, and unnecessary; Palpation has no validity.

D. Loading strategy – Type
Centralization can be elicited:

	N	% agree
26. By repeated end-range movement testing only	**29**	**3.4**
27. By repeated and/ or sustained end-range movement testing	**28**	**60.7**
27a. By repeated and/ or sustained end-range movement testing **only**	**29**	**10.3**
28. By single and repeated movements. Sustained movements are used with caution to prevent confusion with instability	**28**	**14.3**
29. By repeated and/or sustained end-range movement testing, or manual techniques executed by therapist if appropriate	**29**	**79.3**
30. By repeated and/or sustained end-range movement testing, overpressure by patient or therapist or manual techniques executed by therapist if appropriate	**30**	**90.0**
31. Regardless of testing methods used. This can involve test movements or after immediate treatment application, or post treatment over time	**28**	**17.9**

COMMENTS
Statement 28: I have no idea what the statement means as instability has not been defined; Centralization should be distinct from instability and manual techniques, therefore should avoid sustained movements and therapist overpressure.
Statements 29, 30: Overpressure is a manual technique so there is no distinction between 29 and 30; It is essential to allow for manual overpressure by the therapist because some patients cannot achieve the appropriate end range for many different reasons. In the case of correction of a lateral shift, self correction by the patient often fails, but manual shift correction causes the centralization phenomenon easily; I think a good compromise would be to put a period in #30 after the phrase "over pressure by patient to therapist." And the delete the manual techniques wording
Statement 31: I have no idea what this statement means or what it refers to

E. Loading strategy - Planes and directions

Testing for centralization should:

	N	% agree
32. Involve only the sagittal plane	**28**	**3.6**
33. Involve the standard planes of movement available to the spinal region involved	**27**	**33.3**
34. Involve the standard planes of movement available to the spinal region and/or a combination of movements if appropriate	**29**	**93.1**

COMMENTS
Statement 33: What are the standard planes of movement?
Statement 34: Side gliding is not considered a standard plane of movement by many, but an essential inclusion in the lumbar spine. Retraction in the cervical spine suffers the same divergence of opinion and is very important in eliciting centralization; This statement does not include sustained positions

F. Loading strategy - Movement testing

Movement testing for centralization should:

	N	% agree
35. Not be based solely on single movements. Test movements must be performed repeatedly, to the fullest of the patient's available end-range	**30**	**86.7**
36. Not include more than 10 repetitions for each test movement	**30**	**20.0**
37. Include repetitions that are continued long enough to ensure that the status of the symptoms has changed and/or is clear.	**30**	**93.3**
38. Not necessarily be performed in a specific order. The sequence of movement testing does not affect the outcome or the ability to detect centralization	**30**	**40.0**

COMMENTS
Statement 35: Does not include sustained positons.
Statement 37: The exact number of repetitions is not important, the pattern recognition is
Statement 38: This is supported by published data (Donelson et al. in Spine vol 16 1991); We do not know, it has not been studied; In some instances eg patient with a lateral shift, the order of movements is important but not in all cases

G. Criteria for positive test

When defining the presence of centralization:

	N	% agree
39. Changes in the intensity of symptoms should not be considered	**30**	**20.0**
40. Changes in the neurological status (symptoms or signs) should not be considered	**31**	**12.9**
41. Nonspinal conditions or underlying disease states potentially causing or influencing symptoms should be excluded or taken into consideration	**31**	**87.1**
41a. Underlying disease states potentially causing or influencing symptoms should be excluded	**31**	**67.7**
41b. Nonspinal conditions potentially causing or influencing symptoms should be taken into consideration	**31**	**87.1**
42. The reduction of peripheral symptoms towards the spinal midline should be progressive and stable. Distal pain which simply abolishes rather than progressively receding toward the spine should not be considered in the definition	**31**	**22.6**
43. Observed changes should be retained over time (lasting change after testing)	**29**	**86.2**

COMMENTS
Statements 39 & 40: The clinician should differentiate between a directional preference and centralization. Both terms overlap but these terms are not synonymous; Further work is required to operationally define these terms to decrease the confusion between clinical identification of directional preference and/ or centralization. Without this important clinical discussion and continued research confusion will continue regarding how best to define centralization.
Statement 41, 41a, 41b: I'm not sure nonspinal conditions/disease states are criteria for a positive test. They definitely need to be considered as part of complete evaluation
Statement 42: More testing is usually needed
Statement 43: Most appropriate for pure criteria for positive test; Statement needs to be rephrased: over how much time and under what conditions? How long after testing?; Does not apply to centralization, but if referring to the process of centralization or centralizing in the direction of full centralization, that is another topic; It is just the ability of changes to predict longer lasting changes that are of primary concern. Otherwise one is at risk of a circular argument simply suggesting that if someone improves over time that they have improved over time.

H. Testing – Timeframe

When testing for centralization:

	N	% agree
44. The response to testing may be obvious during the first examination or may require evaluation over a period of time to confirm the phenomenon	**31**	**100.0**
45. Symptoms may be sequentially and lastingly abolished at each session (pure response) or gradually abolished in a progressive manner but not sequentially (partial response)	**27**	**74.1**

COMMENTS
Statement 45: This statement is too complex and confusing; Sequentially should be defined; Can't say I've seen a non-sequential abolition moving gradually toward complete abolition; Centralization rarely if ever takes more than a week to be identified, unless the patient is slow to adopt the postural changes and exercise regime between assessments; I agree with 45, but it does not include all possibilities.
Statements 44 & 45: These questions are hard to answer because they both combine the timeframes. I can disagree with both of them and still answer yes.

I. Safety issues

	N	% agree
46. The patient's status (improving or worsening) should be considered beyond symptom relief i.e. neurological picture, range of motion. For example, if the patient's symptoms are improving but his neurological status is compromised, centralization should be considered an undesirable outcome for this patient.	**27**	**59.3**
47. Symptoms getting worse proximally but improving distally should be considered a positive sign	**30**	**100.0**
48. Education of patients is essential following the use of these techniques so that movement is normalized as symptoms settle and patients do not develop fear of movement	**30**	**90.0**

COMMENTS
Statement 46: Unclear question. Was neural status compromised at the outset and is unchanged, or is it deteriorating as a result of the assessment? If the latter, one must cease, although, again, that never happens in the face of centralization. That's one of the wonderful safe guards about the MDT assessment. Certainly, monitoring neural status is paramount and deterioration cannot be accepted; Neurolgical status should be monitored additionally e.g. if neurologic status is worsening and only pain is improving, this should not be considered a positive sign; Question 46 seems to me a purely "researcher" question. Never happened to me that centralization were linked to a worsening of the general status; True centralization with concurrent worsening of neurological status must be vanishingly rare. I cannot recall a case in 30 years of doing this; In this scenario centralization is not the undesirable outcome - the worsening neurological status is the undesirable outcome and centralization of symptoms is besides the point - this is an important distinction but not a good description of it. The point is that if an intervention is worsening the patient's neurological status then it may be inappropriate despite centralization of symptoms. This appears to happen occasionally in patients with spinal stenosis- the little used 'pheasant's sign' was reported to look for this phenomenon.

Statement 48: This seems to speak to intervention, not examination; I would agree with 48, but it would appear to be a general principle and have little to do with concepts of centralization; Education is certainly essential. However, some therapists make the error of discontinuing the movement causing centralization and the expected increase in proximal pain. The end result is that the wrong patient education is then provided

---

Section 3: Other issues related to centralization

	N	% agree
49. Centralizers have a good prognosis	**30**	**96.7**
50. Centralization may be a stronger prognostic factor than psychosocial variables	**29**	**65.5**
51. Centralizers have an internal disc disruption	**27**	**14.8**

COMMENTS
Statement 50: Psychosocial variables are another issue
Statement 51: This is supported by Laslett M et al TSJ 2005; There is currently evidence in the lumbar spine but not the cervical spine; More evidence is needed in support of #51; I disagree strongly with 51 and think its inclusion potentially reduces the credibility of the concept of centralization; I don't know, I don't care
Statements 49, 50 & 51: All three statements are true to some extent. However, those with severe psychologcial distress may require psychological intervention before being able to adhere to the self-treatment program required to centralize and abolish pain. Those with severe IDD (i.e., grade IV on the Dallas Discogram Scale) or who have satellite fissures not connected to the nucleus of the disc may not centralize.

Section 4: Future research

Future research should look at:

	N	% agree
52. Operational definition for centralization:	**29**	**86.2**
57. Mechanism causing the centralization phenomenon (including physiological mechanism)	**29**	**79.3**
58. Diagnostic accuracy of centralization:	**28**	**92.9**
60. Cross validation of diagnostic accuracy in different patient populations and examiners	**30**	**86.7**
62. Centralization and prevalence:	**28**	**96.4**
63. Expected rates of centralization in clinical settings for acute and chronic patients with LBP	**29**	**86.2**
64. Effect of training and procedures in the prevalence and outcomes of centralizers	**28**	**100.0**
65. Centralization and course/ prognosis:	**28**	**92.9**
66. Do centralizers have a favourable course when left untreated (natural history)?	**28**	**92.9**
67. Clinical response versus natural course of centralizers	**30**	**93.3**
68. Prescriptive validity of centralization:	**28**	**89.3**
69. Do centralizers benefit from interventions other than directional movement exercise and which intervention provides the greatest benefit?	**29**	**89.7**
70. Comparison of outcomes of non-responders to centralization for various interventions	**30**	**96.7**
71. Centralization and outcomes (i.e. ability to return to work, psychosocial outcomes, economic outcomes, health care utilization, QoL, recurrences and ability to self-manage ; Cost effectiveness of the McKenzie assessment prior to disc surgery, injections, imaging or any other treatments)	**28**	**85.7**
72. Centralization and subgroups:	**28**	**85.7**
73. Clinical predictors (CPR) of patients responding (or not) to directional preference exercises (including: Effect of patient compliance and attitudes of centralizers on outcomes; Effect of severe disability and psychosocial distress on the diagnostic, prognostic and treatment characteristics of centralizers; Centralization versus other prognostic factors)	**28**	**82.1**
74. Are centralizers and other clinical subgroups (e.g. patients satisfying CPR for manipulation or trunk stabilization treatment or candidates for disc surgery) made up of the same patients? How often do centralization findings co-exist with other findings? (i.e. relationship of centralization with other variables e.g. psychosocial, clinical findings e.g. segmental provocation signs, sciatica and EMG findings). How soon will positive EMGs become normal after centralization has been achieved and maintained?	**28**	**89.3**
74a. Are centralizers and other clinical subgroups (e.g. patients satisfying CPR for manipulation or trunk stabilization treatment or candidates for disc surgery) made up of the same patients?	**28**	**71.4**
74b. How often do centralization findings co-exist with other findings? (i.e. relationship of centralization with other variables e.g. psychosocial, clinical findings e.g. segmental provocation signs, sciatica and EMG findings).	**27**	**77.8**
74c. How soon will positive EMGs become normal after centralization has been achieved and maintained?	**27**	**77.8**
75. Centralization and the cervical spine:	**27**	**92.6**
77. Reliability of detecting, prevalence and outcomes in neck pain	**29**	**96.6**
77a. Reliability of detecting centralization in neck pain	**25**	**92.0**
77b. Prevalence of centralization in neck pain	**27**	**92.6**
77c. Outcomes of centralization in neck pain	**27**	**92.6**

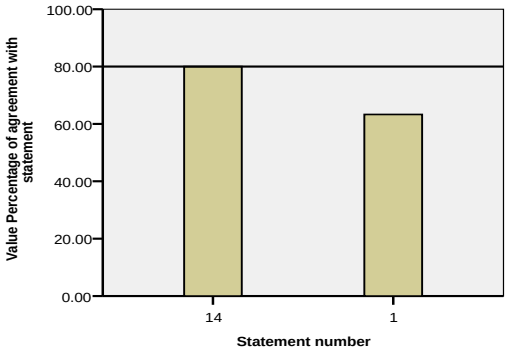
78. Predictive validity of centralization in the management of neck pain	**27**	**92.6**
79. Reproducibility:	**29**	**86.2**
80. Stability/ reversibility of the centralization phenomenon	**28**	**78.6**
81a. Effect of the clinician characteristics on the reliability of centralization	**27**	**74.1**
81b. Effect of the patient characteristics on the reliability of centralization	**28**	**78.6**
81c. Effect of standardization of the test on the reliability of centralization	**28**	**82.1**
81d. Effect of knowledge of the test on the reliability of centralization	**28**	**71.4**
82a. Role of history in predicting the presence of centralization	**28**	**82.1**
82b Role of examiner's training in predicting the presence of centralization	**28**	**85.7**
85. Education and training:	**29**	**86.2**

COMMENTS
Statement 70: "non-responders to centralization"? Makes no sense. Perhaps "non-centralizers"?
Statement 71: Change to 'physical therapy assessment' rather than 'McKenzie assessment' and I will agree.
Statement 74: While useful for prognosis, there has not been any indication in the line of research utilizing clinical prediction rules to determine the effectiveness of various treatment approaches that centralization of symptoms in the absence of neurological signs is a sign that should trump other factors (i.e. time in the manipulation CPR), particularly given different in effect sizes- symptoms distal to the knee seems to be important. These two in combination (symptoms distal to the knee and centralization of symptoms) seem to be likely candidates for patients that will respond best to a direction specific treatment approach more than other treatment approaches. Utilizing 'centralizers' as a stand alone subgroup does not seem to be the best approach; Question 74 contains 3 questions, at which am I supposed to answer? Questions 74 b and c: how is possible to answer with yes or no to questions starting with: "How"?
Statement 85: No answers for question #85?
Section 4: Section 4 is confusing. Too many similar questions. Shot gun approach. What is required is a systematic and reasoned approach to research.

PART 2: GRAPHS

This part presents a graphical display of your responses. The statement numbers are presented in the graphs in order of importance (i.e. best to worse rating in round 3). The graphs refer to the percentage of agreement with each statement. The vertical line in graphs (b) represents the inclusion cut-off point of 80%.

Section 1: General definition

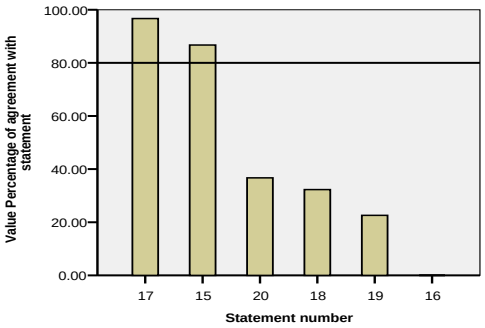


Centralization should generally be defined as:

1. A lasting improvement in patient status (abolishment of distal symptoms or improvement of signs) in response to a defined movement which can vary in terms of direction, degree, duration, repetition and position used; 14. The progressive and stable reduction of the most distal pain towards the spinal midline in response to standardized repeated end-range movement or sustained loading testing procedures. Testing may involve multiple directions and various starting positions, progression of forces or alternative forces (e.g. clinician assistance) as well as more than one visit

Section 2: Operational criteria

A. Population for whom the test is intended

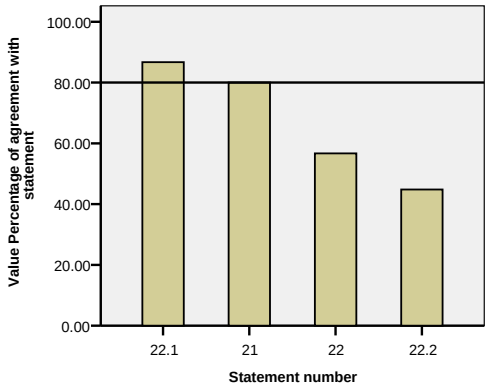


Population:

15. Centralization can be best appreciated only in patients who demonstrate referred or radiating symptoms originating from the spine; 16. Centralization can be best appreciated only in patients with low back pain i.e. and not patients with symptoms originating from other spinal areas; 17. Before testing, patients should be screened for the presence of red flags (i.e. serious pathology indicators); 18. Before testing, patients should be screened for the presence of yellow flags (e.g. Waddell signs); 19. Patients with distal symptoms above the knee that meet other parts of a clinical prediction rule for success with spinal manipulation are most likely to benefit from manipulation regardless of their response to repeated movement testing; 20. The centralization phenomenon becomes complex in chronic cases with an adherent nerve root involved. In such cases, the close monitoring of the

symptoms and potential postural deformities is required and patients are treated on an individual basis

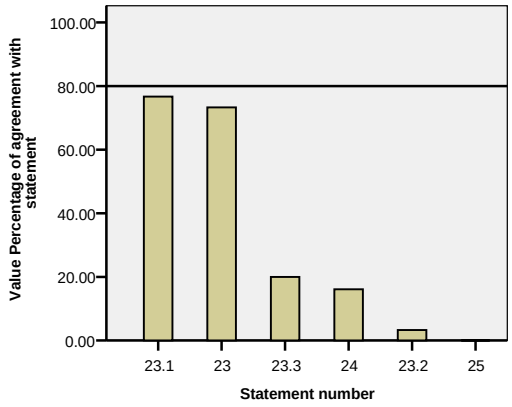
B. Potential test users (examiners)



Test users:

21. Centralization should be recognized by the average clinician rather than requiring years of training; 22. Training and experience of examiners is essential in order to have consistent results among clinicians; 22.1 Training of examiners is essential in order to have consistent results among clinicians; 22.2 The experience of examiners is essential in order to have consistent results among clinicians; 22.1. Training of examiners is essential in order to have consistent results among clinicians; 22.2. The experience of examiners is essential in order to have consistent results among clinicians

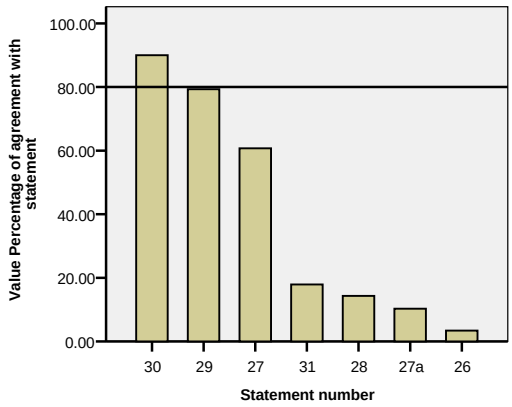
C. Tools used for documentation



Documentation:

23. The pattern of questioning and documenting patient responses i.e. pain intensity and location should be highly standardised; 23.1 The pattern of questioning and documenting patient responses i.e. pain intensity and location should be highly standardised in clinical practice and research; 23.2 The pattern of questioning and documenting patient responses i.e. pain intensity and location should be highly standardised in clinical practice only; 23.3 The pattern of questioning and documenting patient responses i.e. pain intensity and location should be highly standardised in research only; 24. For changes in pain location, a clear overlay numeric template should be used; 25. The area of the most distal symptoms should be marked and reassessed using a measurement from a bony landmark. The determination of the most distal pain should also be confirmed through palpation

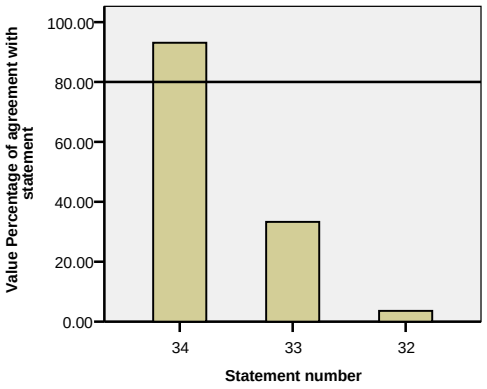
D. Loading strategy – Type



Centralization can be elicited:

26. By repeated end-range movement testing only; 27. By repeated and/ or sustained end-range movement testing; 27a. By repeated and/ or sustained end-range movement testing only; 28. By single and repeated movements. Sustained movements are used with caution to prevent confusion with instability; 29. By repeated and/or sustained end-range movement testing, or manual techniques executed by therapist if appropriate; 30. By repeated and/or sustained end-range movement testing, overpressure by patient or therapist or manual techniques executed by therapist if appropriate; 31. Regardless of testing methods used. This can involve test movements or after immediate treatment application, or post treatment over time

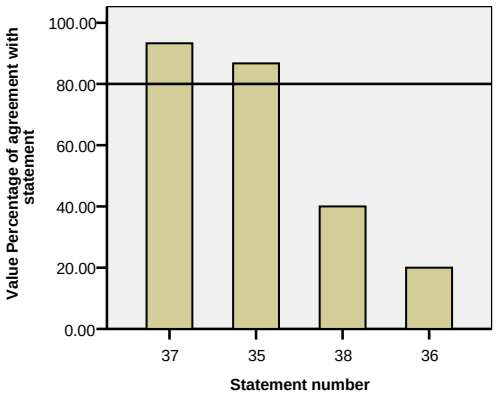
E. Loading strategy - Planes and directions



Testing for centralization should:

32. Involve only the sagittal plane; 33. Involve the standard planes of movement available to the spinal region involved; 34. Involve the standard planes of movement available to the spinal region and/or a combination of movements if appropriate

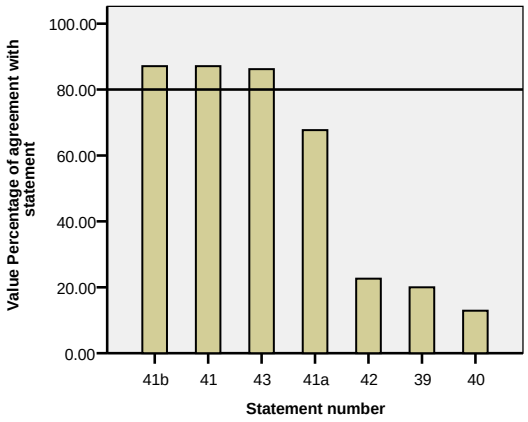
F. Loading strategy - Movement testing



Movement testing for centralization should:

