Musculoskeletal Medicine in The Netherlands

Characteristics of patients and physicians
and validity of outcome measurement instruments

Wouter Schuller

VRIJE UNIVERSITEIT

Musculoskeletal Medicine in The Netherlands

Characteristics of patients and physicians, and validity of outcome measurement instruments

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor of Philosophy
aan de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. V. Subramaniam,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Faculteit der Geneeskunde
op dinsdag 8 december 2020 om 9.45 uur
in de aula van de universiteit,
De Boelelaan 1105

door

Wouter Schuller

geboren te Amsterdam

promotoren: prof.dr.ir. H.C.W. de Vet
prof.dr. R.W.J.G. Ostelo

copromotor: dr. C.B. Terwee

Table of Contents

Chapter 1 Introduction 5

Chapter 2 Physicians using spinal manipulative treatment in The Netherlands: a description of their characteristics and their patients 13

Chapter 3 Pain trajectories and predictors of a favourable course of low back pain in patients consulting musculoskeletal physicians in The Netherlands 29

Chapter 4 Adverse events after spinal manipulative treatment by musculoskeletal physicians in The Netherlands 51

Chapter 5 Smallest Detectable Change and Minimal Important Change of the Neck Disability Index were influenced by population characteristics 67

Chapter 6 Measurement properties of the Dutch-Flemish PROMIS Pain Behaviour item bank in patients with musculoskeletal complaints 81

Chapter 7 Validation of the Dutch-Flemish PROMIS Pain Interference item bank in patients with musculoskeletal complaints..... 101

Chapter 8 General discussion 119

Chapter 9 General summary.....137

Samenvatting..... 143

Dankwoord 149

About the author 152

colofon

cover design
buro RuSt

cover art
Danse Macabre fresco in the Holy Trinity Church, Hrastovlje, Slovenia

layout
Coco Bookmedia, Amersfoort

printing
Wilco BV, Amersfoort

Chapter 1.

Introduction

Background

In The Netherlands, a group of physicians has specialised in Musculoskeletal (MSK) Medicine. These physicians are organised in their own professional organisation, the Dutch Association for Musculoskeletal Medicine (Nederlandse Vereniging voor Artsen Muskuloskeletale Geneeskunde, NVAMG). MSK physicians are mainly concerned with pain and function of the locomotor system, with special attention for complaints of spinal origin. In 2000 the prevalence of musculoskeletal complaints in the Dutch general population was evaluated. Almost three-quarter of the Dutch population reported any musculoskeletal complaint during the past 12 months, most frequently low back pain (43.9%), neck pain (31.4%), and shoulder pain (30.3%), but also knee pain (21.9%), pain in the higher back (18.8%) and pain in the wrist or hand (17.5%)(1). The prevalence of musculoskeletal complaints is high, especially complaints of spinal origin that constitute a large part of the patient population consulting MSK physicians. Over the past decade, the educational programme to become registered as MSK physician has gradually been expanded with more extensive knowledge of neurologic and orthopaedic diagnostic and treatment possibilities, diagnostic imaging, and with invasive treatment options such as applying injections in the spine under X-ray guidance. A key part of the educational programme concerns the use of Spinal Manipulative Treatment (SMT). SMT is a well-known treatment option for spinal/musculoskeletal disorders, but is generally associated with manual therapists and chiropractors. The fact that there are medical doctors professionally concerned with SMT is less known. Primary care guidelines mention SMT as a possible treatment option for conditions such as low back pain, lumbosacral radiculopathies, neck pain and cervicogenic headache(2-4). Other possible treatment options include reassurance and the advice to stay active, exercise treatment, postural corrections, clinical massage, but also medical interventions, such as medication, pain intervention or surgery. The guidelines of the Dutch Federation of Medical Specialists suggest epidural steroid injections, facet denervation or surgery as a treatment option for patients with low back pain and or lumbar radiculopathies(5, 6). The decision as to which treatment options are used is influenced by the clinical background and the specific expertise of the health care professional consulted(7). With their medical background and specific post-graduate education it is likely that MSK physicians treat different patient populations, and use different treatment strategies than other professionals concerned with disorders of the locomotor system. A lot of research has been conducted in the fields of physiotherapy, manual therapy and chiropractic, but there are no studies yet that have addressed MSK medicine. There is a need, therefore, to study the characteristics of MSK physicians and their preferences concerning possible treatment options, to evaluate the characteristics of their patients, and to study the course of patients' complaints after consulting MSK physicians. We studied the characteristics of MSK physicians with a survey distributed to all registered

members of the NVAMG, and we studied patient characteristics and the course of patients complaints in a large observational cohort study. The study was funded almost entirely by the Dutch Association for Musculoskeletal Medicine itself, and was facilitated by the former EMGO research institute of the VU medical centre.

Data collection

To enable a large observational cohort study with limited funding, part of the data collection was automated. The first step in our study set-up was to build a web-based register for the physicians. In this web-based register, MSK physicians could enter basic details about all new patients consulting them, including type and duration of the main complaint, and the existence of concomitant complaints. At the end of treatment the physician was asked to enter further details about the type of treatment administered, and the number of treatment sessions. The second step was to build a web-based register to collect data from the patients. In the register filled in by the physician a field was added to enter the email address of the patient. This email address was automatically transferred to a server at the VU. On this server, a custom build programme (Readmail) stored this email address in a database, from where it was used to send emails to patients with a link to questionnaires. This approach for data collection was approved by the Medical Ethical Research Committee (METc).

The third step was to engage MSK physicians to participate in the study. Being initiated by the NVAMG itself there was a lot of attention for the study within the group of MSK physicians, and all 138 registered MSK physicians were invited to participate. A group of 31 physicians eventually decided to take part in the study. They were instructed in special meetings on how to use the register. In this way, data-collection was automated, opening ways to collect a multitude of data from a large population of patients. While the physician would fill in the same register during the whole study period, the researchers could choose and change the type of patient questionnaires and the follow-up intervals.

This study set-up was used in four different phases. In each phase different data were collected using a variety of patient questionnaires. And in each consecutive phase, the complexity of the measurement programme increased. In the first phase, only the web-based register was used, collecting data from the physicians. This phase was used to collect general data about the patients presenting in MSK practice. In the second phase, data collection from the patient was started, using the Readmail programme, with only a single follow-up questionnaire after a follow-up period of three months. This phase was used to test the programme, and to collect data about global perceived effect and patient satisfaction. In the third phase, more patient questionnaires were added, at baseline and after a follow-up period of three months.

This phase was used to evaluate psychometric properties of newly developed (PROMIS) questionnaires. In the fourth phase, a much more complex measurement programme was used, including several follow-up intervals, and using a flow chart in which questionnaires were tailored to the patients' main complaint. This phase was used to collect data about the course of patients' complaints, and about adverse events experienced after MSK treatment. The different phases are described in more detail in the first paper of this thesis.

Research topics

First topic of this thesis was to study the characteristics of MSK physicians, to evaluate the characteristics of their patients, and to study the course of patients complaints after consulting MSK physicians. For this purpose a survey was conducted among MSK physicians, and a large observational cohort study was set-up to measure patient characteristics, and to measure the course of their complaints after consulting MSK physicians. The survey focused on the clinical training and experience, the type of diagnostic and treatment options used, and further referral to other health care providers. This information could be used to evaluate the characteristics of MSK physicians. The observational cohort study focused on baseline characteristic of patients and the course of their complaints after consulting with an MSK physician. This information could be used to characterise the patient population, to evaluate baseline variables as possible predictors of a favourable course, and to evaluate adverse events after MSK treatment. The results of a descriptive study of the characteristics of MSK physicians and their patients are presented in chapter 2. Chapter 3 concerns the course of complaint in patients with low back pain after consulting an MSK physician. Chapter 4 concerns a descriptive study evaluating adverse events reported by patients after consulting MSK physicians.

The second topic of this thesis was the validation of questionnaires used in musculoskeletal research. Existing questionnaires may need re-evaluation and new questionnaires have to be validated(8, 9). Many existing questionnaires have been developed more than 20 years ago. These questionnaires may be outdated in terms of content or methodologic requirements and they often have not been validated properly. Furthermore, new questionnaires have to be validated. As a part of this validation process questionnaires have to be tested on a study population to assess whether they measure what they are supposed to measure. Can they identify patients with different levels of a certain condition (or trait)? And how do they compare to other, known questionnaires? An exciting development in clinimetrics is the use of item banks based upon Item Response Theory (IRT). Item banks are large collections of questions (items), measuring a specific construct, such as pain interference, pain behaviour, or physical functioning. They are calibrated on a population including the general population

and clinical samples. In this way, item banks are supposed to cover the whole range of the construct on a common metric, with scales that are centred around a general population average. Once calibrated, subsets of items can be used giving scores on the same common metric(10-13). Item banks can be used in Computer Adaptive Testing (CAT), in which a computer algorithm decides on the basis of previous answers which question would be most informative next. In this way, questions can be tailored to the individual patient, and only a few questions need to be answered to arrive at a reliable score(14-17). Unidimensionality, local independency, and monotonicity are important assumptions in IRT item banks(9) that are considered necessary in order to be able to scale the items on the same interval scale. These assumptions need to be evaluated before fitting an IRT model. The fit of this model and the fit of individual items to this model is evaluated. Further evaluation of IRT item banks concerns analyses of Differential Item Functioning (DIF). DIF analyses may reveal whether different groups of patients have different interpretations of the items. To be able to compare scores between these groups the amount of DIF should be limited. Several item banks, based upon IRT, were recently developed by the PROMIS (Patient Reported Outcome Measurement Information System) initiative(18). A number of these item banks have been translated into Dutch-Flemish(19).

Chapter 5 concerns a study evaluating the psychometric properties of the Neck Disability Index (NDI), developed by Vernon in 1991, and frequently used to study the course of patients with neck pain treated by means of SMT(20, 21). Chapter 6 and 7 concern studies in which the psychometric properties of the PROMIS Pain Behaviour(22) and the PROMIS Pain Interference(23) item banks were evaluated.

Research questions

1. How can MSK physicians be characterised? What is their background, how were they trained, what type of treatments do they apply, and what is the profile of their patients?
2. What is the Smallest Detectable Change and the Minimal Important Change of the NDI in patients treated by MSK physicians? How does this relate to other reports in literature?
3. What are the psychometric properties of the PROMIS Pain Interference item bank?
4. What are the psychometric properties of the PROMIS Pain Behaviour item bank?
5. What is the course of pain in patients with low back pain treated by MSK physicians? Can predictors of a favourable course be identified?
6. Which adverse events are reported by patients after MSK treatment?

References

1. Picavet HS, Schouten JS. Musculoskeletal pain in the Netherlands: prevalences, consequences and risk groups, the DMC(3)-study. Pain. 2003;102(1-2):167-78.
2. Cohen SP, Hooten WM. Advances in the diagnosis and management of neck pain. BMJ. 2017;358:j3221.
3. Cote P, Yu H, Shearer HM, Randhawa K, Wong JJ, Mior S, et al. Non-pharmacological management of persistent headaches associated with neck pain: A clinical practice guideline from the Ontario protocol for traffic injury management (OPTIMa) collaboration. Eur J Pain. 2019;23(6):1051-70.
4. Wong JJ, Cote P, Sutton DA, Randhawa K, Yu H, Varatharajan S, et al. Clinical practice guidelines for the noninvasive management of low back pain: A systematic review by the Ontario Protocol for Traffic Injury Management (OPTIMa) Collaboration. Eur J Pain. 2017;21(2):201-16.
5. Federatie Medsich Specialisten . Richtlijn Lumbo-Sacraal Radiculair Syndroom. 2008.
6. Federatie Medisch Specialisten . Wervelkolom gerelateerde pijnklachten van de lage rug. 2012.
7. Cherkin DC, Deyo RA, Wheeler K, Ciol MA. Physician variation in diagnostic testing for low back pain. Who you see is what you get. Arthritis Rheum. 1994;37(1):15-22.
8. de Vet HC, Terwee CB, Mokkink LB, Knol D. Measurement in Medicine. Cambridge: Cambridge University Press; 2011.
9. Streiner DL, Norman GR. Health measurement scales. Oxford: Oxford University Press; 2008
10. Fries J, Rose M, Krishnan E. The PROMIS of better outcome assessment: responsiveness, floor and ceiling effects, and Internet administration. J Rheumatol. 2011;38(8):1759-64.
11. Fries JF, Krishnan E, Rose M, Lingala B, Bruce B. Improved responsiveness and reduced sample size requirements of PROMIS physical function scales with item response theory. Arthritis Res Ther. 2011;13(5):R147.
12. Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. Med Care. 2000;38(9 Suppl):II28-II42.
13. Tugwell P, Knottnerus JA, Idzerda L. Tailoring patient reported outcome measurement. J Clin Epidemiol. 2010;63(11):1165-6.
14. Chakravarty EF, Bjorner JB, Fries JF. Improving patient reported outcomes using item response theory and computerized adaptive testing. J Rheumatol. 2007;34(6):1426-31.
15. Fayers PM. Applying item response theory and computer adaptive testing: the challenges for health outcomes assessment. Qual Life Res. 2007;16 Suppl 1:187-94.
16. Haley SM, Ni P, Hambleton RK, Slavin MD, Jette AM. Computer adaptive testing improved accuracy and precision of scores over random item selection in a physical functioning item bank. J Clin Epidemiol. 2006;59(11):1174-82.
17. Reise SP, Waller NG. Item response theory and clinical measurement. Annu Rev Clin Psychol. 2009;5:27-48.

18. Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. J Clin Epidemiol. 2010;63(11):1179-94.
19. Terwee CB, Roorda LD, de Vet HC, Dekker J, Westhovens R, van Leeuwen J, et al. Dutch-Flemish translation of 17 item banks from the patient-reported outcomes measurement information system (PROMIS). Qual Life Res. 2014;23(6):1733-41.
20. Vernon H. The Neck Disability Index: state-of-the-art, 1991-2008. J Manipulative Physiol Ther. 2008;31(7):491-502.
21. Vernon H, Mior S. The Neck Disability Index: a study of reliability and validity. J Manipulative Physiol Ther. 1991;14(7):409-15.
22. Revicki DA, Chen WH, Harnam N, Cook KF, Amtmann D, Callahan LF, et al. Development and psychometric analysis of the PROMIS pain behaviour item bank. Pain. 2009;146(1-2):158-69.
23. Amtmann D, Cook KF, Jensen MP, Chen WH, Choi S, Revicki D, et al. Development of a PROMIS item bank to measure pain interference. Pain. 2010;150(1):173-82.

Chapter 2.

Physicians using spinal manipulative treatment in The Netherlands: a description of their characteristics and their patients

Wouter Schuller, MD, Raymond W.J.G. Ostelo, PhD, Daphne C. Rohrich MSc, A.T. Apeldoorn, PhD, Henrica C.W. de Vet, PhD

BMC Musculoskeletal Disorders (2017), 18:512

Abstract

Background: Various health care professionals apply Spinal Manipulative Treatment (SMT) in daily practice. While the characteristics of chiropractors and manual therapists and the characteristics of their patient populations are well described, there is little research about physicians who use SMT techniques. A distinct group of physicians in The Netherlands has been trained in musculoskeletal (MSK) medicine, which includes the use of SMT. Our objective was to describe the characteristics of these physicians and their patient population.

Methods: All registered MSK physicians were approached with questionnaires and telephone interviews to collect data about their characteristics. Data about patient characteristics were extracted from a web-based register. In this register physicians recorded basic patient data (age, gender, the type and duration of the main complaint, concomitant complaints and the type of referral) at the first consultation. Patients were invited to fill in web-based questionnaires to provide baseline data about previous treatments and the severity of their main complaint. Functional impairment was measured with Patient Reported Outcome Measures (PROMs).

Results: Questionnaires were sent to 138 physicians of whom 90 responded (65%). Most physicians were trained in MSK medicine after a career in other medical specialities. They reported to combine their SMT treatment with a variety of diagnostic and treatment options part of which were only permissible for physicians, such as prescription medication and injections. The majority of patients presented with complaints of long duration (62.1% > 1 year), most frequently low back pain (48.1%) or neck pain (16.9%), with mean scores of 6.0 and 6.2, respectively, on a 0 to10 numerical rating scale (NRS) for pain intensity. Mean scores on all PROMs showed moderate impairment. Patients most frequently reported previous treatment by physical therapists (68.1%), manual therapists (37.7%) or chiropractors (17.0%).

Conclusion: Our study showed that MSK physicians in The Netherlands used an array of SMT techniques. They embedded their SMT techniques in a broad array of other diagnostic and treatment options, part of which were limited to medical doctors. Most patients consulted MSK physicians with spinal pain of long duration with moderate functional impairment.

Background

Spinal Manipulative Treatment (SMT) is used world-wide to treat musculoskeletal problems such as low back pain and neck pain(1). Given the socioeconomic impact of these conditions and the wide spread use of SMT, determining the efficacy of SMT is a priority for all health care stakeholders. However, determining the efficacy of SMT is challenging. Cochrane reviews for SMT in the treatment of neck pain and chronic low back pain have concluded that there is evidence for some effect, but the size of this effect is small(2-4). Outcomes may be influenced by the heterogeneity of the patient population, or by the clinical setting wherein SMT is used. That is, it is possible that SMT is only effective in subgroups of patients, or that the efficacy is influenced by the variety of clinical settings in which SMT techniques are applied by various health care professionals(5). Whilst SMT is generally associated with chiropractors and manual therapists, SMT techniques are also applied by groups of specially trained physicians. Currently, characteristics of chiropractors and manual therapists and their patients are well described(6-17); however, little is known about physicians trained in the use of SMT(18).

In The Netherlands, there is a group of physicians who have been trained in musculoskeletal (MSK) medicine, including the use of SMT. These physicians are titled "physician for musculoskeletal medicine" and united in the Dutch Association for Musculoskeletal Medicine. To obtain registration as a physician for MSK medicine a two year training program consisting of both theoretical and practical work must be successfully completed after qualifying as a medical doctor. The theoretical component covers specialist knowledge of manipulative treatments, orthopaedics, neurology, radiology, epidemiology, research methodology and medico-legal aspects. The practical training consists of working as a trainee at a designated training practice for at least two days a week for a period of two years. During this time the trainee specialises in at least one of two types of SMT techniques. One SMT technique, manual medicine, is mainly derived from chiropractic and manual therapeutic techniques, diagnosing and correcting limitations in segmental motion. The other SMT technique, orthomanual medicine, has been developed more recently in The Netherlands, and identifies and corrects alterations in joint positions. These joint positions are considered to be interconnected throughout the spine, and are corrected in a strict sequence of specific mobilising techniques. The technique has been shown to differ from manual therapy and chiropractic treatment(18).

The objective of our study was to describe the characteristics of physicians for musculoskeletal medicine in The Netherlands and the characteristics of their patients.

Methods

Study design

We conducted a descriptive study of the characteristics of Dutch MSK physicians and their patients. All members of the Dutch Association for Musculoskeletal Medicine (N=138) were invited to participate. First, we contacted the physicians by mail to participate in a survey to collect physician characteristics. In addition, we contacted all physicians by telephone to stimulate response. We asked participating physicians to provide written informed consent. Second, we asked physicians to participate in a web-based patient registry and to invite all consecutive patients who presented for the first time in MSK practice. If patients gave informed consent, the treating physician entered email addresses of the recruited patients in the registry. Thereafter, we used a specially designed computer program (Readmail) to automatically distribute invitations to patients by email to fill in web-based questionnaires.