35. Not be based solely on single movements. Test movements must be performed repeatedly, to the fullest of the patient's available end-range; 36. Not include more than 10 repetitions for each test movement; 37. Include repetitions that are continued long enough to ensure that the status of the symptoms has changed and/or is clear; 38. Not necessarily be performed in a specific order. The sequence of movement testing does not affect the outcome or the ability to detect centralization

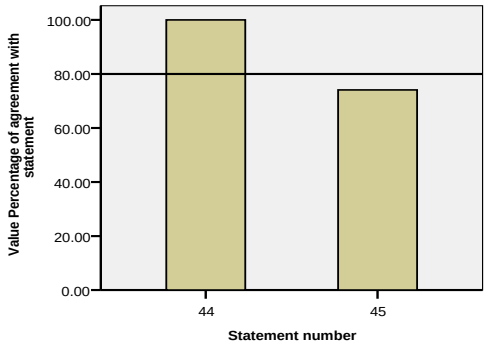
G. Criteria for positive test



Criteria for a positive test:

39. Changes in the intensity of symptoms should not be considered; 40. Changes in the neurological status (symptoms or signs) should not be considered; 41. Nonspinal conditions or underlying disease states potentially causing or influencing symptoms should be excluded or taken into consideration; 41a. Underlying disease states potentially causing or influencing symptoms should be excluded; 41b. Nonspinal conditions potentially causing or influencing symptoms should be taken into consideration 42. The reduction of peripheral symptoms towards the spinal midline should be progressive and stable. Distal pain which simply abolishes rather than progressively receding toward the spine should not be considered in the definition; 43. Observed changes should be retained over time (lasting change)

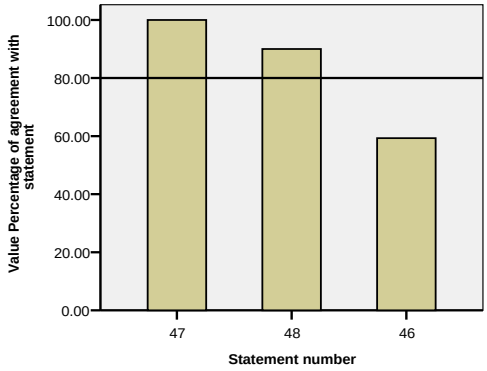
H. Testing – Timeframe



Timeframe:

44. The response to testing may be obvious during the first examination or may require evaluation over a period of time to confirm the phenomenon; 45. Symptoms may be sequentially and lastingly abolished at each session (pure response) or gradually abolished in a progressive manner but not sequentially (partial response)

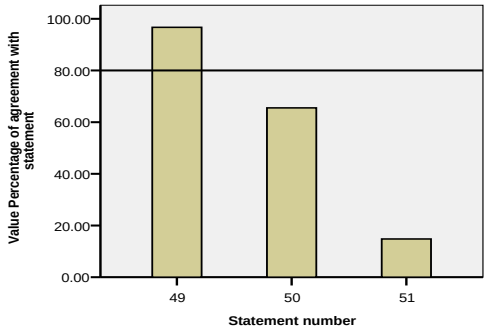
I. Safety issues



Safety issues:

46. The patient's status (improving or worsening) should be considered beyond symptom relief i.e. neurological picture, range of motion. For example, if the patient's symptoms are improving but his neurological status is compromised, centralization should be considered an undesirable outcome for this patient; 47. Symptoms getting worse proximally but improving distally should be considered a positive sign; 48. Education of patients is essential following the use of these techniques so that movement is normalized as symptoms settle and patients do not develop fear of movement

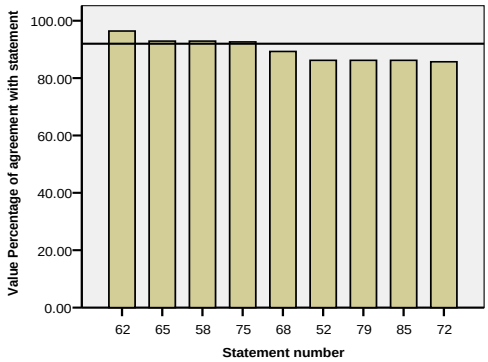
Section 3: Other issues related to centralization



Other issues:

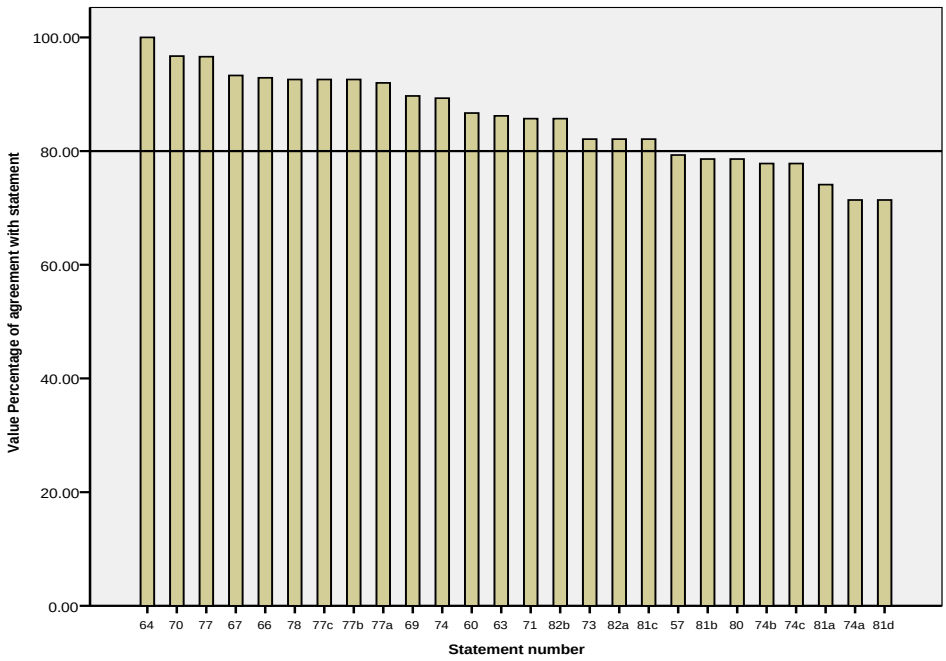
49. Centralizers have a good prognosis; 50. Centralization may be a stronger prognostic factor than psychosocial variables; 51. Centralizers have an internal disc disruption

Section 4: Future research



Areas for future research:

52. Operational definition for centralization; 58. Diagnostic accuracy of centralization; 62. Centralization and prevalence; 65. Centralization and course/ prognosis; 68. Prescriptive validity of centralization; 72. Centralization and subgroups; 75. Centralization and the cervical spine; 79. Reproducibility; 85. Education and training



Future research questions:

57. Mechanism causing the centralization phenomenon (including physiological mechanism); 60. Cross validation of diagnostic accuracy in different patient populations and examiners; 63. Expected rates of centralization in clinical settings for acute and chronic patients with LBP; 64. Effect of training and procedures in the prevalence and outcomes of centralizers; 66. Do centralizers have a favourable course when left untreated (natural history)? 67. Clinical response versus natural course of centralizers; 69. Do centralizers benefit from interventions other than directional movement exercise and which intervention provides the greatest benefit? 70. Comparison of outcomes of non-responders to centralization for various interventions; 71. Centralization and outcomes (i.e. ability to return to work, psychosocial outcomes, economic outcomes, health care utilization, QoL, recurrences and ability to self-manage ; Cost effectiveness of the McKenzie assessment prior to disc surgery, injections, imaging or any other treatments;73. Clinical predictors (CPR) of patients responding (or not) to directional preference exercises (including: Effect of patient compliance and attitudes of centralizers on outcomes; Effect of severe disability and psychosocial distress on the diagnostic, prognostic and treatment characteristics of centralizers; Centralization versus other prognostic factors); 74. Are centralizers and other clinical subgroups (e.g. patients satisfying CPR for manipulation or trunk stabilization treatment or candidates for disc surgery) made up of the same patients? How often do centralization findings co-exist with other findings? (I.e. relationship of centralization with other variables e.g. psychosocial, clinical findings e.g. segmental provocation signs, sciatica and EMG findings). How soon will positive EMGs become normal after centralization has been achieved and maintained? 74a. Are centralizers and other clinical subgroups (e.g. patients satisfying CPR for manipulation or trunk stabilization treatment or candidates for disc surgery) made up of the same patients? 74b. How often do centralization findings co-exist with other findings? (I.e. relationship of centralization with other variables e.g. psychosocial, clinical findings e.g. segmental provocation signs, sciatica and EMG findings); 74c. How soon will positive EMGs become normal after centralization has been achieved and maintained? 76. Centralization and contained cervical disc pathology; 77. Reliability of detecting, prevalence and outcomes in neck pain; 77a. Reliability of detecting centralization in neck pain; 77b. Prevalence of centralization in neck pain; 77c. Outcomes of centralization in neck pain 78. Predictive validity of centralization in the management of neck pain; 80. Stability/ reversibility of the centralization phenomenon; 81a. Effect of the clinician characteristics on the reliability of centralization; 81b. Effect of the patient characteristics on the reliability of centralization; 81c. Effect of standardization of the test on the reliability of centralization; 81d. Effect of knowledge of the test on the reliability of centralization; 82a. Role of history in predicting the presence of centralization; 82b. Role of and examiner's training in predicting the presence of centralization

---

This is the end of the feedback report for Round 3.

Please feel free to send your comments back to us!

Thank you for all your time and patience with this process

Appendix 4.11 Final comments made by the Delphi participants.

'Thank you for the elaboration of the material ...'
'…I found the process interesting and challenging – it has made me really think about centralization in perhaps a more in-depth way than I ever have previously. Thank you for the opportunity to participate in the study.'
'…Interesting results from your study. What I think is particularly interesting is viewing the process that it appeared at the outset than many participants had not considered beyond standard definitions and as the process progressed it appears that some people considered information and appeared to expand their definitions/ concepts…'
'… When the paper(s) is/are published I would greatly appreciate pdf copies please…'
'…Very nice study ☺ -should have no problem with publication. Some final thoughts: The results do not surprise me in respect to 1) centralization can be defined by change in pain intensity (80% agree) and 2) change in pain location does not require objective measurement 74% agree). Despite numerous papers and a systematic review supporting the reliability and discriminant and predictive validity of a stricter operational definition for CEN, clinicians do not find the systematic review's recommended definition for CEN clinically useful. Encouraging physical therapists and other clinicians to merge science with clinical experience (EBM) will always be a challenge. Until there are larger multi-clinic studies validating and comparing different operational definitions for both CEN and directional preference I doubt we will progress with this topic nor standardise an acceptable operational definition for both clinicians and researchers …'
'…What a job that was! I found the summary very interesting. There are still a range of opinions out there for sure, and as always more questions than answers. There was even some "comedy". I had a good chuckle at Section 3: Other issues related to centralization', Statement 51 comments included a WIDE range of opinions from: "This is supported …., There is currently evidence …., more evidence is needed …, I disagree strongly …. potentially reduces the credibility of the concept of centralization, I don't know, I don't care. Is it just me, or is that not funny? Good luck writing this up and I hope you are successful in publication.

Appendix 4.12 Operationalisation of centralisation.

Domain	*Relevant literature*	*Delphi support*	*Delphi panel's comments*	*Relevant research issues*	*Delphi support*	*Delphi panel's comments*	*Author's comments*
A. General definition							
1. The progressive and stable reduction of the most distal pain towards the spinal midline in response to standardised repeated end-range movement or sustained loading testing procedures. Testing may involve multiple directions and various starting positions, progression of forces or alternative forces (e.g. clinician assistance) as well as more than one visit		**80%**	R2 I consider this phenomenon only as regards to pain Pain most usually centralises rather than other types of symptoms Is pain the only symptom? Centralisation should focus on end-range patient movements and their effect on symptoms, including pain. However, should not be limited to pain. The definition needs to state that the symptoms (usually pain) recedes proximally toward the midline of the spine The words 'movement' or moving should be used with caution because to those unfamiliar with the concept it may be misunderstood as the creation of new symptoms. Clarification that the more proximal pain is present from the start should be made. There is both a process of centralisation and an end-point defined as 'centralised symptoms'; Lasting/ persisting improvement is yet another characteristic of the underlying pain generator that also affects patients' prognosis but is not required for pain to centralize. Lasting improvement requires a stability of the centralizing change and should not be required in the definition of centralization Centralization is a dynamic phenomenon: symptoms are sequentially going to the spine, abolishing from their more peripheral position they are usually not produced in every part of the limb (for example a patient who has leg pain at the beginning and have back pain only at the end... and the pain has not gone to the thigh). Cardinal points are: it applies to symptoms, it is the effect of repeated movements or static loading (both at end range), it remains after, is a phenomenon that appears usually quickly; one week? R3 Signs should not be considered Stability and change should be defined better Potential confusion with natural course or non-specific element	a. Operational definition for centralisation b. Head to head comparisons of different operational definitions c. Clarification of the term centralization i.e. number of repetitions, type of change, magnitude of change (e.g. belt line pain), duration and timeframe of required change, necessity of provocative testing for central symptoms	**86.2%** **58.6%** **55.2%**	R2 a. A new definition is unnecessary. McKenzie has provided an operational definition Coming to some sort of agreement by this method will most likely produce a product on which none of us can agree whole heartedly A good operational definition is not as critical to clinically managing patients as it is to performing research b. What would be the purpose? c. McKenzie has provided clarification R3 No comments were made in this round	Operational definition should be tested. Variations in criteria should be explored.
B. Population for whom the test is intended							
1.Centralisation can be best appreciated only in patients who demonstrate referred or radiating spinal symptoms	Werneke et al., 2008	**86.7%**	R2 Centralization does not occur with central symptoms only Radicular pain does not centralize - somatic referred pain does	a. Centralisation and prevalence b. Effect of patient	**96.4%** **78.6%**	R2 a. The CP and directional preference are very	Combine with item J1. Examine spinal

			R3 Best and only contradict each other Problematic with central pain CPR for manipulation	characteristics on reliability of centralisation		common in LBP cases (acute or chronic) and research is already available	patients regardless of symptom distribution Explore further in pilot
2.Centralisation can be best appreciated in patients with low back pain and patients with symptoms originating from other spinal areas	Studies involving patients with neck pain: Cleland et al., 2006a; Piva et al., 2006 (reliability) Werneke et al., 1999 (reliability & prognostic power) May et al., 2008a (prognostic power) Kjellman & Oberg, 2002(RCT)	**0%**	R2 Though anecdotally reported, not aware of any peer review studies suggesting centralization in other areas of the spine R3 No relevant comments were made for this item	a. Reliability of detecting, prevalence and outcomes in neck pain b. Centralisation and the cervical spine	**96.6%** **92.6%**	R2 a. Some data exist R3 No comments were made in this round	Test definition in the cervical spine
3. Before testing, patients should be screened for the presence of red flags (i.e. serious pathology indicators)	Reviews of international guidelines: Koes et al., 2001 Staal et al., 2003	**96.7%**	R2 Screening for red flags is mandatory to all patients prior to the initiation of any testing or treatment Screening for red flags should be done on all patients, but this does not preclude some provisional testing for centralization, though a referral is always warranted Absence of centralization supports the need for further investigation. R3 Minimal dangers associated with testing procedure Patients with red flags should not be excluded from testing unless they can not perform the movements If a patient is not appropriate for mechanical therapy, they will not centralise				History taking and physical examination to exclude red flags Use red flags before testing for centralisation and confirm with physical examination
4. Before testing, patients should be screened for the presence of yellow flags[1] (e.g. Waddell signs)	Reviews of international guidelines: Koes et al., 2001 Staal et al., 2003 Studies on centralisation and yellow flags: Karas et al. 1997; Werneke et al., 1993; Werneke & Hart, 2005 (prognostic power) Laslett et al., 2005 (diagnostic power);Christiansen et	**32.3%**	R2 Yellow flags should be considered on all patients, but this does not affect the desirability of their assessment for centralization Waddell's testing should be performed only after a thorough and unbiased clinical exam including testing for centralization. Waddell's signs may improve when the patient can effectively control pain and gain a sense of hope over their condition. R3 Psychosocial factors can have an impact on musculoskeletal pain and may confound the clinical interpretation of physical findings Prefer FABs to Waddell signs				Variation in how to assess yellow flags, optimal timing and specific tools for identifying these factors. Lack of recommendations on specific course of action once yellow flags are identified Consider yellow flags as a supplement and not an exclusion criterion in spinal

[1] Yellow flags 'indicate psychosocial barriers to recovery that may increase the risk of long-term disability and work loss' (New Zealand Guidelines Group, 2004, p.4)

	al., 2009		Research has shown that physical therapists do not fully understand how to integrate and evaluate the psychosocial domains with the physical domain during the clinical examination of patients with low back pain				assessment
5. The centralisation phenomenon becomes more complex in chronic cases with an adherent nerve root involved. In such cases, the close monitoring of the symptoms and potential postural deformities is required and patients are treated on an individual basis	Prevalence rates in acute & subacute cases 52-77%: Kopp et al., 1986 Delitto et al., 1993 Erhard et al., 1994 Werneke et al., 1999 Prevalence rates in chronic cases 21-87%: Long, 1995;Donelson et al., 1997; Kilpikoski et al., 2002; Laslett et al., 2003; Laslett et al., 2005; Laslett et al., 2006b	**36.7%**	R2 This applies to all patients, not just those tested for centralization Chronic cases are usually complex regardless of the assessment done Complexity is reduced if CP is found Little utility in evidence based clinical decision making since there is not a method that has been validated in the medical literature to reliably identify patients with an adherent nerve root R3 Close monitoring and individualised treatment would apply to all patients anyway No reliable/ valid method for diagnosing an adherent nerve root If a patient has an adherent nerve root, there is no centralisation	Expected rates of centralisation in clinical settings for acute and chronic patients with LBP	**86.2%**		Chronic cases may be more prone to presenting atypical responses? Include patients regardless of symptom duration Explore symptom behaviour in acute and chronic cases
C. Test users							
1. Centralisation should be recognised by the average clinician rather than requiring years of training	UK Delphi study on important clinical examination items: McCarthy et al. 2006 Reliability of judgements on symptom status change: Fritz et al., 2000	**80%**	R2 The literature supports this statement There is evidence of no significant difference for years and between practitioners and students The average clinician should recognize centralization when it occurs but will not have the skills to elicit centralization in many cases Should not require extensive training to recognize once a suitable definition is determined The average clinician should be trained in identifying CP at the undergraduate level, just as assessment for evidence of a neurologic deficit should be part of basic education The average clinician should have training and experience. This is not specialized knowledge or only gained through courses in one method Can not answer question unless “average” and “years” are defined Some training and clinical practice is required as with any of our skills, this is best available in continuing education courses What is essential is the time “practicing” the skill in the clinic. Good instruction only starts the process R3 To be useful and widely used it should be recognised by average and entry-level clinicians We do not have a definition of what “average” is. More research is needed to clarify the issues around training and	a. Effect of training and procedures in the prevalence and outcomes of centralisers b. Education and training c. Reproducibility d. Role of examiner’s training in predicting the presence of centralisation e. Case studies f. How do we best educate healthcare professionals that abolishment of leg or arm symptoms in a patient with central or foraminal stenosis is not a “centralizer”, but a separate subgroup of their own?	**100%** **86.2%** **86.2%** **85.7%** **55.6%** **48.3%**	R2 f. This is a question which cannot be answered by the choices available Centralization as a result of a single direction of repeated test movements (lumbar or cervical) is still centralization, regardless of the imaging findings of central or lateral stenosis. There are patients with a pseudoclaudication history whose pain centralizes with flexion or extension and they find their walking tolerance improves as a result. No one has data on this but it is a good area of research. Until that data is	Single movements advocated in UK Delphi study rather than more complex procedures. Reliability of judgements not significantly different across students / professionals or according to experience but this was not formally tested and only reflects discrepancies in interpretation Explore further in pilot study

			experience Depends on whether clinicians have sufficient education in entry level physiotherapy programs Post-graduate training may be required			collected and analyzed, we can only provide education based on our anecdotal experience, but that does not include teaching that stenotics cannot centralize their pain R3 b. No answers for question [number]	
2. Training of examiners is essential in order to have consistent results among clinicians	Reliability studies using trained examiners (>0.7): Werneke et al., 1999 Wilson et al., 1999 Fritz et al. 2000 Kilpikoski et al., 2002 Reliability studies using trained examiners (<0.7): Kilby et al., 1990 (more relaxed definition) Fritz et al., 2006 (greater time interval between examinations)	**86.7%**	R2 Training is essential to improve inter-examiner reliability Training and experience is less important than clear and agreed upon definitions and reduction of complexity for consistent results among clinicians The literature supports these choices ; See Kilby et al 1990 R3 Standardisation (e.g. specificity and clarity of definitions) is required to achieve the same shared baseline knowledge of procedures Consistency of results might depend on the quality, level and amount of training and experience Prior research has demonstrated that appropriate training is needed in order to have consistency and reliability Current research does support training but there is no research on the level of experience Training is required, and should be at undergraduate level. All physiotherapists should know how to do the test procedures and interpret the patient's responses	a. Effect of clinician characteristics on reliability of centralisation b. Effect of knowledge of the test on the reliability of centralisation	**74.1%** **71.4%**		Examiners should be trained in procedures
3.The experience of examiners is essential in order to have consistent results among clinicians	Reliability studies using experienced examiners: Kilby et al., 1990 Fritz et al., 2000	**44.8%**	R2 Training is necessary for consistency but experience is not necessary There is evidence of no significant difference for years and between practitioners and students Training and experience is less important than clear and agreed upon definitions and reduction of complexity for consistent results among clinicians The literature supports these choices ; See Kilby et al 1990 R3 Training is important for consistency, experience is not Experience is desirable and improves matters, but we all start out as novices Consistency of results might depend on the quality, level and amount of training and experience				Experience not a prerequisite for centralisation testing, but explore further in pilot

D. Setting						
				Expected rates of centralisation in clinical settings for acute and chronic patients with LBP	**86.2%**	
E. Tools used for documentation						
1.The pattern of questioning and documenting patient responses i.e. pain intensity and location should be highly standardised in clinical practice and research		**76.7%**	R2 Standardization is essential to ensure reliability, but statements [E2] and [E3] are unnecessary This may apply to research protocols but not clinical practice Not sure what you mean by 'highly', but standardization helps R3 The word 'highly' should be defined or excluded Standardisation may help, but we do not know. Standardisation does not necessarily have to be prescriptive. It may just imply consistency and clarity of reporting method Failure to recognise the pain distribution and intensity prior to test manoeuvres is common among clinicians, therefore standardisation of assessment is important Research may require standardisation but with different standards depending on the research question Standardisation is essential for some research studies in order to be replicated Routine clinical work may require less standardisation In research, a researcher may choose highly standardised procedure different from McKenzie 's recommendations for the purposes of determining if another method is superior	Effect of standardisation of the test on the reliability of centralisation	**82.1%**	Use body diagrams and numerical scales to describe location and intensity of symptoms in pilot. Quality of symptoms should also be reported, to distinguish pain from other symptoms. Consider testing 'body diagrams v. oral reporting' in VideoNeck study
2.For changes in pain location, a clear overlay numeric template should be used	Use of overlay template: Werneke et al., 1999	**16.1%**	R2 Template for research purposes but not required for clinical assessment For changes in pain location, a categorical response option should be given (e.g. foot, below knee/not foot, etc.) Perhaps in research protocols but not routine clinical practice R3 Too prescriptive for clinical practice, and unnecessary For research purposes only The evidence and a recent systematic review on centralisation support a measurement tool to document centralisation. Perhaps one of several reasons for the large variance in the prevalence rates of centralisation across studies is the lack of a standardised documentation process This is an objective measurement tool and physical therapy guidelines encourage therapists to use objective measurement procedures Pain overlay template is one possibility however, there are			Explore reliability and errors when not using an overlay template Explore sources of prevalence variation Consider comparing methods in pilot study

			other ways especially if using computer based assessment	
3. The area of the most distal symptoms should be marked and reassessed using a measurement from a bony landmark. The determination of the most distal pain should also be confirmed through palpation	Reviews including palpation investigations: Hestboek and Leboeuf-Yde, 2000 (reliability & validity); Najm et al., 2003 (validity); Seffinger et al., 2004 (reliability); van Trijffel et al., 2005 (reliability); Hollerwöger, 2006 (reliability); May et al., 2006 (reliability); Stochkendahl et al., 2006 (reliability); Myburgh et al., 2008 (reliability)	**0.00%**	R2 Precise measurement not required in clinical exam, only in research Palpation cannot determine the location of the most distal pain Palpation is unreliable and invalid This would be time consuming Unclear Is it spinal or peripheral palpation? The underlying structures, not being involved, shouldn't be painful Palpation has never been documented as part of the centralization testing R3 Too prescriptive for clinical practice and unnecessary Palpation has no validity	Do not use palpation to confirm area of the most distal symptoms
F. Type of loading strategy				
1. Movement testing for centralisation should not be based on single movements. Test movements must be performed repeatedly, to the fullest of the patient's available end-range	Studies using single movements: Cleland et al., 2006a; Piva et al., 2006; Fritz et al., 2006 (reliability) Flynn et al., 2002; Cleland et al., 2007 (prognostic power)	**86.7%**	R2 Single may not be sufficient to evaluate True, but ignores static positioning Repeated movements may not be necessary, but I can't think of an example R3 No comments for this item were made in this round	Perform movement testing to possible end-range Include single movements for safety but not for determining centralisation judgements
2.Centralisation can be elicited by repeated end-range movement testing only		**3.4%**	R2 Leaves out other interventions R3 No comments for this item were made in this round	Not considered adequate for all patient presentations
3.Centralisation can be elicited by repeated and/or sustained end-range movement testing		**60.7%**	R2, R3 No comments for this item were made in this round	Acceptable testing procedures but not adequate for all patient presentations
4.Centralisation can be elicited by repeated and/or sustained end-range movement testing only	Studies using other procedures: Werneke et al., 1999 Cleland et al., 2006b Tuttle, 2005; Tuttle et al., 2006	**10.3%**	R2 This sign should appear only in response to repeated movements or sustained positioning at end-range R3 No comments for this item were made in this round	Acceptable but not considered adequate for all patient presentations
5. By single and repeated movements. Sustained movements are used with caution to prevent confusion with instability	Clinical prediction rule for stabilization: Hicks et al., 2003 (reliability); Hicks et al., 2005 (predictive power)	**14.3%**	R2 Definitions of centralisation should not involve static postures or therapist overpressure because this confounds with other competing diagnoses Centralization should be mutually exclusive from sustained and therapist overpressure	Acceptable but not adequate for all patient presentations

			Single movements don't tell much and I don't understand what is meant by instability R3 Centralisation should be distinct from instability and manual techniques, therefore should avoid sustained movements and therapist overpressure		
6.Centralisation can be elicited by repeated and/or sustained end-range movement testing, overpressure by patient or therapist or manual techniques executed by therapist if appropriate	Werneke et al., 1999	**90%**	R2 Disagree with manipulation, but agree with overpressure or manual traction with the cervical spine Depends on definition of traction. Manual traction can be part of the therapist intervention, but motorized traction should not be included. Manual techniques i.e. manipulation/mobilization are not included in the definition R3 It is essential to allow for manual overpressure by the therapist because some patients can not achieve the appropriate end range for many different reasons. In the case of correction of a lateral shift, self correction by the patient often fails, but manual shift correction causes the centralisation phenomenon easily I think a good compromise would be to put a period after the phrase "over pressure by patient to therapist". And delete the manual techniques wording		Consider repeated and/or sustained end-range movement testing, overpressure by patient or therapist or manual techniques executed by therapist as acceptable procedures for eliciting centralisation in pilot
7. Regardless of testing methods used. This can involve test movements or after immediate treatment application, or post treatment over time		**17.9%**	R2 This statement is unclear Depends on purpose of the testing i.e. initial decision making or determine response to treatment. E.g. I use repeated movements during initial examination to test, but would look upon a centralization response after treatment as favourable, regardless of the treatment used. R3 I have no idea what this statement means or what it refers to		Consider distinction between centralisation as a physical sign and centralisation as an outcome measure
8.Movement testing for centralisation should not include more than 10 repetitions for each test movement		**20%**	R2 No comments were made in this round R3 Does not include sustained positions		Perform more than 10 repetitions if appropriate
9.Movement testing for centralisation should include repetitions that are continued long enough to ensure that the status of the symptoms has changed and/or is clear		**93.3%**	R2 May require several sets of exercises and more than one treatment session Some cases with relatively minor pain will require many repetitions 30-50 even. If clear evidence of CP is found with 5-10 movements and it is a reversible phenomenon, then further testing is not required. The number of repetitions or the duration of sustained loading is chosen on a case by case basis depending on the pain response, severity of pain, age		Perform movement testing until a clear response is elicited

			and general condition of the patient, patient willingness etc. R3 The exact number of repetitions is not important, the pattern recognition is				
10. Movement testing for centralisation should not necessarily be performed in a specific order. The sequence of movement testing does not affect the outcome or the ability to detect centralisation	Donelson et al., 1991	**40%**	R2 It might, if one has aggravated the symptoms It is not necessary to perform tests in a specific order to assess centralization. The order of testing can, however influence the results Changing the order of testing can be done for a number of reasons based on a clinical reasoning process. Using clues from the history may minimize the testing needed. If symptoms are vague or minimal provocative testing may be needed so that deduction/centralization can be better observed. Provocative testing can be skipped when the history is very clear or if the patient is very acute or pain is high R3 This is supported by published data (Donelson et al. in Spine vol 16 1991) We do not know, it has not been studied In some instances e.g. patient with a lateral shift, the order of movements is important but not in all cases	Role of history in predicting the presence of centralisation	**82.1%**	Don't know what you mean. Predicting prior to examination?	Perform sequence of testing according to patient presentation
G. Planes and directions of loading strategy							
1.Testing for centralisation should involve only the sagittal plane		**3.6%**	R2 If not elicited with sagittal plane motions the patient may be served better with a different approach Non-sagittal testing is not required if centralization is elicited with sagittal plane testing Sagittal is primary, but may need to pivot pelvis asymmetrically R3 No comments were reported for this item in this round				Sagittal plane movements inadequate for all patient presentations
2.Testing for centralisation should involve the standard planes of movement available to the spinal region involved		**33.3%**	R2 Confusing statement R3 What are the standard planes of movement?				Standard planes not always adequate
3.Testing for centralisation should involve the standard planes of movement available to the spinal region and/or a combination of movements if appropriate		**93.1%**	R2 Confusing statement A full examination often only requires standard planes of motion, but combinations of planes are required for maybe 30% of cases in order to identify the CP R3				All spinal planes & combined movements can be considered in testing

			Side gliding is not considered a standard plane of movement by many, but an essential inclusion in the lumbar spine. Retraction in the cervical spine suffers the same divergence of opinion and is very important in eliciting centralisation This statement does not include sustained positions	
H. Criteria for a positive test				
1. When defining centralisation, changes in the intensity of symptoms should not be considered	Studies using intensity as well as location as a criterion for a positive test: Karas et al., 1997 Laslett and van Wijmen (1999); Laslett et al. (2005)	**20%**	R2 Intensity of pain is an important component, not just pain location Changes in intensity are the starting point indicating when to continue testing. It might be too early to use the term "centralization" YET, but keep testing because a decrease in intensity is commonly seen before the symptom actually abolishes or moves to a more proximal location. R3 The clinician should differentiate between a directional preference and centralisation. Both terms overlap but these terms are not synonymous Further work is required to operationally define these terms to decrease confusion between clinical identification of directional preference and/or centralisation. Without this important clinical discussion and research confusion will continue regarding how best to define centralisation	Changes in intensity should be considered when defining symptom response groups. Explore in pilot
2. When defining centralisation, changes in neurological status (symptoms or signs) should not be considered	Studies including neurological status in the criteria for a positive test: Delitto et al., 1993 Delitto et al., 1995 Fritz, 1998 Fritz et al., 2000	**12.9%**	R2 Some definitions demonstrate a failure to differentiate between signs and symptoms Neurologic status should always be monitored. Occasionally a reflex will return or disappear during testing and this is important R3 No comments were made in this round	Consider neurological symptoms or signs where appropriate
3. Nonspinal conditions or underlying disease states potentially causing or influencing symptoms should be excluded or taken into consideration	Werneke et al., 2005 (presence of non-centralisation associated with some behavioural signs) Laslett et al., 2005 (diagnostic power reducing in the presence of psychosocial distress)	**87.1%**	R2 Unclear context One should always monitor and consider co-morbidities Centralization is an indication related to mechanical spine issues, not disease More than one factors in statement Underlying disease states should be excluded, and non-spinal conditions should be taken into account R3 I'm not sure nonspinal conditions/ disease states are criteria for a positive test. They definitely need to be considered as part of complete evaluation	Co morbidities or other e.g. psychosocial issues should be reported as part of the assessment process but not considered when defining criteria for positive test
4. Underlying disease states potentially causing or influencing symptoms should be excluded	Guidelines on diagnostic triage: Koes et al., 2001	**67.7%**	R2 Underlying disease states should be excluded, and non-spinal conditions should be taken into account	Not always possible because most spinal pain has a non-specific nature

			R3 I'm not sure nonspinal conditions/ disease states are criteria for a positive test. They definitely need to be considered as part of complete evaluation			Co morbidities or other e.g. psychosocial issues should be reported as part of the assessment process but not considered when defining criteria for positive test
5. Nonspinal conditions potentially causing or influencing symptoms should be taken into consideration	Guidelines on diagnostic triage: Koes et al., 2001	**87.1%**	R2 Underlying disease states should be excluded, and non-spinal conditions should be taken into account R3 I'm not sure nonspinal conditions/ disease states are criteria for a positive test. They definitely need to be considered as part of complete evaluation			Co morbidities or other e.g. psychosocial issues should be reported
6. The reduction of peripheral symptoms towards the spinal midline should be progressive and stable. Distal pain which simply abolishes rather than progressively receding toward the spine should not be considered in the definition	Young et al. 2003	**22.6%**	R2 Basically correct but still debatable to some extent Stability is required only for a good prognosis but should not be required in determining the presence of centralization Instability may be underlying but does not necessarily contraindicate in the acute stage That's what McKenzie says Ideal if changes are preserved over time, but looking for longer periods of centralisation and stabilisation of the condition are also important Reference supporting this statement is Young S, et al. Correlation of physical examination characteristics with three sources of chronic low back pain. The Spine Journal 2003; 3: 460-465 R3 More testing is usually needed	Stability reversibility of centralisation	**78.6%**	Combine with item A1.
7. Observed changes should be retained over time (lasting change after testing)		**86.2%**	R2 CP is primarily a spinal (disc) phenomenon (Laslett M et al. Eur Spine J 2006) Even transient evidence of centralization is useful in the determination of motion preference Retained assuming that the patient has been compliant and avoid the provocative posture or movements R3 Most appropriate for pure criteria for positive test Statement needs to be rephrased: over how much time and under what conditions? How long after testing? Does not apply to centralisation, but if referring to the process of centralisation or centralising in the direction of full centralisation, that is another topic			Symptom status should be established after testing Consider cues from the history and clinical examination potentially suggesting centralising symptoms

			It is just the ability of changes to predict longer lasting changes that are of primary concern. Otherwise one is at risk of a circular argument simply suggesting that if someone improves over time that they have improved over time				
8. When testing for centralisation, the response to testing may be obvious during the first examination or may require evaluation over a period of time to confirm the phenomenon	Werneke & Hart, 2003	**100%**	R2 May require several sets of exercises and more than one treatment session Should occur during initial exam but not over time Need to define "time" 3 days of testing OK, not three weeks. R3 These questions are hard to answer because they both combine the timeframes. I can disagree with both of them and still answer yes				The multiple-visit classification is more precise for discriminating pain & disability than the first-visit classification procedure. Combine with item A1 and H6
9. Symptoms may be sequentially and lastingly abolished at each session (pure response) or gradually abolished in a progressive manner but not sequentially (partial response)		**74.1%**	R2 Unclear statement The meaning of sequentially is not clear. Is this temporal or spatial or type of symptoms or distalness? "Lasting" abolition is not a requirement to identify centralization. R3 This statement is too complex and confusing These questions are hard to answer because they both combine the timeframes. I can disagree with both of them and still answer yes Sequentially should be defined Can't say I've seen a non-sequential abolition moving gradually toward complete abolition Centralisation rarely if ever takes more than a week to be identified, unless the patient is slow to adopt the postural changes and exercise regime between assessments I agree [with this statement] but it does not include all possibilities				Applicable for multiple visit definition Check prevalence and predictive value of one v. multiple visit definition
I. Safety							
1. The patient's status (improving or worsening) should be considered beyond symptom relief i.e. neurological picture, range of motion. For example, if the patient's symptoms are improving but his neurological status is compromised, centralisation should be considered an undesirable outcome for this patient.		**59.3%**	R2 Important not only what improves status but also the opposite should worsen status. One must differentiate between decrease in symptoms v. improvement of status e.g. stenosis where flexion may decrease symptoms and extension worsen status but on resumption of ambulation the symptoms recur thus not improved status Confusing statement Centralization is never undesirable; however worsening of neurological status would be considered failure to centralize. Neurological signs should be given priority over pain response Centralization is usually accompanied by improving neurological status. It has been shown that ROM improves with centralization not the opposite	a. Are centralisers and other clinical subgroups made up of the same patients? How often do centralisation findings co-exist with other findings? How soon will positive EMGs become normal after centralisation has been achieved and maintained? b. Potential	**89.3%** **69.0%**	R2 b. This is unlikely to occur. Patients whose pain will only peripheralise and not centralize have other things going on instead of primary discogenic pain (examples taken from my unpublished research data includes: gas filled disc, swollen dorsal	

Some patients have neurological signs that do not improve as pain centralizes but as long as the neurological status is not worse, centralization is desirable. I have never seen a patient who centralized and remained better whose neurological status deteriorated If neurological status is compromised at the beginning of the exam, then improving symptoms i.e. centralization should be pursued as a favourable sign There is no magic in centralization. It is best not considered as a separate entity, but is simply a means of assessing change in symptoms which is appropriately included as part of an overall framework or hierarchy of information used in assessment of symptoms R3 Unclear question. Was neural status compromised at the outset and is unchanged, or is it deteriorating as a result of the assessment? If the latter, one must cease, although again, that never happens in the face of centralisation. That's one of the wonderful safe guards about the MDT assessment. Certainly, monitoring neural status is paramount and deterioration cannot be accepted Neurological status should be monitored additionally e.g. if neurologic status is worsening and only pain is improving, this should not be considered a positive sign Question seems to me a purely "researcher" question. Never happened to me that centralisation were linked to a worsening of the general status True centralisation with concurrent worsening of neurological status must be vanishingly rare. I cannot recall a case in 30 years of doing this In this scenario centralisation is not the undesirable outcome – the worsening neurological status is the undesirable outcome and centralisation of symptoms is besides the point – this is an important distinction but not a good description of it. The point is that if an intervention is worsening the patient's neurological status then it may be inappropriate despite centralisation of symptoms. This appears to happen occasionally in patients with spinal stenosis – the little used 'pheasant sign' was reported to look for this phenomenon.	irreducible derangement i.e. pain moving centrally but neurological symptoms moving distally: at what point should patients be referred for further investigation?	root ganglion, end plate avulsion fractures) R3 a. While useful for prognosis, there has not been any indication in the line of research utilizing clinical prediction rules to determine the effectiveness of various treatment approaches that centralization of symptoms in the absence of neurological signs is a sign that should trump other factors (i.e. time in the manipulation CPR), particularly given different in effect sizes- symptoms distal to the knee seems to be important. These two in combination (symptoms distal to the knee and centralization of symptoms) seem to be likely candidates for patients that will respond best to a direction specific treatment approach more than other treatment approaches. Utilizing 'centralizers' as a stand alone subgroup does not seem to be the best approach; Question 74 contains 3 questions, at which

						am I supposed to answer? Questions 74 b and c: how is possible to answer with yes or no to questions starting with: "How"?	
2. Symptoms getting worse proximally but improving distally should be considered a positive sign		**100%**	R2 This is usually correct, but only if the peripheral pain rapidly abolishes and the increased proximal pain is not too severe or prolonged more than a few days There is usually an increase in proximal pain during centralization. However, patients who have chemically sensitized discs can have partial centralization with tremendous increases in proximal pain. Their pain does not remain centralized when the spine is loaded R3 No comment was made for this item				Monitor both distal and proximal symptoms but give priority to status of distal symptoms
J. Intervention							
1. Patients with distal symptoms above the knee meeting CPR for manipulation are most likely to benefit from manipulation regardless of their response to repeated movement testing	Clinical prediction rule for manipulation (based on single movements): Flynn et al., 2002 Childs et al., 2004a Fritz et al., 2005 Cleland et al., 2006c Independent evaluation of CPR: Hancock et al., 2008 Comparison of interventions for people whose symptoms centralise: Long et al., 2008a; Kilpikoski et al., 2009 Case study: May & Rosedale, 2007 Guidelines: Arnau et al., 2006 Koes et al., 2001	**22.6%**	R2 A description of the clinical prediction rule for success is not provided [in the questionnaire] This is unknown The clinical prediction rule was developed in comparison to stabilization exercise Most will respond to manipulation (Childs et al.) but unknown how patients' response to repeated movement testing will affect this. R3 Childs et al. Ann Int Med, 2004 have provided evidence for this statement and also good evidence that patients with symptoms distal to the knee are not as likely to respond to manipulation In a study by Browder et al. PT 2007, the treatment effect was dramatic in comparison to an extension-oriented treatment approach The manipulation CPR has not been thoroughly tested and we know of one published example that shows manipulation not to be the optimal treatment for a patient who fits the manipulation CPR characteristics Any benefit from manipulation would be overshadowed by the benefit of teaching self-care to centralise and abolish symptoms as well as empowerment for prevention of recurrences using directional exercises and posture modifications This statement is more relevant to the McKenzie method	a. Comparison of outcomes of non-responders to centralisation for various interventions b. Do centralisers benefit from interventions other than directional movement exercise and which intervention provides the greatest benefit? c. Prescriptive validity of centralisation	**96.7%** **89.7%** **89.3%**	R2 No comments were made in this round R3 a. "non-responders to centralization"? Makes no sense. Perhaps "non-centralizers"?	International guidelines do not recommend specific exercises for spinal pain Not all guidelines support manipulation for acute LBP Childs et al., 04a may imply a positive comparison with an ineffective treatment rather than an effective alternative; limited generalisability of results (Deyo, 2004)
2. Education of patients is	Koes et al., 2001	**90%**	R2				Patient advice and