During three consecutive time periods, this registry was used to collect data about patient characteristics. In each period different sets of outcome measures were used, resulting in three cohorts of patients with specific sets of outcome measures (Table 1).

Data collection of physician characteristics

We collected data about physician characteristics using a paper survey sent by mail. In this survey, physicians were asked about their age, gender, their medical background, additional training in other medical specialties, the use of specific techniques and cooperation with other healthcare providers. In addition, we contacted all physicians by telephone to collect data about the number of days per week spent in MSK practice (Table 1).

Data collection of patient characteristics

Both the treating physician and the individual patients provided data, which were recorded in the web-based registry (Table 1). The treating physicians registered the following baseline data of patients: age, gender, type and duration of the main complaint, and the existence of concomitant complaints. The treating physicians coded the main and concomitant complaints according to the International Classification of Primary Care (ICPC)(19).

Three consecutive cohorts of patients were presented with three different sets of baseline and outcome measures. The first cohort of patients provided information about the pain intensity on a Numerical Rating Scale (NRS). The second cohort provided data regarding functional limitations due to their main complaint. Patients with low back pain completed the Roland-Morris Disability Questionnaire (RDQ), patients with neck pain completed the Neck Disability Index (NDI), patients with upper extremity complaints completed the

Table 1; Overview of physician and patient related data collection

Type of data	Source of data	Outcome measures	
Physician characteristics	Survey	Demographics, training, treatment and referral patterns	
	Telephone call	Number of days per week spent in MSK practice	
Patient characteristics (web-based registry)	Sample	Patient questionnaires	Treating physician
	Cohort 1 (09/12-03/13)	Numerical Rating Scale	Demographics, source of referral, type and duration of complaints, and treatment
	Cohort 2 (04/13-01/14)	RDQ, NDI, LEFS, DASH, HIT-6*	
	Cohort 3 (02/14-02/16)	Previous treatments	

*RDQ (Roland Disability Questionnaire); 24 items, range 0-24, higher scores indicate more disability
NDI (Neck Disability Index); 10 items, range 0-50, higher scores indicate more disability
LEFS (Lower Extremity Function Scale); 20 items, range 0-80, higher scores indicate less disability
DASH (Disabilities of the Arm, Shoulder, and Hand); 30 items, range 0-100, higher scores indicate higher disability
HIT-6 (Headache Impact Test); 6 items, range 36-78, higher scores indicate more disability

Disabilities of the Arm, Shoulder and Hand (DASH) questionnaire, patients with lower extremity complaints completed the Lower Extremity Function Scale (LEFS), and patients with headache or migraine completed the Headache Impact Test (HIT-6). All instruments are commonly used in research and have been validated in Dutch populations(20-26). The third cohort provided data about previous treatments.

Data analyses

We analysed data using descriptive statistics in SPSS, version 22.

Results

Characteristics of physicians

Our survey was sent to all 138 members of the Dutch Association for Musculoskeletal Medicine, and returned by 90 physicians (65%). One physician did not tick the informed consent box and was removed from the analyses. Physician characteristics are presented in Table 2. After finishing medical training and before training in MSK medicine the majority of MSK physicians had worked in other medical specialties. Some had finished specialist training in other fields, most frequently in general practice (32.2%) or occupational medicine (16.7%). Of the two SMT techniques taught in the training program, a higher proportion of physicians had finished training in the manual medicine technique (63.3%) than the orthomanual medicine technique (58.9%). A number of MSK physicians were familiar

with other musculoskeletal treatment options, for example, McKenzie(27-30) or the use of protocols developed by the Spine Intervention Society (SIS)(30-33).

Table 2; Physician characteristics

Number of registered MSK physicians	138
Number of respondent	90 (65%)
Demographics	
Gender (male)	77.5 %
Age (range)	57 (38-75)
MSK consultations per week (range)	51 (5-150)
>3 days in MSK practice	57.3 %
Background training (%)	
Trained as General Practitioner	32.2
Still registered as General Practitioner	12.2
Trained in occupational medicine	16.7
Still registered in occupational medicine	8.9
MSK training (finished training, %)	
Orthomanual technique	58.9
Manual technique	63.3
McKenzie	13.3
Marsman	13.3
Spine Intervention Society	11.3

Table 3 presents an overview of treatments used by MSK physicians, as reported by the physicians in our survey. SMT techniques were used predominantly. Although a higher proportion of physicians had followed training in the manual technique, the orthomanual technique was used more frequently in daily practice (used often or regularly in 70.6% versus 56.2%). Regular use of McKenzie treatment was reported by 41.7% of respondents. Other commonly used supportive treatment options were training advice (e.g. advice on sports activities that could support the treatment) and postural advice (e.g. advice about how to perform ADL activities). Regular use of general medical injections (e.g. steroid injections for acute bursitis of the shoulder), prescription medication, and injection treatment according to SIS guidelines under X-ray guidance was reported by 34.8%, 37.1%, and 15.3% of respondents, respectively. Complementary treatment such as homeopathy or acupuncture was used regularly by less than 8% of the respondents.

Referral patterns, reported by the physicians in the survey, are presented in Table 4. Regular referral to physical therapy, exercise therapy, and postural therapy was reported by 46.1%,

62.2%, and 47.1% of the responding physicians, respectively. Physicians also reported further referral to other MSK physicians (referral from manual medicine to orthomanual medicine 20.5%, referral from orthomanual medicine to manual medicine 16.7%). Regular cooperation with medical specialists was mainly reported for orthopaedics, neurology and (anaesthetic) pain clinics (30.3%, 25.6%, and 28.7% respectively).

Table 3; Self-reported treatments used in daily practice by 89 Musculoskeletal physicians

Technique	Never/ Seldom (%)	Sometimes (%)	Regular/ Often (%)
Spinal Manipulative Treatment			
Orthomanual medicine technique	20.0	9.4	70.6
Manual medicine technique	20.2	23.6	56.2
McKenzie	29.8	28.6	41.7
Marsman	66.3	17.5	16.3
Supportive Treatments			
Training advice	9.0	15.7	75.3
Postural advice	4.4	12.2	83.3
Dietary advice	42.5	33.3	24.1
Prescribed medication	25.8	37.1	37.1
Injections			
Injections general medical	43.8	21.3	34.8
Injections SIS	82.4	2.4	15.3
Injections trigger point	70.1	16.1	13.8
Injections neural therapy	69.0	20.7	10.3
Complementary Treatments			
Homeopathy	84.1	11.4	4.5
Acupuncture	87.4	5.7	6.9
Dry needling	87.2	5.8	7.0
Podology	81.4	10.5	8.1