essential following the use of these techniques so that movement is normalised as symptoms settle and patients do not develop fear of movement			This is a true statement, but what does it have to do with centralization This is a question likely to elicit false positive responses as who could possibly disagree with the need for patient education. To say that it is essential in this situation more than others or for all patients, however I don't think is warranted or necessary R3 This seems to speak to intervention, not examination I would agree, but it would appear to be a general principle and have little to do with concepts of centralisation Education is certainly essential. However, some therapists make the error of discontinuing the movement causing centralisation and the expected increase in proximal pain. The end result is that the wrong patient education is then provided				information plays an important role in most guidelines Address this issue in subsequent intervention
K. Attributes							
1. Centralisers have a good prognosis	Reviews: Aina et al., 2004 Wetzel & Donelson, 2003 Berthelot et al., 2007 Chapter 3	**96.7%**	R2 Stability is required only for a good prognosis but should not be required in determining the presence of centralization Research is clear on this statement R3 All three statements are true to some extent. However, those with severe psychological distress may require psychological intervention before being able to adhere to the self-treatment program required to centralise and abolish pain. Those with severe IDD (i.e. grade IV on the Dallas Disco gram Scale) or who have satellite fissures not connected to the nucleus of the disc may not centralise.	a. Clinical response versus natural course of centralisers b. Do centralisers have a favourable course when left untreated? (natural history) c. Centralisation and outcomes	**93.3%** **92.9%** **85.7%**	R2 a, b. How will you leave them untreated when they find out that they can positively affect their symptoms with simple movements? R3 No comments were made in this round	Investigate prognostic value Consider population & stability issues
2. Centralisation may be a stronger prognostic factor than psychosocial variables	Werneke and Hart, 2001; Long et al., 2008b	**65.5%**	R2 Unknown at this time I am not sure that it has been studied to the degree necessary to be able to make such a statement Research still in dispute for this statement R3 All three statements are true to some extent. However, those with severe psychological distress may require psychological intervention before being able to adhere to the self-treatment program required to centralise and abolish pain. Those with severe IDD (i.e. grade IV on the Dallas Disco gram Scale) or who have satellite fissures not connected to the nucleus of the disc may not centralise.	a. Predictive validity of centralisation in the management of neck pain b. Centralisation and subgroups c. Clinical predictors (CPR) of patients responding (or not) to directional preference exercises	**92.6%** **85.7%** **82.1%**	b. Too many variables	Examine relative importance to other established prognostic indicators
3. Centralizers have an internal disc disruption	Laslett et al. 2005 Hancock et al. 2007	**14.8%**	R2 At present, it appears to be so, but more evidence is needed This is correct in patients who are not severely disabled (Roland Morris category) or distressed (Distress Risk Assessment Method) see Laslett M et al Eur Spine J 2006. All undistressed and minimally disabled patients who	a. Diagnostic accuracy of centralisation b. Cross validation of diagnostic accuracy in different patient	**92.9%** **86.7%**	R2 a. Validation of the Laslett M et 2006 results is urgently required	

centralize satisfy ISIS criteria for internal disc disruption Grade 1-3 internal disc disruption (using the Dallas Disco gram scale) will centralize. Grade 4 will not Many centralizers have a herniated disc and sciatica. Many others likely have an internal disc problem. R3 All three statements are true to some extent. However, those with severe psychological distress may require psychological intervention before being able to adhere to the self-treatment program required to centralise and abolish pain. Those with severe IDD (i.e. grade IV on the Dallas Disco gram Scale) or who have satellite fissures not connected to the nucleus of the disc may not centralise. This is supported by Laslett M et al. TSJ 2005 There is currently evidence in the lumbar spine but not the cervical spine More evidence is needed in support of [statement] I disagree strongly with [statement] and think its inclusion potentially reduces the credibility of the concept of centralisation I don't know, I don't care	populations and examiners		b,c,d,e,f. No comments were made in this round
	c. Mechanism causing the centralisation phenomenon (including physiological mechanism)	**79.3%**	h. Is [statement number] a question? This is not a rare patient in my experience –it's common for a centralizing patient not to improve until they control their posture sufficiently to avoid worsening between exercise sessions R3 a, b, c, d, e, f, g, h. No comments were made in this round
	d. Criterion validity of centralization (e.g. using provocation discography as a standard; centralization as a tool for assessing the severity of a disc lesion; centralization as a tool for ruling out pathologies other than discogenic pain)	**69.0%**	
	e. Conceptual model for centralisation	**65.0%**	
	f. Is centralization an anatomical phenomenon? Centralization in relation to tissue response	**55.2%**	
	g. Centralization and contained cervical disc pathology	**55.2%**	
	h. The rare patient whose pain centralizes up to the lumbar spine but then remains unchanged and may even worsen with exercise. Generally improves over time and is usually initiated by long sitting time	**35.7%**	

Note:CP, Centralisation Phenomenon; CPR, Clinical prediction rule; FABs, Fear Avoidance Beliefs; IDD, Internal Disc Disruption; LBP, Low Back Pain; R2, Round 2; R3, Round 3; ROM, Range of Motion.

Appendix 5.1 Combined results from the systematic reviews.

ITEM	STATISTICALLY SIGNIFICANT ASSOCIATION WITH OUTCOMES (STRONG EVIDENCE)	RELIABILITY [LEVEL OF EVIDENCE, κ ≥ 0.4]
Changes in pain location with repeated lumbar motion testing	**Symptoms**	**Changes in pain location with repeated movement: κ = 0.15-0.70 [Conflicting]**
Changes in pain location or intensity with lumbar motion testing	**Work status**	**Changes in pain location or intensity with repeated movement: κ = 0.51-0.90 [Strong]**
ITEM	**STATISTICALLY SIGNIFICANT ASSOCIATION WITH OUTCOMES (LIMITED EVIDENCE)**	**RELIABILITY [LEVEL OF EVIDENCE, κ ≥ 0.4]**
Changes in neck pain location with manual therapy	**Symptoms**	**Directional preference in the cervical spine κ = 0.46 [Moderate]**
Changes in low back pain intensity with manual therapy	**Symptoms**	**Directional preference in the lumbar spine κ = 0.90 [Moderate]**
Pain during lumbar motion testing	Work loss	Pain during lumbar motion testing: κ = 0.10-1.00 [Conflicting]
Prone instability test	Disability	Prone instability test: κ=0.75-0.94 [Strong]
Pain during neurodynamic testing (lumbar spine)	Disability	Pain on SLR testing: κ =0.36-0.81 [Conflicting]
ITEM	**STATISTICALLY SIGNIFICANT ASSOCIATION WITH OUTCOMES (CONFLICTING EVIDENCE)**	**RELIABILITY [LEVEL OF EVIDENCE, κ ≥ 0.4]**
Changes in pain location with lumbar motion testing (single, repeated)	**Disability, Healthcare use**	**Changes in pain location or intensity with single lumbar movement: κ = 0.28-0.81 [Moderate]** **Changes in pain location with repeated movement: κ = 0.15-0.70 [Conflicting]** **Changes in pain location or intensity with repeated movement: κ = 0.51-0.90 [Strong]**
Pain during neurodynamic testing (lumbar spine)	Symptoms	Pain on SLR testing: κ =0.36-0.81 [Conflicting]
ITEM	**NO STATISTICALLY SIGNIFICANT ASSOCIATION WITH OUTCOMES (STRONG EVIDENCE)**	**RELIABILITY [LEVEL OF EVIDENCE, κ ≥ 0.4]**
Changes in pain location with repeated lumbar motion testing	Work loss	Changes in pain location with repeated movement: κ = 0.15-0.70 [Conflicting]

ITEM	NO STATISTICALLY SIGNIFICANT ASSOCIATION WITH OUTCOMES (LIMITED EVIDENCE)	RELIABILITY [LEVEL OF EVIDENCE, κ≥ 0.4]
Changes in pain location with single cervical motion testing	Perceived global change	Changes in pain location with single movement: κ = -.005 – 1.00 [Moderate]
Cervical compression testing	Symptoms, Perceived global change or satisfaction	No evidence
Cervical distraction testing	Symptoms, Perceived global change or satisfaction	No evidence
Pain during neurodynamic testing (cervical spine)	Perceived global change	No evidence
Changes in neck pain location with manual treatment	Symptoms (intensity), Range of Motion, Disability, Perceived Global change	Directional preference in the cervical spine: κ = 0.46 (0.43, 0.49) [Moderate]
Changes in neck pain intensity with manual treatment	Symptoms (location or intensity), Range of Motion, Disability, Perceived Global change	
Changes in pain location with repeated lumbar motion testing	Strength	Changes in pain location with repeated movement: κ = 0.15-0.70 [Conflicting]
Pain on spring testing	Disability	Pain on segmental mobility (including spring) testing κ = 0-0.67 [Strong]
Posterior shear test (lumbar spine or SIJ)	Disability	Posterior shear test (lumbar spine): κ = 0.20-0.51 [Moderate] Posterior shear test (SIJ): κ = 0.88 [Limited]
SIJ tests (Gaenslen test, Sacral thrust test, Resisted hip abduction, Compression distraction test, Sacral sulcus test, Patrick test)	Disability	No evidence on: Gaenslen test, Resisted hip abduction, Sacral sulcus test , Patrick test Sacral thrust test: κ = 0.52 [Limited] Compression: κ = 0.09-0,73 [Conflicting] Distraction test : κ = 0.11-0.69 [Moderate]
Pain on neurodynamic testing	Work loss	Pain on SLR testing: κ =0.36-0.81 [Conflicting]
Changes in pain intensity with physical examination	Symptoms, Disability	No evidence

Note: K, kappa statistic; SIJ, Sacro-Iliac Joint; SLR, Straight Leg Raise.

Appendix 5.2 Study information leaflet for the reliability study.

Symptom response pilot study: Reliability and role in predicting outcome in neck pain

Patient Information Leaflet

Version 3.0 - 12/03/2008

For further information please contact:
Ms Angeliki Chorti, Warwick Medical School,
University of Warwick, Coventry CV4 7AL
Tel: 024 761 50405

You are being invited to take part in this research being carried out as part of a PhD project at the Warwick Medical School. Before you decide whether you want to participate or not, it is important for you to understand why the research is being done and what it will involve. Please take time to read the following information carefully, and discuss it with friends, relatives and/or your GP if you wish. If there is anything that is not clear or if you would like to know about the study, please do ask and we will be happy to provide more information.

Thank you for reading this.

1. What is the purpose of the study?

We are looking at the usefulness of information deriving from the clinical examination and in particular, information related to the change in your complaints at your neck and shoulders. The study is recruiting adult employees through the Safety and Occupational Health Departments. All people who enter the study will receive advice about their condition and in particular, information on pain, self-management, exercises, and how to carry on with daily activities.

2. Why have I been chosen?

You have been chosen because you work for a company that is supporting the study, and are an employee who is experiencing pain in your neck, with or without symptoms at your shoulder and/or head. Approximately 50 people are being asked to take part.

3. Do I have to take part?

No. It is up to you to decide whether or not to take part. If you decide not to take part you will receive usual care from your occupational health service. This will not affect the standard of care you receive from the service in any way.

If you decide to take part, you will be asked to sign a consent form. A copy of this information sheet together with one signed copy of the consent form will be given to you to keep. You are still free to withdraw at any time and without giving any reason and this having no implications on your rights and benefits or the standard of care you receive in the service.

4. What will happen to me if I take part?

Usually, when you visit health professionals for a problem in your spine, they examine you with the aim of determining the nature and management of your condition. Often, the choice of the various questions and tests used to help clinicians make a decision rely on the clinician's background and not on scientific evidence. The way clinicians make their assessments and reach a conclusion may also differ across health professionals. In order to ensure the information in your assessment is useful and the results are interpreted in the same way across clinicians, we need to make comparisons. All patients will be assessed separately by two clinicians and the results will be cross - checked in order to identify where there is agreement or disagreement between them. This is called a reliability study.

Once the assessments are finished, you will receive information about your condition and how to manage it. We are also interested to see how your condition is progressing and whether/how it has affected you. For this reason, you will be contacted again over a 3-month period (please see below).

5. What do I have to do?

This leaflet has been given to you by the Safety Officer of your company. If you are interested in participating, an appointment will be arranged for a meeting. In this meeting, you will have the opportunity to ask further questions if you wish to. If you decide to take part, you will be assessed by our research clinicians. You will be asked to fill in two questionnaires and answer questions about your condition, and how it is affecting you, before your assessment and at 3-months after your assessment. In addition to the clinical questions, you will also be asked questions about your personal life (you can choose not to answer these personal questions if you wish to). If you consent to take part in this study, your GP will, with your permission, be notified of your participation in the study.

6. What are the possible disadvantages of taking part?

There are very few risks associated with this research. Occasionally, you might experience some increase in pain, but this is normal when people start to move again after a period of pain. Every step has been taken to keep any discomfort and inconvenience to a minimum.

7. What are the possible benefits of taking part?

We hope that the advice you get will help to improve your condition and/ or prevent it from happening again. However, this cannot be guaranteed. The information we get from this study may help clinicians assess future patients with neck pain more effectively and make more accurate decisions about the management of this condition.

8. What if new information becomes available?

Sometimes, during the course of a research project, new information becomes available that answers all the questions being asked. If this happens, you will be informed about it and discuss whether you want to continue in the study. If you decide to withdraw your occupational health department will make arrangements for your care to continue. If you decide to continue in the study you will be asked to sign an updated consent form. If the study stops for any other reason, you will be informed about the reasons for this action.

Also, on receiving new information your occupational health nurse/physician might consider it to be in your best interests to withdraw you from the study. He/she would explain the reasons and arrange for your care to continue.

9. What happens when the research study stops?

After the research study has finished, your occupational health nurse would continue to provide you with advice where necessary or when appropriate, to refer you on to other health professionals.

10. What if something goes wrong?

It is unlikely that you will experience problems by taking part in this research. If you are concerned about the assessment or treatment you received you should contact your occupational health nurse or physician straight away. If you wish to complain or have any concerns about any aspect of the way you have been approached or treated during the course of this study, the Safety Office or Warwick University can be contacted (details can be found at the back of this leaflet).

11. Will my taking part in this study be kept confidential?

All information that is collected about you during the course of the research will be kept strictly confidential. This information will be kept in a secure place and only people involved in the study will have access to it.

12. What will happen to the results of the research study?

The data collected will be analysed and the results will be used to write a research report and articles to scientific journals. Presentation of any findings will be made without using your real name or any details that could identify you. We can also send participants a summary of the findings on request.

14. Who is organising and funding the research?

The person responsible for this study is Ms Angeliki Chorti from the Warwick Medical School. This study is sponsored by the University of Warwick.

15. Contact for further information or concerns

Safety Office Contact:

Bill Leslie & Karen Lawrence

02476 88 7341

w.leslie@coventry.ac.uk & *karen.lawrence@warwick.ac.uk*

Study Co-ordinator:

Ms Angeliki Chorti, Warwick Medical School, University of Warwick, Coventry CV4 7AL

Tel: 024 761 50405

E-mail: *A.Chorti@warwick.ac.uk*

This is your information sheet to keep. If you require any other information please do not hesitate to contact Ms Angeliki Chorti at the above address. **Thank your for taking the time to read this leaflet**.

Appendix 5.3 Information leaflet for the VideoNeck study.

Symptom response identification in neck pain: a reliability study using videotaped assessments across multiple examiners (VideoNeck study)

PATIENT INFORMATION SHEET

We are asking you to take part in the second part of a research project on the usefulness of information deriving from the clinical examination. This study is undertaken as part of a PhD degree at the Medical School of the University of Warwick. Before you decide, it is important for you to understand why the research is being done and what it will involve. Please take time to read the following information carefully and discuss it with others if you wish. Please feel free to contact us if you would like more information or you have any concerns regarding this research. Take time to decide whether or not you wish to take part.

Thank you for reading this.

What is the purpose of this study?
This study is looking at the utility of various aspects of the neck pain assessment, this time across multiple examiners. We are comparing the findings and the clinical opinions of different examiners based on the videotaped assessments of patients with neck pain.

Why is the study being done?
Neck pain is a very common condition and therefore, any clinical information that contributes to the decisions made about diagnostic practice and effective treatments is very important. Pain and symptom response to spinal testing have been argued to be important clues in the assessment of spinal conditions. However, limited evidence exists regarding their utility in neck pain or the influence of clinicians' characteristics on the assessment findings.

Why have I been invited to participate?
You have been chosen because you are experiencing pain in your neck, with or without symptoms at your shoulder and/or head.

Do I have to take part?
It is up to you to decide whether or not to take part. If you do, you will have to sign a consent form for this study. You are still free to withdraw at any time and without giving a reason. This decision will not affect you or your rights in any way.

If you consent to take part in this study, your GP will, with your permission, be notified of your participation in the study. Also, on receiving new information your GP and/or your occupational health nurse/physician might consider it to be in your best interests to withdraw you from the study. He/she would explain the reasons and arrange for your care to continue.

What will happen to me if I take part?
Your neck assessment by our research clinicians will be videotaped and the videotapes edited so that your identity and any personal information you do not wish to be disclosed is concealed. Access to the edited videos will be granted to approximately 35 participating health professionals to provide their opinions about your condition.

What do I have to do?
You will need to complete, sign and return the consent form to our research team in order for us to have written permission for the videotaping of your neck assessment.

What are the possible benefits of taking part?
There are no direct benefits to you from taking part in this study. However, we are hoping that the data collected will result in providing information about and contribute to the standardisation and improvement of neck pain assessments.

What are the possible disadvantages of taking part?
Every effort has been made to keep any inconvenience or risk to the minimum.

Will my taking part in the study be kept confidential?
All information which is collected about you during the course of the research will be kept strictly confidential. This information will be kept in a secure place and only people involved in the study or authorised individuals will have access to it.

What happens when the research stops?
The data obtained will be used for internal publication for a PhD Project and submitted for assessment with a view to being published in scientific journals/ conferences. We can also send participants a summary of the study results on request.

What if there is a problem?
In the unlikely event that there is a problem and you wish to complain about any aspect of the way you have been approached or treated during the course of this study, the normal University complaints mechanisms are open to you. If you do have any complaints or you do not wish to continue this study, please contact Ms Angeliki Chorti (Tel: 02476150405, Email: A.Chorti@warwick.ac.uk) or Dr. Chris McCarthy, on (Tel: 02476575856, Email: C.J.McCarthy@warwick.ac.uk).

Who has reviewed this study?
The study has been reviewed by the Biomedical Research Ethics Sub-Committee of Warwick University.

Contact details
If you would like any further information please contact:

Angeliki Chorti
Warwick Medical School
University of Warwick
Gibbet Hill Road
Coventry, CV4 7AL

Email: A.Chorti@warwick.ac.uk
Tel: 02476150405

Thank you once again for taking the time to read this information.

Section 4: Assessment

Appendix 5.4 Selected sections from the physiotherapist manual.

For the assessment of patients, a standardised form will be used. Although the form indicates a specific order of testing, the assessment should still be guided by each individual's presentation.

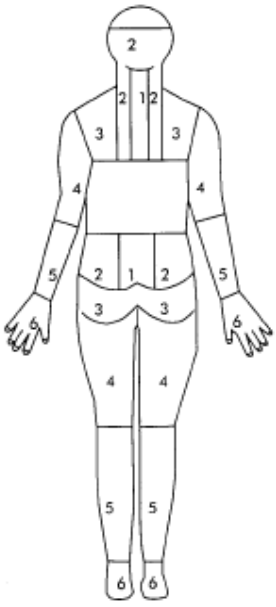
A version of the assessment form as it will be used in the study is provided below. Examples of the assessment form, with some additional notes to assist in its completion are provided in Appendix 2.

4.1 The Assessment form

History

Current Date:

ID number: Date of Birth:



Present Symptoms _____

Present since _____ *Improving / Unchanging / Worsening*

Commenced as a result of _____ *Or no apparent reason*

Symptoms at onset: *neck / arm / forearm / headache* _____

Constant symptoms: *neck / arm / forearm / headache* Intermittent symptoms: *neck / arm / forearm / headache*

Worse *bending* *sitting* *turning* *lying / rising*

am / as the day progresses / pm *when still / on the move*

other _____

Better *bending* *sitting* *turning* *lying*

am / as the day progresses / pm *when still / on the move*

other _____

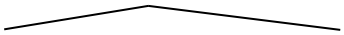
Disturbed Sleep *Yes / No* Pillows _____

Sleeping postures *Prone / sup / side R / L* Surface *Firm / soft / sag*

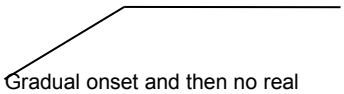
Previous Episodes 0 1-5 6-10 11+ Year of first episode _____

Previous History _____

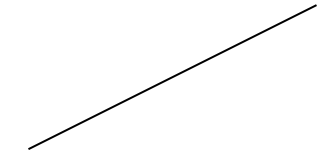
Course of symptoms until now

O 

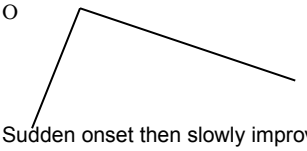
Gradual onset and then slowly improving but it is still a problem

O 

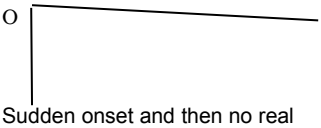
Gradual onset and then no real improvement

O 

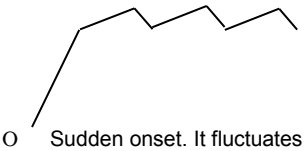
It is getting worse since it started

O 

Sudden onset then slowly improving but it is still a problem

O 

Sudden onset and then no real improvement

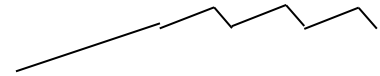
O 

Sudden onset. It fluctuates in intensity but never goes away.

O 

Recurrent episodes of the same problem. It comes on and then gradually improves but it keeps coming back.

Recurrent episodes where the pain comes on quickly but then settles quickly. Pain free between episodes but it keeps coming back.



O Gradual onset. It fluctuates in intensity but never goes away.