Table 4; Referral of MSK physicians (N=89) to other specialists and practitioners

Specialism	Never/ Seldom (%)	Sometimes (%)	Regular/ Often (%)
SMT			
Orthomanual medicine technique	47.0	32.5	20.5
Manual medicine technique	66.7	16.7	16.7
Chiropractor	96.4	2.4	1.2
Manual therapist	66.7	27.2	6.2
McKenzie	47.0	28.9	24.1
Marsman	84.1	13.4	2.4

Supportive treatment			
Physiotherapy	16.9	37.1	46.1
Exercise therapy	5.7	32.2	62.1
Postural therapy	12.6	40.2	47.1
Dietician	61.9	31.0	7.1
Medical specialists			
Neurologist	12.2	62.2	25.6
Orthopaedic surgeon	16.9	52.8	30.3
Rehabilitation	64.7	25.9	9.4
Pain clinic/ SIS	36.8	34.5	28.7
Complementary treatments			
Trigger point therapy	90.4	8.4	1.2
Neural therapy	91.6	7.2	1.2
Homeopathy	82.1	16.7	1.2
Acupuncture	79.5	18.1	2.4
Dry needling	83.3	15.5	1.2
Insoles	41.4	37.9	20.7

Characteristics of patients

A group of 31 MSK physicians volunteered to register patient data in our web-based registry, and to recruit patients. Demographic characteristics of the participating physicians (79% male, average age 54) were comparable to the demographic characteristics of both the whole population of MSK physicians (81% male, average age 57) and the part of the population that had answered to the physician survey (79% male, average age 56). Patient characteristics are presented in Table 5. The first cohort consisted of 1704 patients, of whom 1498 completed a baseline questionnaire (80%). The data registered by the treating MSK physician showed that forty-two percent of patients were male, and the predominant main complaint was low back pain without sciatica (30.0%), followed by low back pain with sciatica (18.1%) and neck pain (16.9%). Most patients (62.1%) had a main complaint that had been present for more than one year, only 16.3% had a main complaint that had lasted for less than three months. More than half of the patients (61.0%) sought care through self-referral, while 16% was referred by a general practitioner. The baseline questionnaire answered by the patients showed average NRS scores for the subgroups of patients with low back pain, neck pain and other complaints of 6.0, 6.2, and 6.0, respectively.

The second cohort consisted of 2610 patients, of whom 1701 patients answered to a baseline questionnaire (65%). Average baseline scores on the specific functional PROMs showed a moderate level of functional disability.

A sample of 433 patients was extracted from the third cohort, in which patients provided data about previous treatments. The majority of patients (82.1%) had been treated otherwise before consulting a MSK physician. Patients most frequently reported previous treatment by physical therapists (68.1%), followed by manual therapists (37.7%), medication (25.6%) and chiropractors (17.0%). Almost half (45.9%) of the patients had previously been treated by manual therapists or chiropractors, and 8.8% had been treated by both manual therapists and chiropractors.

Table 5; Patient characteristics

Cohort 1 data	**N**	**Main Complaint (ICPC code)**	**Percent**
Number of registrations	1704	Spine (L2, L3, and L86)	73.9
Number of respondents	1498	Low Back without sciatica (L3)	30.0
		Low Back with sciatica (L86)	18.1
		Neck (L1)	16.9
		Headache (N01, N02, and N89)	4.6
		Upper Extremity (L8-L12)	7.3
		Lower Extremity (L13-L17)	8.6
		Other	5.6
		Duration	
		< 3 months	16.3
		3-12 months	21.6
		> 1 year	62.1
		Source of referral	
		General practitioner	16.1
		Physiotherapist	8.7
		Medical Specialist	3.2
		Self-referral	61.0
		Other	11.0
		NRS pain*	**Mean (sd)**
		Low Back (N=722)	6.0 (2.0)
		Low Back without sciatica (N=449)	5.9 (1.9)
		Low Back with sciatica (N=273)	6.2 (2.0)
		Neck (N=250)	6.2 (2.0)
		Other (N=526)	6.0 (2.2)

Cohort 2 data		Function measures*	Mean (sd)
Number of registrations	2610	RDQ (N=827)	8.9 (5.3)
Number of respondents	1701	NDI (N=269)	13.1 (7.2)
		LEFS (N=159)	55.0 (15.8)
		DASH (N=102)	31.6 (16.5)
		HIT-6 (N=54)	60.0 (7.5)
Cohort 3 data		**Previous treatments**	**Percent**
Sample of respondents	433	Physical therapy	68.1
		Manual therapy	37.7
		Chiropractic treatment	17.0
		MT or chiropractor	45.9
		MT and chiropractor	8.8
		Medication	25.6
		Injections (pain clinic)	6.7
		Surgery	4.4

*Patient Reported Outcome Measures were tailored to the main complaint

Discussion

While the characteristics of chiropractors and manual therapists and their patients are well described, little is known about MSK physicians who use SMT. Our study is a first step to address this knowledge gap: we described MSK physicians in The Netherlands and their patients. Most MSK physicians in The Netherlands had previous experience in other medical specialties. They were trained in a variety of SMT techniques that, in part, differed from the techniques used by chiropractors and manual therapists. Furthermore, they used an array of other diagnostic and treatment options, part of which were, by law, restricted to medical doctors, such as prescription medication and general medical injections or injections under X-ray guidance. Physicians reported frequent use of training advise or postural advise and further referral for exercise therapy or physiotherapy.

The majority of patients consulting MSK physicians reported spinal pain of long duration, with moderate functional disability. This is comparable to the patient population reported in a previous study to consult chiropractors in The Netherlands(17). Patients consulting manual therapists(16) reported, on average, musculoskeletal pain of shorter duration (59% < 3 months, 21% > 1 year) than the patients seen by chiropractors (24% < 3 months, 58% > 1 year) and MSK physicians (16% < 3 months, 62% > 1 year). Patients consulting MSK physicians had frequently been treated previously by other SMT professionals. This could be reflective of the practice in The Netherlands where, traditionally, the general practitioner

refers patients with musculoskeletal complaints to physical therapists. In The Netherlands, manual therapy is a subspecialty of physical therapy, and thus manual therapists are likely to be consulted by patients with less severe complaints at an earlier stage. Only when complaints are refractory to treatment do patients consult chiropractors or MSK physicians. This practice is supported by health care insurance policies, which generally cover a number of physiotherapy treatments; the costs of chiropractic care and MSK medicine are only reimbursed for patients with additional coverage.

It must be noted that our study described the situation in The Netherlands. Due to differences in health care organisation, recognition of the various professional groups and reimbursement of the costs of treatment, respective patient populations may vary between countries. In Denmark, for example, chiropractic treatment is embedded in regular primary care, with strong academic connections(34), while in The Netherlands and Belgium chiropractic treatment is considered to be complementary medicine(13, 17). Furthermore, in other countries, the various professional groups might have different licensing requirements for prescribing medication or applying injections. Comparable variations exist in the position of MSK physicians. MSK medicine is practised in other European countries as an additional competence to other medical specialities, while in The Netherlands it is put forward as a medical profession in its own right.

Strengths and limitations

The main strengths of this study are that the whole population of physicians registered in MSK medicine was approached for our study, and the large number of patients who provided data. Nearly all physicians using SMT in The Netherlands are members of the Dutch Association for Musculoskeletal Medicine, because registration in the Register for Musculoskeletal Medicine is necessary to have the costs of treatment reimbursed, and this registration can only be obtained after completing the professional training program. A limitation of our study could be that only 65% of the members returned our survey. However, demographic characteristics (age and sex) of the responding physicians were comparable to non-responders. Another limitation could be that data on physician characteristics was self-reported. Lastly, we obtained patient data from a subset of MSK practices, as not all MSK practices were willing to collect patient data. However, we consider the data to be representative as the demographic characteristics of the participating physicians were comparable to the demographic characteristics of all members of the Dutch Association for Musculoskeletal Medicine.

Further study

Additional studies describing physicians who are trained to use SMT in other countries are needed. There are differences in the type of SMT technique used by various professionals(18). Future studies should clearly report the SMT techniques in detail. The CIRCLe SMT study presented criteria for reporting SMT techniques(35). Lastly, studies in which various SMT techniques are embedded within different treatment protocols are warranted.

Conclusion

MSK physicians in The Netherlands reported to use an array of SMT techniques. They had embedded their SMT techniques in a broad array of other diagnostic and treatment options, part of which were limited to medical doctors. Most patients consult MSK physicians with spinal pain of long duration with moderate functional impairment.

Abbreviations

SMT, Spinal manipulative treatment; MSK, Musculoskeletal; NRS, Numerical rating scale; ICPC, International classification of primary care; RDQ, Roland Morris Disability Questionnaire; NDI, Neck Disability Index; LEFS, Lower Extremity Function Scale; DASH, Disabilities of the Arm, Shoulder and Hand; HIT-6, Headache Impact Test; SIS, Spine intervention society; CIRCLe SMT, Consensus on interventions reporting criteria list spinal manipulative therapy.

Declarations

Ethics approval and consent to participate

The Medical Ethical Committee of the VU medical center decided that this observational study did not require the strict procedure for written and signed informed consent based on the law for Scientific Medical Research (WMO). Nonetheless, we used a form of informed consent. Verbal informed consent was obtained from all patients in this study, which was recorded by the treating physician. Written informed consent was obtained from all physicians in this study. This study protocol was approved by the Medical Ethical Committee (METc) of the VU medical center (no 2013/133).

Acknowledgements

We would like to thank all members of the Dutch Association for Musculoskeletal Medicine who cooperated in this study. We also would like to thank K. Uegaki for reviewing the manuscript.

References

1. Posadzki P, Ernst E. Spinal manipulation: an update of a systematic review of systematic reviews. N Z Med J. 2011;124(1340):55-71.
2. Gross AR, Hoving JL, Haines TA, Goldsmith CH, Kay T, Aker P, et al. A Cochrane review of manipulation and mobilization for mechanical neck disorders. Spine (Phila Pa 1976). 2004;29(14):1541-8.
3. Rubinstein SM, Terwee CB, Assendelft WJ, de Boer MR, van Tulder MW. Spinal manipulative therapy for acute low back pain: an update of the cochrane review. Spine (Phila Pa 1976). 2013;38(3):E158-E77.
4. Rubinstein SM, van Middelkoop M, Assendelft WJ, de Boer MR, van Tulder MW. Spinal manipulative therapy for chronic low-back pain: an update of a Cochrane review. Spine (Phila Pa 1976). 2011;36(13):E825-E46.
5. Deyo RA. The Role of Spinal Manipulation in the Treatment of Low Back Pain. JAMA. 2017;317(14):1418-9.
6. Assendelft WJ, Pfeifle CE, Bouter LM. Chiropractic in The Netherlands: a survey of Dutch chiropractors. J Manipulative Physiol Ther. 1995;18(3):129-34.
7. Blum C, Globe G, Terre L, Mirtz TA, Greene L, Globe D. Multinational survey of chiropractic patients: reasons for seeking care. J Can Chiropr Assoc. 2008;52(3):175-84.
8. Dunn AS, Passmore SR. Consultation request patterns, patient characteristics, and utilization of services within a Veterans Affairs medical center chiropractic clinic. Mil Med. 2008;173(6):599-603.
9. French SD, Charity MJ, Forsdike K, Gunn JM, Polus BI, Walker BF, et al. Chiropractic Observation and Analysis Study (COAST): providing an understanding of current chiropractic practice. Med J Aust. 2013;199(10):687-91.
10. Hurwitz EL, Coulter ID, Adams AH, Genovese BJ, Shekelle PG. Use of chiropractic services from 1985 through 1991 in the United States and Canada. Am J Public Health. 1998;88(5):771-6.
11. Leboeuf-Yde C, Hennius B, Rudberg E, Leufvenmark P, Thunman M. Chiropractic in Sweden: a short description of patients and treatment. J Manipulative Physiol Ther. 1997;20(8):507-10.
12. Mootz RD, Cherkin DC, Odegard CE, Eisenberg DM, Barassi JP, Deyo RA. Characteristics of chiropractic practitioners, patients, and encounters in Massachusetts and Arizona. J Manipulative Physiol Ther. 2005;28(9):645-53.
13. Ailliet L, Rubinstein SM, de Vet HC. Characteristics of chiropractors and their patients in Belgium. J Manipulative Physiol Ther. 2010;33(8):618-25.
14. Coulter ID, Shekelle PG. Chiropractic in North America: a descriptive analysis. J Manipulative Physiol Ther. 2005;28(2):83-9.
15. Malmqvist S, Leboeuf-Yde C. Chiropractors in Finland--a demographic survey. Chiropr Osteopat. 2008;16:9.