Comorbidity:

Red flags:

Spinal malignancy	Age > 50 years
	Previous history of cancer
	Unexplained weight loss
	Constant progressive pain at night
	Pain lasting more than one month
	No improvement after one month of conservative management
	Elderly person with neck pain for first time
	Elderly person with rapidly increasing pain and/or stiffness in the neck
	Dysphagia
	Multiradicular weakness
Spinal infection	Age > 50 years
	Cause for infection - urinary tract, skin or respiratory infection, intravenous drug use, tuberculosis, surgery
	Fever / systemic illness
Fracture	History of violent trauma
	Age > 70 years
	Corticosteroid use
Spinal cord lesion	Bladder or bowel dysfunction
	Widespread progressive motor weakness, disturbed gait, clumsiness, loss of dexterity
	Widespread paraesthesia
	Increased tone / spasticity / hyperreflexes / clonus
	Positive Babinski sign (extensor plantar response)
Inflammatory arthropathy	Gradual onset < 40 years of age
	Marked morning stiffness
	Persisting limitation of movement
	Peripheral joint involvement
	Iritis, skin rashes, colitis, urethral discharge
	Family history
Vascular/ neurological	Extreme dizziness
	Abnormal speech, sight or swallowing
	Blackouts or falls
	Positive cranial nerve signs

N.B. If suspicion of serious spinal pathology is not clear from the history, it should quickly become apparent that loading strategies produce no lasting reduction. Worsening of symptoms in response to all loading strategies is likely (McKenzie & May 2006)

Have you been absent from work due to your current neck or shoulder symptoms?
Yes/no

If yes
For how many days? ……

Did you already return to work?
Yes/no

If not when are you expecting to return? (number of days or date)

Are your complaints caused by work: yes/no
If yes, in what sense (what type of work/tasks; work stress)

Aggravated by work: yes/no
If yes, what type of work/tasks

Your regular work

	never	sometimes	often	always
1. Can you plan your own work?	O	O	O	O
2. Do you have enough variation in your work?	O	O	O	O
3. Do you have too much to do?	O	O	O	O
4. Do you work against the clock?	O	O	O	O

5. How do you rate your relationship with your colleagues? Please circle the relevant number

0	1	2	3	4	5	6	7	8	9	10

Very bad Very good

6. How do you rate your relationship with your supervisor? Please circle the relevant number

0	1	2	3	4	5	6	7	8	9	10

Very bad Very good

7. Have there been any changes in personal situations leading to higher demands in personal life over the last 12 months?

8. Any life events in past 12 months
   - O Birth
   - O Death
   - O Divorce
   - O Miscarriage
   - O Lost job
   - O Major accident / Ill health
   - O

9. Would you like to mention any other things I didn't ask about?

Physical examination

POSTURE

Sitting: *Good / Fair / Poor* Standing: *Good / Fair / Poor* Protruded Head: *Yes / No* Wry Neck: *Right / Left / Nil*

Correction of Posture: *Better / Worse / No effect* _____ Relevant: *Yes / No*

Other Observations: _____

NEUROLOGICAL

Motor Deficit _____ Reflexes _____

Sensory Deficit _____ Dural Signs _____

MOVEMENT LOSS	Maj	Mod	Min	Nil	Pain
Protrusion					
Flexion					
Retraction					
Extension					

	Maj	Mod	Min	Nil	Pain
Lateral flexion R					
Lateral flexion L					
Rotation R					
Rotation L					

TEST MOVEMENTS **Describe effect on present pain – During:** produces, abolishes, increases, decreases, no effect, centralising, peripheralising, **After:** better, worse, no better, no worse, no effect, centralised, peripheralised.

	Symptoms During Testing	Symptoms After Testing	Mechanical Response ↑Rom	↓Rom	No Effect
Pretest symptoms sitting:					
PRO					
Rep PRO					
RET					
Rep RET					
RET EXT					
Rep RET EXT					
Pretest symptoms lying:					
RET					
Rep RET					
RET EXT					
Rep RET EXT					
If required pretest pain sitting:					
LF - R					
Rep LF - R					
LF - L					
Rep LF - L					
ROT - R					
Rep ROT - R					
ROT - L					
Rep ROT - L					
FLEX					
Rep FLEX					

STATIC TESTS

Protrusion _____ Flexion _____

Retraction _____ Extension: *sitting / prone / supine* _____

OTHER TESTS _____

Area

o Spine			
o Cervical	o Thoracic	o Lumbar	o Other (please specify)
o Peripheral			
Symptom response classification			
o Centralization	o Non-centralization	O Other	

4.2 History

The history taking usually starts with the patient on the treatment table or a backless chair so that the true relaxed sitting posture is revealed. Please make sure that the patient is relaxed and try to avoid using any medical jargon that may be unfamiliar. Start with open ended questions first, rather than using leading questions. More specific questions can then be used if particular aspects need clarification.

The aims of history – taking are to (1):

- obtain an overall impression of the patient's presentation and response to his/her problem
- determine the functional limitations on the patient's life
- determine the painful sites i.e. neck, arm, symmetrical or not
- determine the stage of the disorder (e.g. acute, subacute, chronic) and the status of symptoms (improving, worsening, same)
- identify any red flags or contraindications
- identify movements or positions that improve or worsen the patient's symptoms
- determine how severe is the problem (clues as to the vigour of the physical examination)

Current date

Please insert date of history taking using a dd/mm/yy format and your (research clinician) initials in brackets.

ID number/ DOB

This is the ID number and date of birth of the patient being assessed. This information should be already available. Please make sure you have the correct information by asking the patient his date of birth.

Body diagram

Used to record 'all symptoms this episode' i.e. all the symptoms the patient is complaining of, not signs. Please note that all symptoms may still not be present at the time of the assessment. The aim is to determine the most distal extent of any pain, which to some extent provides an indication of the severity of the problem (2).

Examples of questions used (1):

- Where have you had symptoms this episode?
- Where have you had pain or aching?
- Have you had any pins and needles, tingling or numbness?
- Have you had any weakness in your arm?
- Where are you still having symptoms?

Present symptoms
Record the location/ type of symptoms that are still concerning the patient at the time of the assessment. This may differ from the body chart as not all symptoms may still be present. Central or bilateral symptoms usually need sagittal plane forces (e.g. flexion or extension) whereas unilateral symptoms may require movement testing in the lateral plane e.g. lateral flexion (1).

Present since
This is usually given in weeks or days. Alternatively you can write a specific date if known. If a patient has had recurrent problems, please record only the date of the present episode. The present episode, for example, may be indicated by an acute exacerbation of a problem that may have caused the patient to seek assistance in chronic cases (1).

Improving/ unchanging/ worsening
Circle as appropriate, and ask patient how, or in what way if they say they are improving or worsening (i.e. time: constant or intermittent? frequency and intensity: increased or decreased? referral of pain: towards the spine/ midline or moving to the extremities? movement and activities: increased, decreased, the same?) (1).

Commenced as a result of
This can not always be determined, and careful questioning is usually required to determine the true relationship between the event and the onset of symptoms (1). If appropriate describe the mechanism of injury e.g. lifting or twisting or circle no apparent reason.

Symptoms at onset
This question aims to determine whether the location of symptoms has changed since onset (1). Please circle area of symptoms and give a timeframe of onset of distal pain e.g. circle neck two days later shoulder.

Constant/ Intermittent symptoms
Circle as appropriate. For intermittent symptoms, clarify timeframe or frequency outside the circle e.g. 2 hours a day. Some patients can interpret troubling symptoms as constant even though they are not; if so, please clarify with questions such as: ‘is there any time day or night when you have no pain or discomfort?’ (1)

Aggravating/ Easing factors (Worse/ Better)
What makes your symptoms worse? What makes your symptoms better? These questions determine the movements, positions or activities that influence symptoms. It is important to record these, as well as what type of loading strategies (movement or postures?) mostly affects the symptoms(1). Make sure you clarify if activities consistently influence symptoms in the same way and what happens when the activity stops.

If there are two unrelated areas of pain, you may need to indicate if dealing with neck or arm pain in each activity

- Circle for always
- Line under – sometimes
- Oblique line through – no effect
- Put a ? above activity if patient still unsure even after further questions, rather than leaving blank

Other: It may also be useful to find out if there are any other movements or activities the patient finds painful and/or is avoiding. This is helpful to gain information about fear avoidance behaviour.

24 picture (am/ as the day progresses/ pm)
Are you better or worse on waking in the morning? Are you better or worse as the day progresses? Please indicate when symptoms are worse/ better during the day by putting a circle in the appropriate answer. The diurnal pattern may provide some information as to the effect of different activities and the effect of general activity compared to rest (1)

Disturbed sleep
Does the pain wake you at night? If always, circle yes. Sometimes, underline yes. Not affected, circle No. If sleep was previously disturbed, please circle 'Yes' but write previously. It is also useful to find out if disturbed sleep is caused because of pain, usual sleeping pattern or other reasons.

Sleeping postures
Circle usual, indicate if unable to use this because of current pain and indicate current position – best and worse

Sleeping surface
Please circle as appropriate

Pillows
Describe number and type of pillows e.g. thin

Previous episodes
Have you had neck pain before? If so, approximately how many episodes? Circle most representative answer (0, between 1-5, 6-10, 11+) and indicate year of first episode. This information should be used in conjunction with 'Course of symptoms until now'.

Previous history
This section should be used in conjunction with 'Course of symptoms until now'. Write if episodic, which areas affected before and what was it like between episodes e.g. 100% between episodes. Document previous treatments and diagnostic investigations for current and previous episodes, if any. What treatment have you found particularly helpful? Indicate what has helped and in what way if appropriate.
Course of symptoms
Ask patient which is the most representative course of his symptoms until now

Comorbidity
Indicate whether there is any upper extremity or musculoskeletal pain including previous traumatic injury to the affected neck/ limb/ or shoulder resulting in a related current or prolonged disability. Also ask for other medical conditions including any severe psychiatric or personality disorders diagnosed by a psychiatrist (exclusion criterion). Indicate whether the employee is expected to receive a major medical or surgical treatment within the next 3 -4 months.

Red flags
Serious pathology causing spinal symptoms is usually rare (3) however, if the patient reports any red flags, they need to be explored in relation to their neck symptoms and findings from the physical examination.

Questions in relation to red flags (1):
- Are you working normally? Is there any weakness or clumsiness in your arms and/or legs?
- Are there any pins and needles or altered sensations in your arms and/or legs?
- Is your bladder and bowel function normal as always?
- Is there any history of serious illness?
- Do you have to leave your bed at night because of pain?
- Has there been any unexplained weight loss recently?
- Is there any systemic ill health or malaise?
- Has there been any major surgery?
- Have there been any major recent accidents?

Work absence (have you been absent from work due to your current neck or shoulder symptoms?)
Circle as appropriate. If Yes, document the number of days and if person has returned to work or not. Expected return should be in number of days or date

Work as a cause (are your complaints caused by work?)
Circle as appropriate. If yes, what type of work/ tasks/ stress start the symptoms

Work aggravation (Aggravated by work)
Circle as appropriate. If yes, indicate the tasks that worsen symptoms

Your regular work
It is important to know the patient's occupation and what type of activities it involves. You should aim at the predominant activities of the patient's working hours so that factors potentially influencing symptoms are determined. A change in activities may also be useful information.

Describe type of work and job activities/ day schedule and indicate frequency of activity e.g. 50% standing, 50% sitting

Questions 1-4, 8
Tick or cross statement that best answers each question

Questions 5, 6
Mark the scale with a vertical hyphen

Questions 7, 9
Give a brief description and provide explanation if appropriate

4.3 Physical examination
The physical examination involves various observations and movements based on which judgements about symptom response are made. The aims of the physical examination are to expose (1):

- the patient's usual posture
- the symptomatic response to posture or other static loading strategies
- any obvious deformities or asymmetries related to this episode
- neurological examination findings, if appropriate
- baseline measures of mechanical presentation
- symptomatic and mechanical response to movement and other testing

Please note that it is not essential to perform all the components of the physical examination with every patient. If a particular aspect of testing is not performed, please draw an oblique line.

Posture
During the history, the patient is sitting unsupported so you should be able to observe his/ her true seating posture. Please circle appropriate answer. For the recognition of a protruded head posture, try to imagine dropping a plumb line from the patient's chin (1). If this falls in front of the patient's trunk, the head posture is protruded; if the line falls onto the chest, then the head posture is reasonably upright. You should also note whether there is an exaggerated cervico-thoracic kyphosis or a lateral deviation of the head (if so, clarify whether this is fixed or the patient can correct this)

Correction of posture
Having the patient in the unsupported sitting posture, ask: as you are sitting there now, do you have any of the symptoms that you have told me about? It is important to determine the location and intensity of the most distal symptoms, and whether they have worsened or provoked whilst in sitting.

Procedure for posture correction (1), (please see Appendix):

- The lumbar lordosis is restored and then the head is retracted to a neutral position.
- Once in this position for a minute or two, ask: In that position, do symptoms feel better, worse or the same?
- Circle response and indicate which pain changes if appropriate

Posture in standing could also be examined at this point if needed.
N.B: In patients who display a lateral deviation that is too painful to correct, the normal examination procedures are usually abandoned. Examination should continue in an unloaded position e.g. supine with a combination of appropriate positioning and time until movement begins to return (1).

Other observations

Record any significant musculoskeletal differences e.g. wasting, leg difference etc

Neurological testing
The necessity of the neurological examination is based on your clinical judgement. From the history, possible indicative clues may be if the patient reports referred pain (i.e. arm or forearm symptoms), sensory changes (e.g. paresthesia), muscle weakness or gait disturbance (1).

The neurological examination may involve four components (1), p.176-177:
- sensation
- muscle power
- reflexes
- nerve tension tests

Qualify which deficit in each section, recorded if abnormal e.g. decreased S1 reflex. Can add Babinski / Clonus to reflexes if required. Record as NAD if testing is normal. Oblique line through if not applicable (not performed)

Range of movement and Movement loss
Movement testing begins from a standardised seated position with back support. This type of testing will help determine the presence of any movement loss and initial symptom response to treatment. For further tips, please refer to CROM measurement protocol. Please note that you can not use the CROM if a patient has a pacemaker.

The following movements are examined (1):
- Protrusion
- Flexion
- Retraction
- Extension
- Rotation
- Lateral flexion

Movement loss
The boxes Maj/Mod/Min/Nil can be used as a line i.e. more as a continuum. Please compare with prior to current episode and also tick the “pain” box, if the patient is reporting pain (also indicate location of the pain).

Test movements
This section is used to determine the effect that different movements and positions have on symptoms. Start with the sagittal plane (except in the case of a lateral deviation) and then proceed to other planes e.g. frontal if appropriate. Frontal plane movements should be tested if sagittal plane movements do not improve the symptomatic or mechanical presentation (1)

Repeated movements (1) (please see Appendix):
- Protrusion (sitting)
- Retraction (sitting)
- Retraction and extension (sitting)
- Retraction and extension (lying)

- Lateral flexion (sitting)
- Rotation (sitting)
- Flexion (sitting)

Please indicate the order performed by numbering *if order is different to standard as shown in form*. Please also record the number of repetitions performed to gain the response. Baseline: 10 repetitions for each movement

Symptomatic response - Use standard terms only (found at the top of the table). The symptomatic response is recorded 3 times during the assessment: before, during and after testing. Monitor and describe effect on most distal symptoms predominantly e.g. 'Sitting there now, are you feeling any of the symptoms you have mentioned?' Where is your pain now? If you have pain in your arm, how far down the arm does it extend?
(1). Avoid leading questions e.g. is the pain increasing?

Mechanical response – Tick appropriate box. Can indicate which movement has been affected by the change if it is different to the one being tested.

On completion of the repeated tested movement and return to the neutral position, ask the patient about their symptoms (e.g. type, **location and intensity**) and observe whether movement has increased or decreased.

Record as 'better', 'worse', 'no worse', or 'no effect'

If no change in the symptomatic or mechanical response occurs, the process may have to be repeated more vigorously (1). It is also possible that pain may be caused by other areas e.g. shoulder.

N.B: Once a favourable response is elicited, further testing is redundant and unnecessary (1). Some patients with acute or severe presentations may not tolerate testing while sitting and may need to be examined from alternative unloaded positions e.g. lying down. In patients who display a lateral deviation that is too painful to correct, the normal examination procedures are usually abandoned. Examination should continue in an unloaded position e.g. supine with a combination of appropriate positioning and time until movement begins to return (1)

Static tests
If the effect of repeated movement testing on symptomatic or mechanical presentation is not significant, it may be necessary to perform static tests or sustained postures (1) . Standard static evaluation can be conducted in the following postures:
- sitting slouched, head protruded
- sitting upright, head retracted
- retraction and extension in supine lying
- extension in prone lying

Record with standard "After" words (found at the top of the 'test movements' table).

Other tests
State which and the response achieved

Area
Tick the area that is likely to be the cause of the patient's symptoms

Symptom response classification
Please tick as appropriate.

References

1. McKenzie RA, May S. The cervical and thoracic spine: mechanical diagnosis and therapy. Anonymous, editor. Waikanae, New Zealand: Spinal publications; 2006.
2. Spitzer WO, LeBlanc FE, Dupuis M, Abenhaim L, Bloch R, Bombardier C, et al. Scientific approach to the activity assessment and management of activity related spinal disorders. Spine. 1987;12:S1-S55.
3. Clinical Standards Advisory Group C. Back Pain. Report. London: HMSO; 1994 Contract No.: Document Number|.

Research clinicians training - Questions Part 1

1. Centralization was originally described as:

☐ The change in neurological status of patients with spinal pain

☐ The movement of symptoms originating from the spine from a distal to a more proximal location in response to spinal testing

☐ The movement of symptoms originating from the spine from a proximal to a more distal location in response to spinal testing

Comments:

2. **Centralization/ peripheralization of symptoms can be present only in the:**

☐ Postural syndrome

☐ Dysfunction syndrome

☐ Derangement syndrome

Comments:

3. **When recording findings in the assessment form, the line under means:**

☐ Always

☐ Sometimes (it may be useful to document frequency or circumstances) e.g. after 5min

☐ Never

Comments:

4. **Once the whole patient assessment is finished:**

☐ The patient assessment form is placed into an opaque envelope and relevant details are completed at the front side of envelope and the assessment log

☐ The research clinician keeps the assessment form for his own use and completes the assessment log

Comments:

5. **Please give an example of a serious adverse event:**
6. **Please give an example where you would be cautious when undertaking repeated movement testing:**
7. **Please give an example of a contraindication of movement testing:**

8. If a complication or adverse event is discovered:

☐ This should be communicated back to the Chief Investigator and the event notification form must be completed

☐ Participant should be referred back to GP or OH

☐ None of the above

Comments:

9. Testing for range of motion and movement loss takes place:

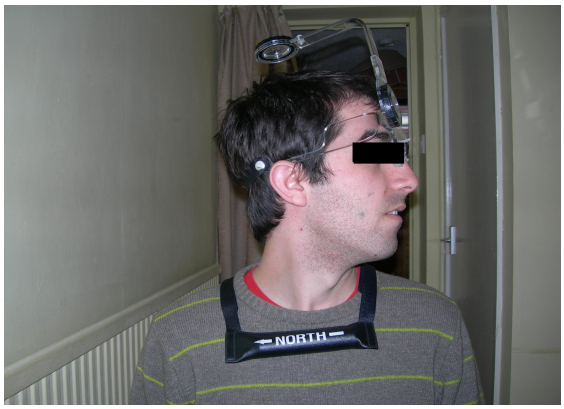
☐ With participant sitting in a chair without back support

☐ With participant sitting in a chair with back support

☐ With participant standing

Please describe body posture of participant:

10. **The following picture illustrates an example of a range of motion measurement (rotation to the left) using the CROM device.**



Which possible errors can you identify in this picture? Please describe

11. Movement loss is established through:

☐ The indication of the CROM

☐ The perception of patient before the current episode

☐ Both of the above

☐ None of the above

Comments:

12. Which would be the expected effect of repeated movement testing on the following:

A. Postural syndrome

☐ Decrease end-range pain - Better

☐ Produce end-range pain - Worse

☐ No effect

B. Dysfunction syndrome

☐ Decrease end-range pain - No better

☐ Produce end-range pain - No worse

☐ Centralise end-range pain - Better

C. Derangement syndrome

☐ Produce pain during or end-range pain – Decrease ROM – Worse

☐ Centralise pain – Increase ROM – Better

☐ No effect

Comments:

13. **If the ROM increases, but symptoms remain unchanged you should:**

☐ Stop the procedure that induced this response immediately

☐ Continue with the same procedure for a few more sets

☐ Apply more vigorous testing immediately

Comments:

14. **If repeated retraction produces symptoms at end range during repeated movements but afterwards the patient's symptoms are no worse:**

☐ Fewer repetitions of retraction are required

☐ The application of more force is most likely necessary

Comments:

15. **In the previous case, if with more force pain is still produced at end-range but is no worse after testing and the range of motion remains unchanged, this is likely a:**

☐ Derangement syndrome

☐ Dysfunction syndrome

☐ None of the above

Comments:

16. **Which is the progression of force for retraction in sitting?**

☐ Retraction in sitting with patient overpressure

☐ Retraction extension in sitting

☐ Retraction extension in lying

☐ Lateral flexion in sitting

Comments:

17. **If following a test movement, symptoms felt in the lower part of the arm are now felt in the upper part of the arm, recording should be as follows:**

☐ Improved

☐ Centralised

☐ Peripheralised

☐ None of the above. It should be clarified whether this is the most distal symptom.

Comments:

18. **During movement testing, the participant's pain is produced on the first movement, but decreases on repetition and by the end of the testing procedure is minimal or absent. Response should be considered and recorded as:**

☐ Favourable – Better

☐ Unclear – No worse

☐ Undesirable – Worse

Comments:

19. When you would use the following testing? What might have been done before?

A. Retraction extension (lying supine)

B. Lying on the side with pillow support on the neck

20. **Participant has neck pain 4/10 and pain in the arm 2/10 at baseline. After repeated retraction, the pain in the arm is abolished but neck pain is 5/10:**

☐ Patient status is worse. The procedure should be stopped

☐ Patient status is unclear. Further testing is required

☐ Patient status has improved.

Comments:

21. **Frontal plane testing should be undertaken if:**

☐ A lateral deviation is present. Sagittal plane testing should not be tested

☐ The symptomatic or mechanical presentation has improved

☐ The symptomatic or mechanical presentation has not improved or worsened

Comments:

22. **Participant complains of pain 'that never goes away even for a few minutes' in their neck. Physical examination reveals symptom fluctuations that after testing never go away. Possible causes of the participant's symptoms may be:**

☐ A non mechanical cause e.g. inflammation

☐ A mechanical cause: irreducible derangement

☐ Both of the above

☐ None of the above

Comments:

23. **Patient reports pain in the arm and complains that he/she can not lift arm as much as the other side. Is this or not an indication for neurological examination?**

☐ Yes

☐ No

☐ Other

Comments:

Research clinicians training - Questions Part 2

24. **When a patient has pain only in the neck, centralisation is defined as:**

☐ The decrease in intensity of the patient's neck symptoms in response to spinal testing

☐ The abolition of the patient's neck symptoms in response to spinal testing

☐ The movement of neck symptoms from a proximal to a more distal location in response to spinal testing

Comments:

25. The following statements represent possible reasons for failure to achieve centralisation. Give an example for each case:

- Movements that are not to the patient's end – range:
- Force progressions required:
- Force alternatives required:
- Reduction achieved but not maintained:

26. Are the following statements true or false?

- Sagittal plane movements are tested first except in the case of a wry neck or lateral deviation
- If symptoms in response to one repetition are worse, this direction of movement altogether is abandoned
- If symptoms in response to movement testing are no worse or no better, this direction of movement altogether is abandoned
- The response to unloaded sagittal movements (e.g. when sagittal plane movements in sitting worsen symptoms) should be assessed before abandoning the sagittal plane altogether
- Lateral plane movements are tested if the patient's symptoms worsen or remain unchanged as a result of sagittal plane loading strategies

27. A lateral deviation of the cervical spine may be present when:

Tick all that apply:

☐ Head and upper cervical spine are visibly shifted to one side

☐ Onset of deviation occurred with onset of neck pain

☐ Patient is unable to correct deviation voluntarily and maintain correction

☐ Correction affects intensity or location of symptoms

28. Are the following statements true or false?

In the presence of a lateral deviation:

- The patient's symptoms may be on one side with the head shifted to the opposite side (contralateral deviation)
- The patient's symptoms may be on one side with the head shifted to the same side (ipsilateral deviation)

Flexion forces:

- Are usually required in the management of headaches and some types of dysfunctions
- May sometimes be combined with lateral procedures e.g. lateral flexion or rotation, when movement testing in one plane has failed to elicit a favourable response

29. **In determining which lateral force to use, the clinician should take account of the patient's:**

☐ Reported aggravating factors

☐ Movement loss (especially the most affected movement)

☐ Response to repeated movement testing

☐ All of the above

☐ None of the above

30. Kyphotic deformity: in the presence of central or symmetrical pain, extension is obstructed and head is fixed in protrusion and flexion. Any attempt to correct this worsens symptoms and patients may avoid movements by maintaining a flexed posture. What would you do in such a case?

☐ Carry on with the usual procedures in sitting until a favourable response is elicited

☐ It is usually impossible to carry out a normal physical examination. Assessment is preferably undertaken in an unloaded position. Head should rest on pillows and/or raised treatment table so that flexion deformity is accommodated. Retraction in

supine is attempted. Depending on symptom response, pillows/ treatment table are gradually lowered letting the head move towards a more neutral → extension direction

31. When distinguishing between a derangement and a dysfunction which of the following factors should be taken into account?

☐ Consistency of aggravating factor
☐ Response of symptoms once the aggravating position is released
☐ Both of the above
☐ None of the above

32. Case study 1. Have a look at the following case study:

Patient 31yrs, Sitting 75%, Moving 25% most days:

History

Present Symptoms: Neck, upper trapezius, headaches

Present since: Unchanging last 2 months

Commenced as a result of: Road accident / Or no apparent reason

Symptoms at onset: Neck, arm, forearm, headache

Constant symptoms: *neck / arm / forearm / headache*

Intermittent symptoms: *neck , L arm*

Worse: *bending* *sitting* *turning L & R* *lying / rising* *when still or moving* *am / as the day progresses / pm* *better after shower* *fast*

other When looking up

Better: *bending* *sitting* *turning when still / on the move* *lying* *am / as the day progresses / pm*

other **When moving slowly or carefully**

Disturbed Sleep: No

Pillows: _____

Sleeping postures: *Prone / sup / side R / L*

Surface: *Firm / soft / sag*

Previous Episodes: 0

Year of first episode: _____

Previous History: _____

Physical examination

POSTURE

Sitting: *Poor* Standing: *Good / Fair / Poor* Protruded Head: *Yes* Wry Neck: *Nil*

Correction of Posture: *Better / Worse / No effect* _____ Relevant: *Yes / No*

Other Observations: _____

NEUROLOGICAL

Motor Deficit ***Not tested*** Reflexes ***Not tested***

Sensory Deficit ***Not tested*** Dural Signs ***Not tested***

MOVEMENT LOSS	Maj	Mod	Min	Nil	Pain
Protrusion		√			
Flexion	√				
Retraction	√				
Extension	√				

	Maj	Mod	Min	Nil	Pain
Lateral flexion R		√			
Lateral flexion L		√			
Rotation R		√			
Rotation L		√			

TEST MOVEMENTS **Describe effect on present pain – During:** produces, abolishes, increases, decreases, no effect, centralizing ,peripheralising. **After:** better, worse, no better, no worse, no effect, centralised, peripheralised.

	Symptoms During Testing	Symptoms After Testing	Mechanical Response		
			↑Rom	↓Rom	No Effect
Pretest symptoms sitting:	**None**				
PRO					
Rep PRO					
RET	Pro ERP neck	NW			
Rep RET	Pro ERP neck	NW			
RET EXT	Pro ERP neck & shoulders	NW			
Rep RET EXT	Pro ERP neck & shoulders	NW			
Pretest symptoms lying:					
RET					
Rep RET					
RET EXT					
Rep RET EXT					
If required pretest pain sitting:					
LF - R					
Rep LF - R					
LF - L					
Rep LF - L					
ROT - R	Pro ERP neck & headache	NW			
Rep ROT - R	Pro ERP neck & headache	NW			
ROT - L	Pro ERP neck & headache	NW			
Rep ROT - L	Pro ERP neck & headache	NW			
FLEX	Pro ERP neck & headache	NW			
Rep FLEX	Pro ERP neck & headache	NW			

STATIC TESTS

Protrusion _____ Flexion _____

Retraction _____ Extension: *sitting / prone / supine* _____

OTHER TESTS _____

Looking at the history only,

- which types/directions of forces appear to make the patient's symptoms **worse:**

- which types/directions of forces appear to make the patient's symptoms **better:**

Looking also at the physical examination:

- Is it the neck or not? ☐ Yes ☐ No
- Is this possibly a centraliser or not? ☐ Yes ☐ No
- Is this possibly a dysfunction or derangement?

33. Case study 2. Have a look at the following case study:

Student 25 years old:

History

Present Symptoms ***neck***

Present since **3 weeks** *Unchanging*

Commenced as a result of *Studying for exams* *Or no apparent reason*

Symptoms at onset: *Neck*

Constant symptoms: *neck / arm / forearm / headache* Intermittent symptoms: *neck*

Worse *Bending 1h* *Sitting after 30'* *turning* *lying / rising*

am / as the day progresses / pm *when still*

other *When studying for long time*

Better *bending* *sitting* *turning* *lying*

am / as the day progresses / pm *on the move*

other ***Rising for chair***

Disturbed Sleep *No* Pillows

Sleeping postures *Prone / sup / side R / L* Surface *Firm / soft / sag*

Previous Episodes 0 Year of first episode

Previous History

Physical examination

POSTURE

Sitting: *Poor* Standing: *Good / Fair / Poor* Protruded Head: *Yes* Wry Neck: *Nil*

Correction of Posture: *Better* _____ Relevant: *Yes / No*

Other Observations: _____

NEUROLOGICAL

Motor Deficit ***Not tested*** Reflexes ***Not tested***

Sensory Deficit ***Not tested*** Dural Signs ***Not tested***

MOVEMENT LOSS	Maj	Mod	Min	Nil	Pain
Protrusion				√	
Flexion				√	
Retraction				√	
Extension				√	

	Maj	Mod	Min	Nil	Pain
Lateral flexion R				√	
Lateral flexion L				√	
Rotation R				√	
Rotation L				√	

TEST MOVEMENTS **Describe effect on present pain – During:** produces, abolishes, increases, decreases, no effect, centralizing ,peripheralising. **After:** better, worse, no better, no worse, no effect, centralised, peripheralised.

	Symptoms During Testing	Symptoms After Testing	Mechanical Response		
			↑Rom	↓Rom	No Effect
Pretest symptoms sitting:	**None**				
PRO	**N.E**	**N.E**			
Rep PRO	**N.E**	**N.E**			
RET	**N.E**	**N.E**			
Rep RET	**N.E**	**N.E**			
RET EXT	**N.E**	**N.E**			
Rep RET EXT	**N.E**	**N.E**			
Pretest symptoms lying:					
RET					
Rep RET					
RET EXT					
Rep RET EXT					
If required pretest pain sitting:	**None**				
LF - R	**N.E**	**N.E**			
Rep LF - R	**N.E**	**N.E**			
LF - L	**N.E**	**N.E**			
Rep LF - L	**N.E**	**N.E**			
ROT - R	**N.E**	**N.E**			
Rep ROT - R	**N.E**	**N.E**			
ROT - L	**N.E**	**N.E**			
Rep ROT - L	**N.E**	**N.E**			
FLEX	**N.E**	**N.E**			
Rep FLEX	**N.E**	**N.E**			

STATIC TESTS

Protrusion _____ Flexion _____

Retraction _____ Extension: *sitting / prone / supine* _____

OTHER TESTS _____

Looking at the history only,

- which types/directions of forces appear to make the patient's symptoms **worse:**

- which types/directions of forces appear to make the patient's symptoms **better:**

Looking also at the physical examination:

- Is it the neck or not? ☐ Yes ☐ No
- Is this possibly a centraliser or not? ☐ Yes ☐ No
- What testing you would do after repeated movement testing?

Appendix 5.6 Critical appraisal of guidelines[1]. APTA, American Physical Therapy Association.

ITEM	TASK FORCE ON NECK PAIN	APTA
I. Scope and purpose		
1. The overall objective(s) of the guideline are specifically described	4	4
2. The clinical question(s) covered by the guideline is (are) specifically described	4	1
3. The patients to whom the guideline is meant to apply are specifically described	4	4
II. Stakeholder involvement		
4. The guideline development group includes individuals from all the relevant professional groups	4	4
5. The patient's view and preferences have been sought	4	1
6. The target users of the guideline are clearly defined	1	2
7. The guideline has been piloted among target users	1	2
III. Rigor of development		
8. Systematic methods were used to search for evidence	4	3
9. The criteria for selecting the evidence are clearly described	4	2
10. The methods used for formulating the recommendations are clearly described	3	3
11. The health benefits, side effects, and risks have been considered in formulating the recommendations	4	3
12. There is an explicit link between the recommendations and the supporting evidence	2	2
13. The guideline has been externally reviewed by experts prior to its publication	2	2
14. A procedure for updating the guideline is provided	1	4
IV. Clarity and Presentation		
15. The recommendations are specific and unambiguous	2	4
16. The different options for management of the condition are clearly presented	4	3
17. Key recommendations are easily identifiable	4	4
18. The guideline is supported with tools for application	3	3
V. Applicability		
19. The potential organizational barriers in applying the recommendations have been discussed	3	1
20. The potential cost implications of applying the recommendations have been considered	1	2
21. The guideline presents key review criteria for monitoring and/or purposes	1	1
22. The guideline is editorially independent from the funding body	4	1
VI. Editorial independence		
23. Conflicts of interest of guideline development members have been reported	4	1
Total	**68 / 92**	**57 / 92**

[1] The Appraisal of Guidelines for Research and Evaluation (AGREE) was used for this purpose. Items have been rated on a 4-point Likert scale (4 = strongly agree to 1= strongly disagree) (for further information: AGREE Collaboration, 2003).

Appendix 5.7 Ethics approval letter for the reliability study.

Warwick
Medical School

19 October 2007

Ms Angeliki Chorti
Warwick Medical School
University of Warwick
Coventry
CV4 7AL

Dear Ms Chorti

Ref: Symptom response pilot study: reliability and role in predicting outcome in neck pain.

Thank you for submitting your revisions for the the above-named project to the University of Warwick Biomedical Research Ethics Sub-Committee for Chair's Approval.

The Chair is pleased to confirm that full approval has been granted and your study may commence. May I remind you that the Committee would like to receive an End of Project Report once your study is completed.

Yours sincerely,

Jane Barlow
Chair
Biomedical Research
Ethics Sub-Committee

Copy:
Lynn Green, Research Governance Facilitator

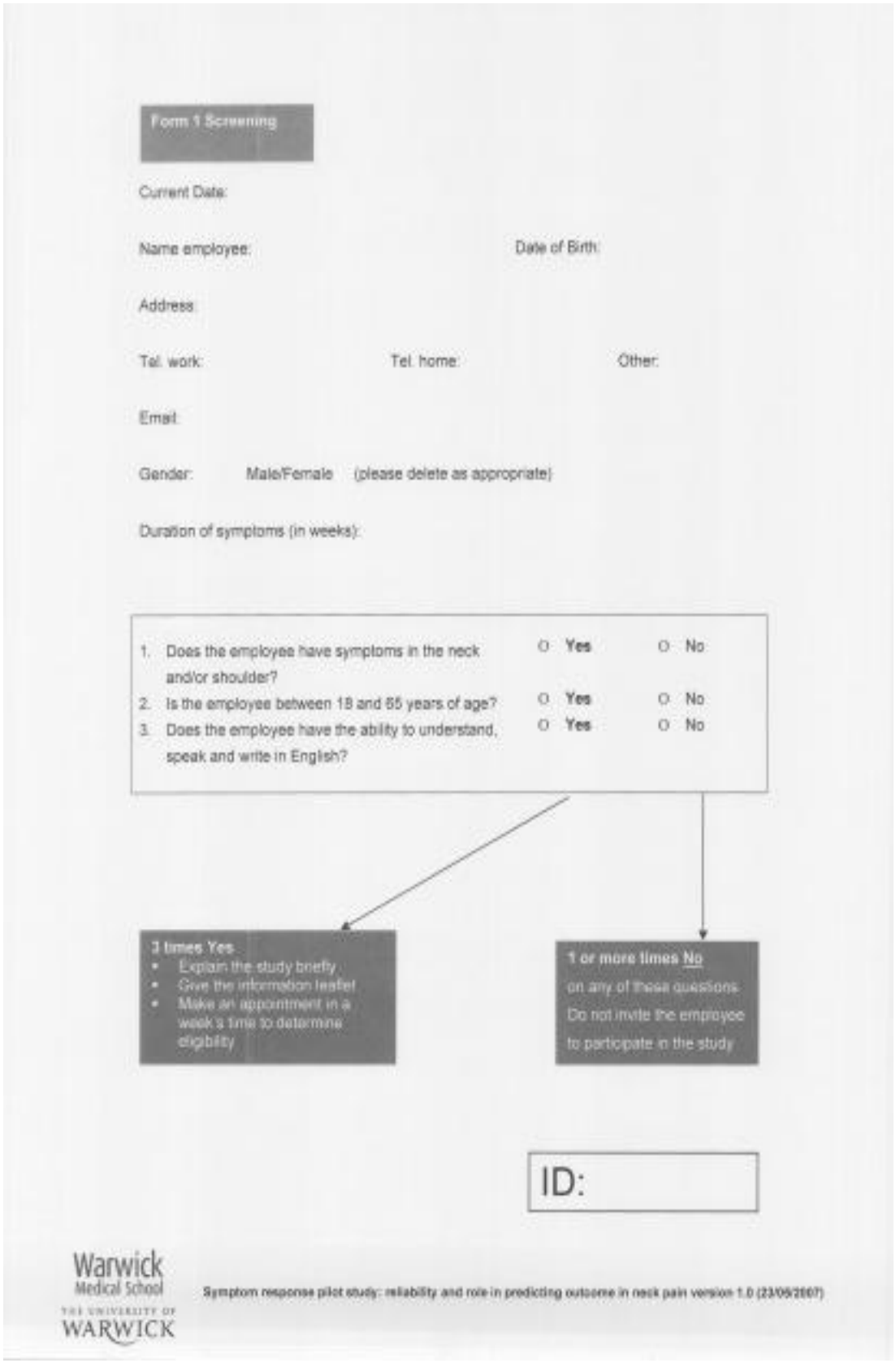
Biomedical Research Ethics Subcommittee
Enquiries: Krysia Saul
Tel: 02476-573163
Email: krysia.saul@warwick.ac.uk

Health Sciences Research Institute
Warwick Medical School

Medical School Building
The University of Warwick
Coventry CV4 7AL United Kingdom
Tel: +44 (0)24 7652 8204
Fax: +44 (0)24 7652 8575

THE UNIVERSITY OF
WARWICK

Appendix 5.8 Screening form



Form 1 Screening

Current Date:

Name employee: Date of Birth:

Address:

Tel. work: Tel. home: Other:

Email:

Gender: Male/Female (please delete as appropriate)

Duration of symptoms (in weeks):

1. Does the employee have symptoms in the neck and/or shoulder?	O Yes	O No
2. Is the employee between 18 and 65 years of age?	O Yes	O No
3. Does the employee have the ability to understand, speak and write in English?	O Yes	O No

3 times Yes
- Explain the study briefly
- Give the information leaflet
- Make an appointment in a week's time to determine eligibility

1 or more times No
on any of these questions
Do not invite the employee to participate in the study

ID:

Warwick Medical School
THE UNIVERSITY OF WARWICK

Symptom response pilot study: reliability and role in predicting outcome in neck pain version 1.0 (23/05/2007)

Centre number:
Study Number:
Participant Identification Number for this study:

Appendix 5.9 Consent forms.

CONSENT FORM

Symptom response pilot study: reliability and role in predicting outcome in neck pain.

Please initial box

1. I confirm that I have read and understand the information leaflet dated …… version …… for the above study. I have had the opportunity to consider the information, ask questions and have had these answered satisfactorily. ☐

2. I understand that my participation is voluntary and that I am free to withdraw at any time, without giving any reason, without my medical care or legal rights being affected. ☐

3. I understand that relevant sections of the data collected during the study may be looked at by individuals involved in the study, or from regulatory authorities, where it is relevant to my taking part in this research. I give permission for these individuals where it is relevant to have access to my records. ☐

4. I agree to my GP being informed of my participation in the study. ☐

5. I am aware that the results of the study may be presented in research reports, scientific conferences and/or journals. However, the information I provide for the study will remain confidential. ☐

6. I am aware that the research team will contact me again at 3 months. ☐

7. I agree to participate in the above study ☐

_____ _____ _____

Name of Participant (BLOCK CAPITALS) Date Signature

_____ _____ _____

Name of person taking consent Date Signature

Warwick Medical School

THE UNIVERSITY OF WARWICK

CONSENT FORM

Symptom response identification in neck pain: a reliability study using videotaped assessments across multiple examiners (VideoNeck study)

Please initial box

8.	I confirm that I have read and understand the Patient Information sheet dated …….. version … for the above study. I have had the opportunity to consider the information, ask questions and have had these answered satisfactorily.	☐
9.	I understand that my participation is voluntary and that I am free to withdraw at any time, without giving any reason and without any of my rights being affected.	☐
10.	I understand that the research team and/or other authorised individuals at the Warwick Medical School may have access to my data. I give permission to these individuals where it is relevant to have access to the data I have provided.	☐
11.	I agree to my GP being informed of my participation in the study.	☐
12.	I understand that my examination will be videotaped for research purposes and that other health professionals participating in the study may view these tapes. However, my identity will be concealed and the information I provide will remain confidential. I give my permission to be videotaped and to participating health professionals viewing my assessment where it is relevant.	☐
13.	I understand that the information that I provide for the study will remain confidential and that I will be given anonymity in any publications or reports that arise from this research.	☐
14.	I agree to participate in the above study	☐

_____	_____	_____
Name of Participant (BLOCK CAPITALS)	Date	Signature

_____	_____	_____
Researcher	Date	Signature

N.B. When completed, 1 for participant; 1 for researcher file site

Appendix 5.10 Baseline questionnaire.

Baseline questionnaire

Symptom response pilot study: reliability and role in predicting outcome in neck pain

ID:

Date of Birth (dd/mm/yy):

Current Date (dd/mm/yy):

Contact details:

Angeliki Chorti

University of Warwick

Medical School, Gibbet Hill Campus

Coventry

CV4 7AL

Tel: 02476150405

Email: A.Chorti@warwick.ac.uk

Thank you very much for agreeing to take part in the study and answering this questionnaire. Please try to answer every question and feel free to ask the research therapist for help if you get stuck. Your answers will be kept strictly confidential and your name will not appear anywhere on the questionnaire.