16. Oostendorp RA. Manual Physical Therapy in the Netherlands: Reflecting on the Past and Planning for the Future in an International Perspective. The Journal of Manual & Manipualtive Therapy. 2007 2007:133-41.
17. Rubinstein S, Pfeifle CE, van Tulder MW, Assendelft WJ. Chiropractic patients in the Netherlands: a descriptive study. J Manipulative Physiol Ther. 2000;23(8):557-63.
18. Veen van de EA, de Vet HC, Pool JJ, Schuller W, de Zoete A, Bouter LM. Variance in manual treatment of nonspecific low back pain between orthomanual physicians, manual therapists, and chiropractors. J Manipulative Physiol Ther. 2005;28(2):108-16.
19. Lamberts H, Wood, M. (Eds.). International Classification of Primary Care (ICPC). Oxford: Oxford University Press; 1987.
20. Beurskens AJ, de Vet HC, Koke AJ. Responsiveness of functional status in low back pain: a comparison of different instruments. Pain. 1996;65(1):71-6.
21. Hoogeboom TJ, de Bie RA, den Broeder AA, van den Ende CH. The Dutch Lower Extremity Functional Scale was highly reliable, valid and responsive in individuals with hip/knee osteoarthritis: a validation study. BMC Musculoskelet Disord. 2012;13:117.
22. Jorritsma W, de Vries GE, Geertzen JH, Dijkstra PU, Reneman MF. Neck Pain and Disability Scale and the Neck Disability Index: reproducibility of the Dutch Language Versions. Eur Spine J. 2010;19(10):1695-701.
23. Jorritsma W, Dijkstra PU, de Vries GE, Geertzen JH, Reneman MF. Detecting relevant changes and responsiveness of Neck Pain and Disability Scale and Neck Disability Index. Eur Spine J. 2012;21(12):2550-7.
24. Martin M, Blaisdell B, Kwong JW, Bjorner JB. The Short-Form Headache Impact Test (HIT-6) was psychometrically equivalent in nine languages. J Clin Epidemiol. 2004;57(12):1271-8.
25. Veehof MM, Sleegers EJ, van Veldhoven NH, Schuurman AH, van Meeteren NL. Psychometric qualities of the Dutch language version of the Disabilities of the Arm, Shoulder, and Hand questionnaire (DASH-DLV). J Hand Ther. 2002;15(4):347-54.
26. Beurskens AJ, de Vet HC, Koke AJ, van der Heijden GJ, Knipschild PG. Measuring the functional status of patients with low back pain. Assessment of the quality of four disease-specific questionnaires. Spine (Phila Pa 1976). 1995;20(9):1017-28.
27. Berthelot JM, Delecrin J, Maugars Y, Passuti N. Contribution of centralization phenomenon to the diagnosis, prognosis, and treatment of diskogenic low back pain. Joint Bone Spine. 2007;74(4): 319-23.
28. Dunsford A, Kumar S, Clarke S. Integrating evidence into practice: use of McKenzie-based treatment for mechanical low back pain. J Multidiscip Healthc. 2011;4:393-402.
29. Machado LA, de SM, Ferreira PH, Ferreira ML. The McKenzie method for low back pain: a systematic review of the literature with a meta-analysis approach. Spine (Phila Pa 1976). 2006;31(9):E254-E62.

30. van Helvoirt H, Apeldoorn AT, Ostelo RW, Knol DL, Arts MP, Kamper SJ, et al. Transforaminal epidural steroid injections followed by mechanical diagnosis and therapy to prevent surgery for lumbar disc herniation. Pain Med. 2014;15(7):1100-8.
31. Baker RM. International Spine Intervention Society (ISIS) presidential address: 20th Annual Scientific Meeting. Wednesday, July 18, 2012. Pain Med. 2012;13(9):1108-9.
32. Bogduk N. International Spinal Injection Society guidelines for the performance of spinal injection procedures. Part 1: Zygapophysial joint blocks. Clin J Pain. 1997;13(4):285-302.
33. Engel A, King W, MacVicar J. The effectiveness and risks of fluoroscopically guided cervical transforaminal injections of steroids: a systematic review with comprehensive analysis of the published data. Pain Med. 2014;15(3):386-402.
34. Hartvigsen J, Sorensen LP, Graesborg K, Grunnet-Nilsson N. Chiropractic patients in Denmark: a short description of basic characteristics. J Manipulative Physiol Ther. 2002;25(3):162-7.
35. Groeneweg R, Rubinstein SM, Oostendorp RA, Ostelo RW, van Tulder MW. Guideline for Reporting Interventions on Spinal Manipulative Therapy: Consensus on Interventions Reporting Criteria List for Spinal Manipulative Therapy (CIRCLe SMT). J Manipulative Physiol Ther. 2017;40(2):61-70.

Chapter 3.

Pain trajectories and predictors of a favourable course of low back pain in patients consulting musculoskeletal physicians in The Netherlands

W. Schuller, R.W. Ostelo, D.C. Rohrich, M.W. Heymans, H.C.W. de Vet

Submitted

Abstract

Background: There are no studies yet reporting on the course of Low Back Pain (LBP) after treatment by musculoskeletal (MSK) physicians.

Methods: In an observational cohort study MSK physicians recorded various baseline and treatment variables. Patient questionnaires included information about previous medical consumption, together with PROMs measuring the level of pain and functional status at baseline, and at 6-weekly intervals during a follow-up period of six months. Latent Class Growth Analysis (LCGA) was used to classify patients into different groups according to their pain trajectories. Baseline variables were evaluated as predictors of a favourable trajectory using logistic regression analyses.

Results: 1377 Patients were recruited, of whom 1117 patients (81%) answered at least one follow-up measurement. LCGA identified three groups of patients with distinct pain trajectories. A first group (N=226) with high pain levels showed no improvement, a second group (N=578) with high pain levels showed strong improvement, and a third group (N=313) with mild pain levels showed moderate improvement. The two groups of patients presenting with high baseline pain scores were compared, and a prediction model of a favourable course was constructed. Male gender, previous specialist visit, previous pain clinic visit, having work, a shorter duration of the current episode, and a longer time since the complaints first started were predictors of a favourable course. The prediction model showed a moderate area under the curve (0.68) and a low explained variance (0.09).

Conclusions: Three different pain trajectories were identified. Baseline variables were of limited value in predicting a favourable course.

Introduction

Low back pain (LBP) is a major health problem, with a point prevalence in Western countries of 12-30%(1). In 2010, in The Netherlands, total costs were estimated to be €3.5 billion, including both direct costs, related to the consumption of medical care, and indirect costs, related to loss of productivity and disability pensions(1). Because in most cases the mechanism of LBP is not known, there is no intervention that can be directed at the cause of the pain, and while many interventions are available, none has shown to be superior(2). Although the course of low back pain has long been considered favourable, recurrences are common(3, 4), and many patients (65%) still reported pain 1 year after onset(5). Considering the recurrent character of LBP, recent research has increasingly focused on identifying LBP trajectories(6, 7). Distinct clusters of pain trajectories were identified(8), and over the course of their LBP, patients showed consistent cluster membership(6). Rather than studying prognostic factors based upon outcome measurements at one follow-up moment, it may therefore be more informative to follow patients for longer periods of time, and to identify prognostic factors that predict the trajectory of LBP. This knowledge can potentially be used in outcome research, studying whether interventions can influence patients pain trajectories, rather than offering momentary improvement. Measuring pain trajectories has become easier with the development of automated systems distributing patient reported questionnaires over internet, or using text messages. A recent study by Ailliet et al., for example, used text messages to study the pain trajectories of patients with low back and neck pain in patients consulting chiropractors in The Netherlands and Belgium(9).