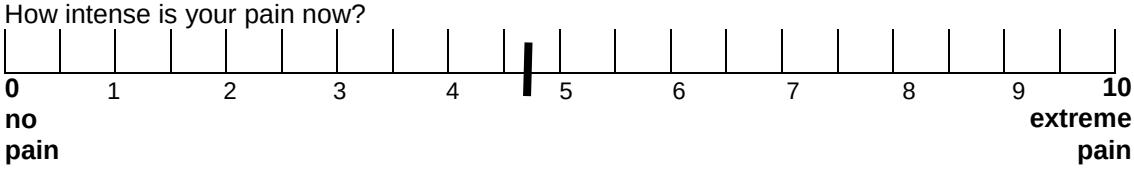
You will receive a similar questionnaire in the post after 3 months for you to complete.

Answering the questions Most of the questions could be answered by circling the most appropriate answer. For example 'How confident are you with ice skating?'

0	**1**	**2**	**3**	**4**	**5**	**6**
Not at all Confident						Completely confident

Two of the questions have **another format**. Mark the scale with a vertical hyphen, like shown below

How intense is your pain now?



Similar questions Two parts of the questionnaire are very similar, the disability arm shoulder hand questionnaire and the neck disability index. This is because we need to know which questions would best measure the effect of your symptoms on your daily life. We kindly ask you to fill in both parts, even if you do have only neck pain or only shoulder pain. This would be of great help to reduce the number of questions in further research.

Disability Arm Shoulder Hand Questionnaire

Instructions

This questionnaire asks about your symptoms as well as your ability to perform certain activities. Please answer every question, based on your condition in the **last week.** If you did not have the opportunity to perform an activity in the past week, please make your best estimate of which response would be the most accurate. It doesn't matter which hand or arm you use to perform the activity; please answer based on your ability regardless of how you perform the task.

	No difficulty	Mild difficulty	Moderate Difficulty	Severe Difficulty	Unable
1. Open a tight or new jar	**1**	**2**	**3**	**4**	**5**
2. Write	**1**	**2**	**3**	**4**	**5**
3. Turn a key	**1**	**2**	**3**	**4**	**5**
4. Prepare a meal	**1**	**2**	**3**	**4**	**5**
5. Push open a heavy door	**1**	**2**	**3**	**4**	**5**
6. Place an object on a shelf above your head	**1**	**2**	**3**	**4**	**5**
7. Do heavy household chores (e.g. wash walls, wash floors)	**1**	**2**	**3**	**4**	**5**
8. Garden or do yard work	**1**	**2**	**3**	**4**	**5**
9. Make a bed	**1**	**2**	**3**	**4**	**5**
10. Carry a shopping bag or briefcase	**1**	**2**	**3**	**4**	**5**
11. Carry a heavy object (over 10 lbs/ 5 kg)	**1**	**2**	**3**	**4**	**5**
12. Change a light bulb overhead	**1**	**2**	**3**	**4**	**5**
13. Wash or blow dry your hair	**1**	**2**	**3**	**4**	**5**
14. Wash your back	**1**	**2**	**3**	**4**	**5**
15. Put on a sweater	**1**	**2**	**3**	**4**	**5**
16. Use a knife to cut food	**1**	**2**	**3**	**4**	**5**
17. Recreational activities which require little effort (e.g. card playing, knitting, etc.)	**1**	**2**	**3**	**4**	**5**
18. Recreational activities in which you take some force or impact through your arm, shoulder and hand. (e.g. golf, hammering, tennis, etc.)	**1**	**2**	**3**	**4**	**5**
19. Recreational activities in which you move your arm freely	**1**	**2**	**3**	**4**	**5**
20. Manage transportation needs (getting from one place to	**1**	**2**	**3**	**4**	**5**

	No difficulty	Mild difficulty	Moderate Difficulty	Severe Difficulty	Unable
another)					
21. Sexual activities	**1**	**2**	**3**	**4**	**5**

	Not at all	Slightly	Moderately	Quite a bit	Extremely
22. During the past week, to what extent has your arm, shoulder or hand problem interfered with your normal social activities with family, friends, neighbours or groups	**1**	**2**	**3**	**4**	**5**

	Not limited at all	Slightly limited	Moderately limited	Very limited	Unable
23. During the past week, were you limited in your work or other regular daily activities as a result of your arm, shoulder or hand problem	**1**	**2**	**3**	**4**	**5**

Please rate the severity of the following symptoms in the last week (circle number)

	None	Mild	Moderate	Severe	Extreme
24. Arm, shoulder or hand pain	**1**	**2**	**3**	**4**	**5**
25. Arm, shoulder or hand pain when you performed any specific activity	**1**	**2**	**3**	**4**	**5**
26. Tingling (pins and needles) in your arm, shoulder or hand.	**1**	**2**	**3**	**4**	**5**
27. Weakness in your arm, shoulder or hand.	**1**	**2**	**3**	**4**	**5**
28. Stiffness in your arm, shoulder or neck	**1**	**2**	**3**	**4**	**5**

	No difficulty	Mild difficulty	Moderate difficulty	Severe difficulty	So much difficulty that I can't sleep
29. During the past week, how much difficulty have you had sleeping because of the pain in your arm, shoulder or neck (circle number)	**1**	**2**	**3**	**4**	**5**

	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
30. I feel less capable, less	**1**	**2**	**3**	**4**	**5**

confident or less useful because of my arm, shoulder or neck problem (circle number)

The following questions ask about the impact of your arm, shoulder or hand problem on your ability to work (including homemaking if that is your main work role)

Please indicate what your job /work is:…….

Please circle the number that best describes your physical ability in the past week. Did you have any difficulty:

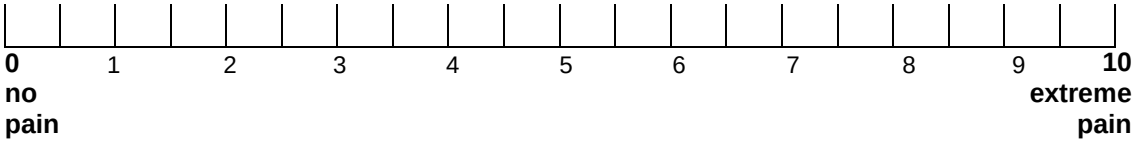
	No difficulty	Mild difficulty	Moderate difficulty	Severe difficulty	Unable
31. Using your usual techniques for your work?	**1**	**2**	**3**	**4**	**5**
32. Doing your usual work because of arm, shoulder or neck pain?	**1**	**2**	**3**	**4**	**5**
33. Doing your work as well as you would like?	**1**	**2**	**3**	**4**	**5**
34. Spending your usual amount of time doing your work?	**1**	**2**	**3**	**4**	**5**

Pain

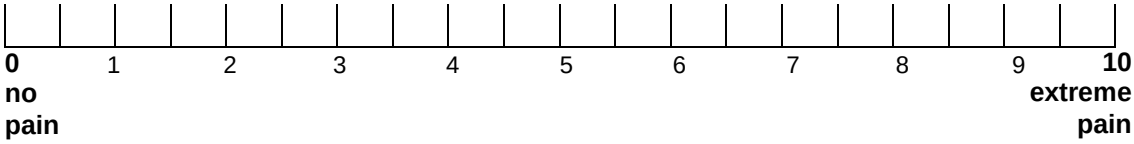
Please mark the scale below to show how intense your pain is (see instructions on page2).

A 0 means no pain and 10 means extreme pain.

35. How intense is your pain now?



36. How intense was your pain on average last week?



37. During the past week, how troublesome has your Neck and/or Shoulder pain been?

1	**2**	**3**	**4**	**5**
Not at all troublesome	Slightly troublesome	Moderately troublesome	Severely troublesome	Very severely troublesome

Neck Disability Index

This questionnaire is designed to enable us to understand how much your neck pain has affected your ability to manage everyday activities. We realise that you may feel that more than one statement may relate to you, but please just circle **the one choice which closely describes** your problem **right now**.

Q38 Pain intensity

1. I have no pain at the moment
2. The pain is very mild at the moment
3. The pain is moderate at the moment
4. The pain is fairly severe at the moment
5. The pain is very severe at the moment
6. The pain is the worst imaginable at the moment

Q39 Personal Care (Washing, Dressing etc.)

1. I can look after myself normally without causing extra pain
2. I can look after myself normally but it causes extra pain
3. It is painful to look after myself and I am slow and careful
4. I need some help, but manage most of my personal care
5. I need help every day in most aspects of self-care
6. I do not get dressed, I wash with difficulty and stay in bed

Q40 Lifting

1. I can lift heavy weights without extra pain
2. I can lift heavy weights, but it gives extra pain
3. Neck pain prevents me from lifting heavy weights off the floor but I can if they are conveniently positioned, for example on the table
4. Neck pain prevents me from lifting heavy weights, but I can manage light to medium weights conveniently positioned
5. I can lift very light weights
6. I cannot lift or carry anything at all

Q41 Reading

1. I can read as much as I want, with no pain in my neck
2. I can read as much as I want, with slight pain in my neck
3. I can read as much as I want, with moderate pain in my neck
4. I cannot read as much as I want, because of moderate pain in my neck
5. I cannot read as much as I want, because of severe pain in my neck
6. I cannot read at all because of the pain in my neck

Q42 Headaches

1. I have no headaches at all
2. I have slight headaches which come infrequently
3. I have moderate headaches which come infrequently
4. I have moderate headaches which come frequently
5. I have severe headaches which come frequently
6. I have headaches almost all of the time

Q43 Concentration

1. I can concentrate fully when I want to with no difficulty
2. I can concentrate fully when I want to with slight difficulty
3. I have a fair degree of difficulty in concentrating when I want to
4. I have a lot of difficulty in concentrating when I want to
5. I have a great, great deal of difficulty in concentrating when I want to
6. I cannot concentrate at all

Q44 Work

1. I can do as much work as I want to
2. I can only do my usual work, but no more
3. I can do most of my usual work, but no more
4. I cannot do my usual work
5. I can hardly do any work
6. I cannot do any work at all

Q45 Driving (please leave blank if you don't have a driver licence or don't drive)

1. I can drive my car without any neck pain at all
2. I can drive my car as long as I want, with slight pain in my neck
3. I can drive my car as long as I want, with moderate pain in my neck
4. I cannot drive my car as long as I want, because of moderate pain in my neck
5. I can hardly drive my car at all because of severe pain in my neck
6. I cannot drive my car at all

Q46 Sleeping

1. I have no trouble sleeping
2. My sleep is barely disturbed (less than 1 hour sleepless)
3. My sleep is mildly disturbed (1-2 hours sleepless)
4. My sleep is moderately disturbed (2-3 hours sleepless)
5. My sleep is greatly disturbed (3-5 hours sleepless)
6. My sleep is completely disturbed (5-7 hours sleepless)

Q47 Recreation

1. I am able to engage in all recreational activities, with no pain in my neck at all
2. I am able to engage in all recreational activities, with some pain in my neck
3. I am able to engage in most, but not all of my usual recreational activities, because of pain in my neck
4. I am able to engage in a few of my usual recreational activities because of pain in my neck.
4. I can hardly do any recreational activities, because of pain in my neck
5. I cannot do any recreational activities at all, because of pain in my neck

---

Tampa Scale

Please read each statement and circle the most appropriate answer (see instructions on page 2).

	Strongly Disagree	Disagree	Agree	Strongly Agree
48. I'm afraid that I might injure myself if I exercise	**1**	**2**	**3**	**4**

	Strongly Disagree	Disagree	Agree	Strongly Agree
49. If I were to try to overcome it, my pain would increase	**1**	**2**	**3**	**4**
50. My body is telling me I have something dangerously wrong	**1**	**2**	**3**	**4**
51. My pain would probably be relieved if I were to exercise	**1**	**2**	**3**	**4**
52. People aren't taking my medical condition seriously enough	**1**	**2**	**3**	**4**
53. My accident has put my body at risk for the rest of my life	**1**	**2**	**3**	**4**
54. Pain always means I have injured my body	**1**	**2**	**3**	**4**
55. Just because something aggravates my pain does not mean it is dangerous	**1**	**2**	**3**	**4**
56. I am afraid that I might injure myself accidentally	**1**	**2**	**3**	**4**
57. Simply being careful that I do not make any unnecessary movements is the safest thing I can do to prevent my pain from worsening	**1**	**2**	**3**	**4**
58. I wouldn't have this much pain if there weren't something potentially dangerous going on in my body	**1**	**2**	**3**	**4**
59. Although my condition is painful, I would be better off if I were physically active	**1**	**2**	**3**	**4**
60. Pain lets me know when to stop exercising so that I don't injure myself	**1**	**2**	**3**	**4**
61. It's really not safe for a person with a condition like mine to be physically active	**1**	**2**	**3**	**4**
62. I can't do all the things normal people do because it's too easy for me to get injured	**1**	**2**	**3**	**4**
63. Even though something is causing me a lot of pain, I don't think it's actually dangerous	**1**	**2**	**3**	**4**
64. No one should have to exercise when he/she is in pain	**1**	**2**	**3**	**4**

Self Efficacy

Please rate how confident you are that you can do the following things at present, despite the pain. To answer circle one of the numbers on the scale under each item, where 0 = 'not at all confident' and 6 = 'completely confident'.

Remember, this questionnaire is not asking whether or not you have been doing these things, but rather how confident you are that you can do them at present, despite the pain.

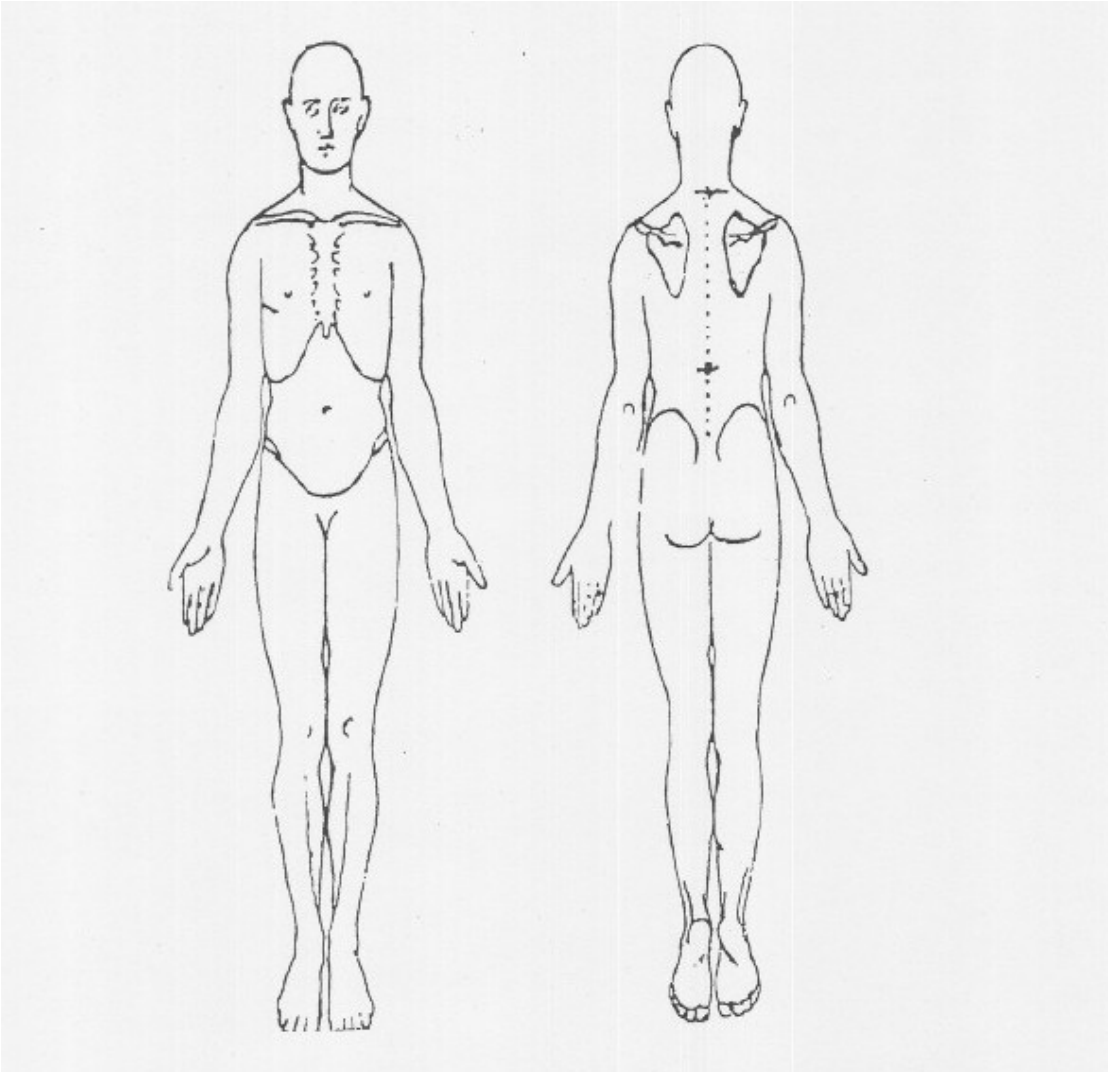
	Not all confident						**Completely confident**
65. I can enjoy things, despite the pain	**0**	**1**	**2**	**3**	**4**	**5**	**6**
66. I can do most of the household chores (e.g. tidying-up, washing dishes, etc.), despite the pain	**0**	**1**	**2**	**3**	**4**	**5**	**6**
67. I can socialise with my friends or family members as often as I used to do, despite the pain	**0**	**1**	**2**	**3**	**4**	**5**	**6**
68. I can cope with my pain in most situations	**0**	**1**	**2**	**3**	**4**	**5**	**6**
69. I can do some form of work, despite the pain. ('Work' includes housework, paid and unpaid work).	**0**	**1**	**2**	**3**	**4**	**5**	**6**
70. I can still do many of the things I enjoy doing, such as hobbies or leisure activity, despite the pain	**0**	**1**	**2**	**3**	**4**	**5**	**6**
71. I can cope with my pain without medication	**0**	**1**	**2**	**3**	**4**	**5**	**6**
72. I can still accomplish most of my goals in life, despite the pain	**0**	**1**	**2**	**3**	**4**	**5**	**6**
73. I can live a normal lifestyle, despite the pain	**0**	**1**	**2**	**3**	**4**	**5**	**6**
74. I can gradually become more active, despite the pain	**0**	**1**	**2**	**3**	**4**	**5**	**6**

Thank you very much for your time.
Please check you have completed all questions.

Patient ID:
DOB:
Research clinician

Appendix 5.11 Body diagrams.

SYMPTOMS THIS EPISODE

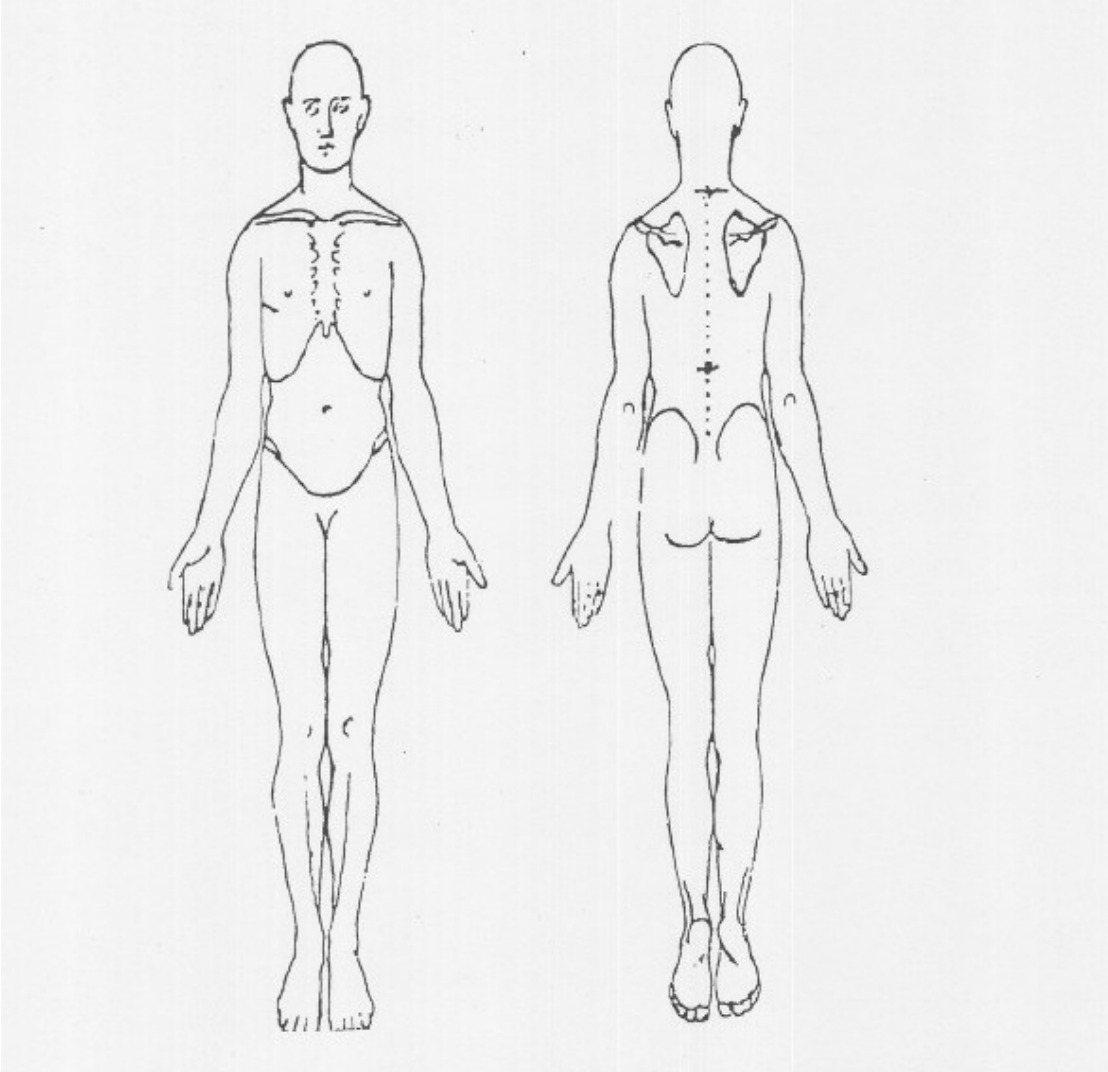


Where have you had discomfort/symptoms and what kind of discomfort/ symptoms at **this episode**?

Please shade all areas where you experience pain/ discomfort **at this episode** and describe the kind of discomfort/pain next to the figure: pins and needles, aching, cramp etc

Patient ID:
DOB:
Research clinician:

SYMPTOMS BEFORE TESTING

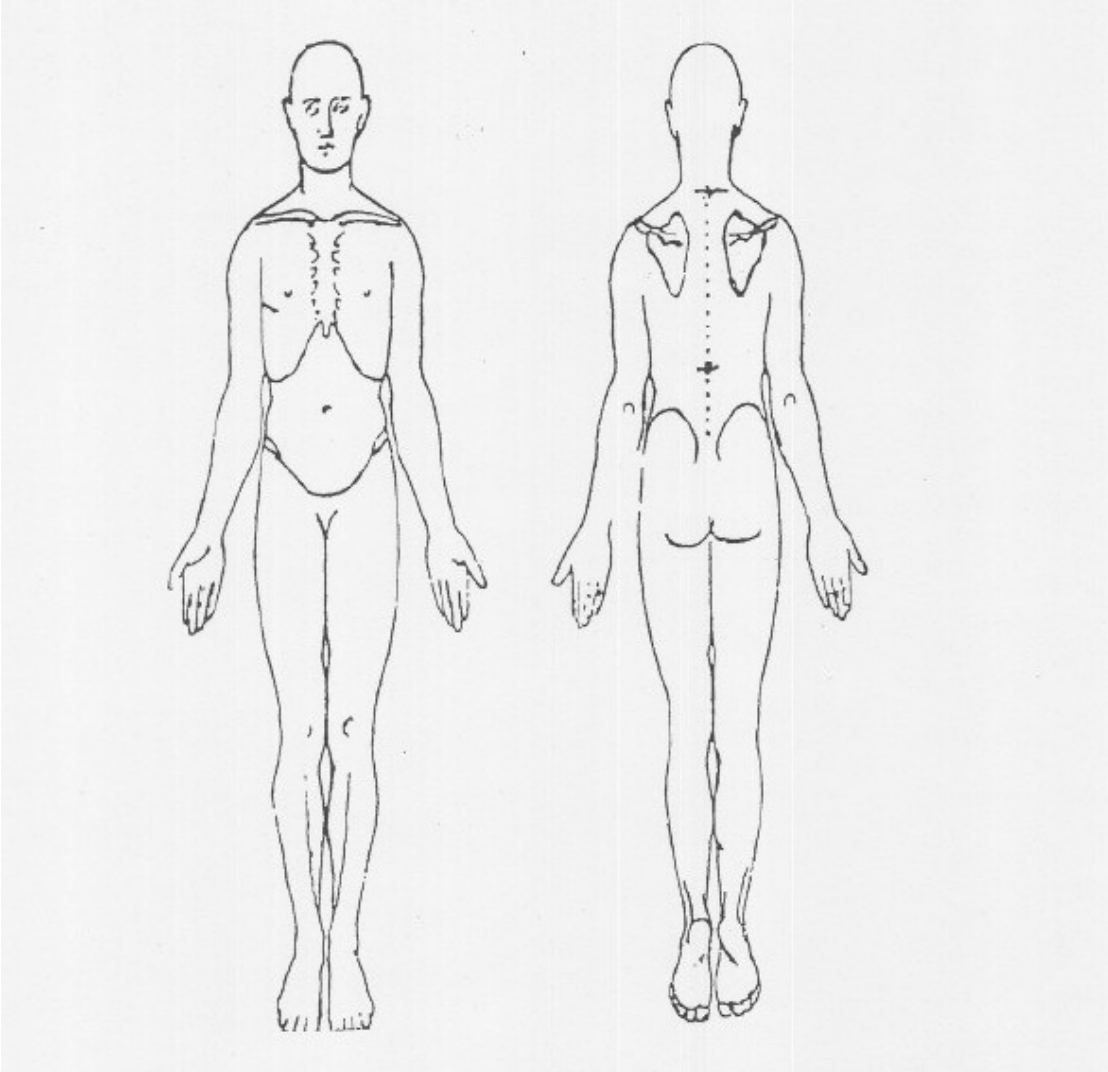


Where do you have discomfort/symptoms and what kind of discomfort **at the moment**?

Please shade all areas where you experience pain discomfort **at the moment** and describe the kind of discomfort next to the figure: pins and needles, aching, cramp etc

Patient ID:
DOB:
Research clinician:

SYMPTOMS AFTER TESTING



Where do you have discomfort/symptoms and what kind of discomfort **at the moment**?

Please shade all areas where you experience pain discomfort **at the moment** and describe the kind of discomfort next to the figure: pins and needles, aching, cramp etc

Appendix 5.12 Advice session instructions.

Section 5: Advice (1)

5.1 Objective of the session:
Encourage the participant to stay active

5.2 Method
One to one session

5.3 Basic messages (based on the neck book):
Keep moving
Do not stay in one position for long
Move about before you stiffen up
Move a little further and faster each day
Don't stop doing things – just change the way you do them

5.4 Attitude
- Show sympathy for their problem
- Give recognition that it may be difficult to handle their problem
- Promote taking an active role to reduce their symptoms and disabilities

5.5 Start of the session
1. Explain the objective of the session.
2. Ask them about difficulties they may experience in their daily lives.
3. Advise them to handle their difficulties in an optimal way, for example
    a. Working for a long time in same position: try to cut down the work in bouts of 20 minutes and do active exercises for 5 minutes
    b. Work in awkward position: try to solve this with other positions, but be aware that this might not be possible. Eyes determine the working posture, so a person will adjust his posture up to optimal sight.
    c. Handle of weights: carry things hugged to the body or split it over two hands, if not possible try to use supporting equipment (e.g. trolley).
    d. Sitting: try to sit comfortably and if the arm or shoulder hurts, use support of pillows if necessary.
    e. Desk work: Explain basic rules of working at a desk with a computer
    f. Reading: if reading gives problems, try to adjust reading height by putting a folder under the book or document or at home use pillows.

    Technique: interactive, request examples, use body postures to explain.

4. Determine if they have any difficulties in coping with pain
    Label positive behaviour

5. Instruct how to perform exercises for their pain
    a. Use the exercises given in the leaflet

6. Ask if there are any questions related to the subjects discussed.

7. Provide the participant with the neck book.

Participants will be advised to act on this strategy and try refrain from other treatments for a 3 month period. Please try to familiarise yourself with the advice booklet given so that the patient is not confused by conflicting information and use the advice log for recording of the details of the session.

References
1. Luime J. Study protocol: Pilot Study Early Rehabilitation of Neck and Shoulder Pain. Coventry: University of Warwick; 2005.

Appendix 5.13 GP letter.

Warwick Medical School

Date

GP Name & Address

Dear Dr. GP name,

Symptom response pilot study: reliability and role in predicting outcome in neck pain

At the Warwick Medical School, one of our research interests relates to the prevention and management of musculoskeletal conditions. This letter is to inform you that your patient, patient name, DOB is involved in the study 'Symptom response pilot study: reliability and role in predicting outcome in neck pain' (Neck Pain Assessment and VideoNeck study).

The study is recruiting employees with neck pain (with or without referred complaints). We are looking at the value of observed symptom changes in the clinical assessment of patients with neck pain with the aim to help improve clinical decision making and the management of neck pain. All participants will receive information on pain, self-management, exercises, and how to carry on with daily activities. We hope that this information will help patients cope with their condition.

If you have any questions about this research, please feel free to contact Ms Angeliki Chorti (details given below).

Yours sincerely,

Angeliki Chorti
Warwick Medical School
University of Warwick, Coventry CV4 7AL
Tel: 024 761 50405,
Email: A.Chorti@warwick.ac.uk

Study number:

Patient ID:

Appendix 5.14 Event notification form.

Symptom response pilot study: Reliability and role in predicting outcome in neck pain

---

Event notification form

Telephone immediately with any notification of withdrawal, death, serious adverse event or complaint*

Ms Angeliki Chorti on 024 761 50405

Or

Dr. Chris McCarthy 024 765 75856

Completed by..... Date of completion …../…../.... …..

Participant details:

Participant ID number |__|__|__| |__|__| |__|__|__|__|

Date of birth |__|__| |__|__| |__|__| (dd/mm/yy)

Only include the participant's name or address if used to notify a name or address change

Old name	New name
……	……
Old address	New address
……	……
……	……
……	……
……	……

Study number:

Patient ID:

---

1. Patient request for withdrawal from *assessment or advice* (see also next item '2')

Date request received |__|__| |__|__| |__|__| (dd/mm/yy)

Reason for request (if given)
……..
Date Warwick informed |__|__| |__|__| |__|__| (dd/mm/yy) *Please phone us with this information as soon as possible*

---

2. Patient request for withdrawal from *follow-up* (i.e., from receiving questionnaires or telephone calls)

Date request received |__|__| |__|__| |__|__| (dd/mm/yy)

Reason for request (if given)
.....
.

Date Warwick informed |__|__| |__|__| |__|__| (dd/mm/yy) *Please phone us with this information as soon as possible*

---

3. Practice request for patient to be withdrawn from *follow-up* (i.e., from receiving questionnaires or telephone calls)

Date request received |__|__| |__|__| |__|__| (dd/mm/yy)

Reason for request (if given)
.....
.

Date Warwick informed |__|__| |__|__| |__|__| (dd/mm/yy) *Please phone us with this information as soon as possible*

---

4. Death notification

Date of death |__|__| |__|__| |__|__| (dd/mm/yy)

Cause of death (if known)…….....…….

Date Warwick informed |__|__| |__|__| |__|__| (dd/mm/yy) *Please phone us with this information as soon as possible*

Message taken by (at Warwick) .....
……

---

5. Serious adverse event notification

Date of event |__|__| |__|__| |__|__| (dd/mm/yy)

Nature of possible adverse event.....……..

.....
.
Date Warwick informed |__|__| |__|__| |__|__| (dd/mm/yy) *Please phone us with this information as soon as possible*

Source of notification of possible adverse event.…….

Message taken by (at Warwick).....……..……

---

6. Complaint notification

Date of notification |__|__| |__|__| |__|__| (dd/mm/yy)

Nature of complaint.....…

.....
.
Date Warwick informed |__|__| |__|__| |__|__| (dd/mm/yy) *Please phone us with this information as soon as possible*

Source of notification of complaint.……..…….

Message taken by (at Warwick)
……

Completed by …… (Block capitals)

Date of completion |__|__| |__|__| |__|__| (dd/mm/yy)

Notes

Participant details

Participant ID number |__|__|__| |__|__| |__|__|__|__|

Date of birth |__|__| |__|__| |__|__| (dd/mm/yy)

Appendix 5.15 Raw data from the reliability study. Note: B, category 'better' determined by changes in distal symptom location; 'CP_L', category 'centralisation' determined by changes in distal symptom location; NC, category 'no change' determined by changes in distal symptom location; 'non-CP_L', category 'non-centralisation' determined by changes in distal symptom location; κ, kappa statistic; '$Other_L$', category 'other' determined by changes in distal symptom location; W, category 'worse' determined by changes in distal symptom location.

For physiotherapist pair 1, κ_1 = 0.86 (0.59 to 1.00), total agreement 92% for classifications based on changes in location ('CP_L', 'non-CP_L', '$Other_L$'); κ_1 = 0.22 (-0.23 to 0.66), 50% total agreement for classifications based on changes in location and / or intensity ('B', 'W', NC').

3x3 table of results for changes in location for physiotherapist pair 1 (n=12).

		Physiotherapist 2			
		CP_L	**Non-CP_L**	**$Other_L$**	**Total**
Physiotherapist 1	**CP_L**	1	1	0	2
	Non-CP_L	0	6	0	6
	$Other_L$	0	0	4	4
	Total	1	7	4	12

Note: Frequencies per category: CP_L= 0.17; Non-CP_L =3.50; $Other_L$ = 1.33´Total/Total = 0.42; se (κ) = 0.14.

Abbreviations: n, number of participants; se (κ), standard error of κ.

3x3 table of results for changes in location and / or intensity for physiotherapist pair 1 (n=12).

		Physiotherapist 2			
		B	**W**	**NC**	**Total**
Physiotherapist 1	**B**	2	2	0	4
	W	0	0	3	3
	NC	0	1	4	5
	Total	2	3	7	12

Note: Frequencies per category: B = 0.67; W =0.75; NC = 2.92´Total/Total = 0.36; se (κ) = 0.23.

Abbreviations: n, number of participants; se (κ), standard error of κ.

For physiotherapist pair 2, κ_2 = 0.49 (0.10 to 0.87), total agreement 74% for classifications based on changes in location ('CP_L', 'non-CP_L', '$Other_L$'); κ_2 = 0.04 (-0.32 to 0.41), 38% total agreement for classifications based on changes in location and / or intensity ('B', 'W', NC').

3x3 table of results for changes in location for physiotherapist pair 2 (n=19).

		Physiotherapist 2			
		CP_L	**Non-CP_L**	**$Other_L$**	**Total**
Physiotherapist 1	**CP_L**	1	1	2	4
	Non-CP_L	1	11	0	12
	$Other_L$	0	1	2	3
	Total	2	13	4	19

Note: Frequencies per category: CP_L = 0.42; Non-CP_L =8.21; $Other_L$ = 0.63´Total/Total = 0.49; se (κ) = 0.20.

Abbreviations: n, number of participants; se (κ), standard error of κ.

3x3 table of results for changes in location and / or intensity for physiotherapist pair 2 (n=16)*

		Physiotherapist 2			
		B	**W**	**NC**	**Total**
Physiotherapist 1	**B**	4	4	1	9
	W	1	1	3	5
	NC	1	0	1	2
	Total	6	5	5	16

*3 unable to classify

Note: Frequencies per category: B = 3.38; W =1.56; NC = 0.63´Total/Total = 0.35; se (κ) = 0.19.

Abbreviations: n, number of participants; se (κ), standard error of κ.

For physiotherapist pair 3, κ_3 = 0.06 (-0.69 to 0.81), total agreement 67% for classifications based on changes in location ('CP_L', 'non-CP_L', '$Other_L$'); κ_3 = 0.47 (0.04 to 0.89), 67% total agreement for classifications based on changes in location and / or intensity ('B', 'W', NC').

3x3 table of results for changes in location for physiotherapist pair 3 (n=12).

		Physiotherapist 2			
		CP_L	**Non-CP_L**	**$Other_L$**	**Total**
Physiotherapist 1	**CP_L**	0	1	1	2
	Non-CP_L	1	8	0	9
	$Other_L$	0	1	0	1
	Total	1	10	1	12

Note: Frequencies per category: CP_L = 0.17; Non-CP_L =7.50; $Other_L$ = 0.08´Total/Total = 0.65; se (κ) = 0.81.

Abbreviations: n, number of participants; se (κ), standard error of κ.

3x3 table of results for changes in location and / or intensity for physiotherapist pair 3 (n=12).

		Physiotherapist 2			
		B	**W**	**NC**	**Total**
Physiotherapist 1	**B**	5	0	1	6
	W	0	2	1	3
	NC	1	1	1	3
	Total	6	3	3	12

Note: Frequencies per category: B = 3.00; W =0.75; NC = 0.75´Total/Total = 0.38; se (κ) = 0.22.

Abbreviations: n, number of participants; se (κ), standard error of κ.

For physiotherapist pair 4, κ_4 = 1.00 (1.00 to 1.00), total agreement 100% for classifications based on changes in location ('CP_L', 'non-CP_L', '$Other_L$'); κ_4 = 0.61 (-0.09 to 1), 86% total agreement for classifications based on changes in location and / or intensity ('B', 'W', NC').

3x3 table of results for changes in location for physiotherapist pair 4 (n=7).

		Physiotherapist 2			
		CP_L	**Non-CP_L**	**$Other_L$**	**Total**
Physiotherapist 1	**CP_L**	2	0	0	2
	Non-CP_L	0	5	0	5
	$Other_L$	0	0	0	0
	Total	2	5	0	7

Note: Frequencies per category: CP_L= 0.57; Non-CP_L =3.57; $Other_L$ = 0.00´Total/Total = 0.59; se (κ) = 0.00.

Abbreviations: n, number of participants; se (κ), standard error of κ

3x3 table of results for changes in location and / or intensity for physiotherapist pair 4 (n=7).

		Physiotherapist 2			
		B	**W**	**NC**	**Total**
Physiotherapist 1	**B**	5	0	0	5
	W	0	1	0	1
	NC	1	0	0	1
	Total	6	1	0	7

Note: Frequencies per category: B = 4.29; W =0.14; NC = 0.00´Total/Total = 0.63; se (κ) = 0.36.

Abbreviations: n, number of participants; se (κ), standard error of κ

Classifications on 48 participants

Category based on location

ID	CP_L	Non-CP_L	$Other_L$	Pi
1	0	2	0	1
2	0	0	2	1
3	2	0	0	1
4	2	2	0	3
5	0	2	0	1
6	0	0	2	1
7	0	0	2	1
8	1	1	0	0
9	0	2	0	1
10	0	2	0	1
11	0	2	0	0
12	0	2	0	1
13	0	2	0	0
14	1	1	0	0
15	0	2	0	1
16	0	2	0	1
17	0	1	1	0
18	1	1	0	0
19	0	2	0	1
20	0	2	0	1
21	0	2	0	1
22	0	2	0	1
23	0	0	2	1
24	1	0	1	0
25	0	2	0	1
26	0	2	0	1
27	2	0	0	1
28	0	2	0	1
29	1	1	0	0
30	0	2	0	1

Classifications on 45 participants

Category based on location and/or intensity

ID	B	W	NC	Pi
1	0	1	1	0
2	0	0	2	1
3	2	0	0	1
4	0	1	1	0
5	0	0	0	0
6	0	1	1	0
7	0	0	2	1
8	2	0	0	1
9	2	0	0	1
10	2	0	0	1
11	1	0	1	0
12	0	0	2	1
13	0	1	1	0
14	2	0	0	1
15	2	0	0	1
16	1	1	0	0
17	0	0	2	1
18	1	1	0	0
19	0	2	0	1
20	0	2	0	1
21	0	0	0	0
22	1	0	1	0
23	0	1	1	0
24	1	1	0	0
25	0	1	1	0
26	0	1	1	0
27	2	0	0	1
28	0	2	0	1
29	2	0	0	1
30	0	2	0	1

31	0	2	0	1
32	0	2	0	1
33	2	0	0	1
34	2	0	0	1
35	0	2	0	1
36	0	2	0	1
37	0	2	0	1
38	0	2	0	1
39	0	0	2	1
40	0	2	0	1
41	0	2	0	1
42	0	2	0	1
43	1	1	0	0
44	0	2	0	1
45	0	0	2	1
46	0	2	0	1
47	0	2	0	1
48	1	0	1	0
49	0	1	1	0
50	1	0	1	0
Total	14	63	17	total 94
Pj	0.149	0.670	0.181	
P	40			
Pa	0.833			
Pe	0.504			
		Se(k) = 0.046		
K=	**0.664**	K/SE(K)= 14.43	p<0.00003	

31	1	0	1	0
32	2	0	0	1
33	2	0	0	1
34	2	0	0	1
35	0	1	1	0
36	1	1	0	0
37	2	0	0	1
38	2	0	0	1
39	0	0	2	1
40	0	1	1	0
41	1	1	0	0
42	2	0	0	1
43	1	1	0	0
44	2	0	0	1
45	0	0	2	1
46	1	0	1	0
47	2	0	0	1
48	1	1	0	0
49	0	0	0	0
50	1	0	1	0
Total	43	23	24	total 90
Pj	0.483	0.258	0.258	
P	26			
Pa	0.578			
Pe	0.366			
		Se(k) = 0.113		
		k/se(k) = 2.96	p<0.0015	
K=	**0.334**			

Note: ID, Identification number; k, kappa statistic; se(k), standard error of kappa.

Statistics for measuring agreement on each of the three categories

Category	Σn_1^2	Pj	P	**κ**	$Var(\kappa)_{1971}$	κ/ SE(κ)	p
Symptom location							
CP_L	28	0.149	0.957	**0.949**	0.159	2.38*	p < 0.0087
Non-CP_L	115	0.670	0.788	**0.357**	0.279	0.68	p< 0.2483
$Other_L$	29	0.181	0.669	**0.596**	0.151	1.53	p<0.0630
Symptom location and / or intensity							
B	75	0.483	0.725	**0.468**	0.194	1.06	P<0.1446
W	31	0.258	0.335	**0.104**	0.156	0.26	P<0.3974
NC	35	0.258	0.507	**0.336**	0.156	0.85	P<0.1977

*indicates significant result.

Note: B, category 'better' determined by changes in distal symptom location; 'CP_L', category 'centralisation' determined by changes in distal symptom location; NC, category 'no change' determined by changes in distal symptom location; 'non-CP_L', category 'non-centralisation' determined by changes in distal symptom location; κ, kappa statistic; '$Other_L$', category 'other' determined by changes in distal symptom location; p, p value; P, probability; Pj, conditional probability; se (κ), standard error of κ; $Var(\kappa)_{1971,}$ approximate variance of κ for a specific category; W, category 'worse' determined by changes in distal symptom location.