In The Netherlands, among the various diagnostic and treatment possibilities, patients can consult physicians specialised in Musculoskeletal Medicine (MSK). MSK physicians use an array of diagnostic and treatment options, almost invariably including a form of Spinal Manipulative Treatment (SMT). About half of the patients consult MSK physicians because of LBP(10). The aim of our study was to assess whether different pain trajectories could be identified in LBP patients after consulting a MSK physicians, and to identify possible predictors of a favourable course.

Methods

Study design and participants

We conducted a prospective cohort study with a follow-up period of six months.

All MSK physicians registered with the Dutch Association for Musculoskeletal Medicine were invited to participate in our study. Participating physicians were instructed to register

all patients who presented for the first time in an MSK practice through a web-based registry, and to invite these patients to take part in the study. Inclusion criteria were LBP, age≥18, and sufficient mastery of the Dutch language to answer questionnaires in Dutch. If patients gave informed consent, the treating physician entered email addresses of the recruited patients in the web-based registry. Thereafter, a specially designed computer program (Readmail) was used to automatically distribute invitations to patients by email to fill in web-based questionnaires. Our study procedures were approved by the Medical Ethical Committee (METc) of the VU Medical Center (2012/414).

Study procedure

Both the treating physicians and the individual patients provided data via web-based registries. The treating physicians recorded data at baseline and at the end of the treatment. Study procedures were explained to participating physicians during specially organised information sessions. In addition, a research assistant visited all participating practices to explain the procedures. Practices that agreed to participate at a later stage were informed by telephone. Instructions were to ask all consecutive patients presented for a first consultation to participate in the study. Recruited patients received invitations to fill in web-based questionnaires within three weeks before the first consultation, and at six weekly intervals during the ensuing six months. When patients did not respond, a maximum of three reminders were sent within a period of two weeks. Both the invitation email and the web-based questionnaires contained links to a leaflet with information about the study.

Measurement

At baseline physicians registered data about age, gender, type and duration of the main complaint, and the existence of concomitant complaints. Complaints were registered according to the International Classification of Primary Care (ICPC)(11). At the end of treatment, data were registered about the number of treatment sessions, the type of treatment used, and further referrals.

At baseline, patients were asked to indicate whether their main complaint was low back pain, neck pain or any other complaint. This question was supported by text and manikins, explaining which area was considered to cover neck pain or low back pain. For other complaints, patients could explain these in text. Patients were asked to indicate whether their pain radiated to the legs or arms, and whether they had numbness or pins and needles in their legs or arms. Patients were also asked about the duration of the current episode, the time since the first episode, educational level, work status, previous specialist consultations, and previous treatments. The effect of previous treatments was measured on an ordinal scale, with four possible answer categories; 1.strong improvement, 2.little improvement,

3.unchanged, 4.deteriorated. Furthermore, all patients were asked to answer a set of Patient Reported Outcome Measures (PROMs), including a Numerical Rating Scale (NRS) for pain severity, the SF6D(12), and the Fear Avoidance Beliefs Questionnaire (FABQ)(13, 14). Patients who indicated LBP as their main complaint were asked to answer the Oswestry Disability Index (ODI)(15, 16). All PROMs have been validated in patients with low back pain, and are frequently used in research. The SF6D is a short version of the SF36, measuring health related quality of life. Scores range from 0-1, with higher scores indicating lower quality of life. The FABQ consists of 16 items, and measures pain related fear in LBP patients. Higher scores indicate more pain related fear. The ODI consists of 10 items with scores ranging from 0-50, with higher scores indicating more disability because of LBP. At all follow-up points patients were asked to answer the same PROMs, except for the FABQ. A question about the Global Perceived Effect (GPE) of treatment was added.

Statistical analyses

Identification of pain trajectories

Our study population consisted of all LBP patients who completed the baseline questionnaire. For the analyses of pain trajectories, patients were selected who completed the baseline questionnaire and at least one of the follow-up questionnaires. Latent Class Growth Analyses (LCGA) were used to explore whether subgroups of patients following distinct pain trajectories could be identified, using the NRS for pain scores(17). Several LCGA models were evaluated with different numbers of trajectories, allowing linear or quadratic pain trajectories, and allowing more or less heterogeneity in pain trajectories within subgroups. A final model was selected based upon model fit and considerations of interpretability and clinical practicality. Model fit was evaluated with the Vuong-Lo-Mendell-Rubin likelihood ratio test (LMR-LRT) and the Bayesian Information Criterion (BIC)(18). The BIC considers both the likelihood of the model as well as the number of parameters in the model, with lower values showing better model fit. The LMR-LRT provides a p-value. A significant p-value indicates that a model with k classes fits better than a model with k-1 classes. LCGA was carried out using Mplus (Version 7)(18, 19).

Predicting a favourable outcome

Descriptive analyses of baseline variables were carried out for the complete population included in the analyses, and for each group of patients with a distinct pain trajectory separately. For the patients that presented high baseline pain scores two distinct pain trajectories were identified (see results). One trajectory identified a group of patients who did not improve (class 1), and one trajectory identified a group of patients who improved (class 2). Logistic regression analyses were conducted to study the univariate relationship

between baseline variables and the dependent variable (i.e. high baseline pain and not improved (class 1) vs high baseline pain and improved (class 2)).

A backward selection procedure was carried out on cases with complete data on all variables to construct a prediction model, based upon a p-value of 0.157 (Akaike criterion). Treatment variables were considered as possible confounders instead of predictors. Although not known at baseline, they could possibly influence the outcome. The relationship between continuous predictors and the outcome was tested for linearity, and non-linear variables were entered as splines. The fit of the final model was evaluated with the loglikelihood and the Hosmer Lemeshow test(20). Discriminative properties of the model were evaluated by calculating the Area Under the Curve (AUC), and explained variance was evaluated with Nagelkerke R^2 (21). Bootstrapping was used for internal validation(22). Descriptive analyses and univariate analyses were carried out using SPSS 22, except for the univariate analyses of non-linear variables. Linearity, univariate analyses of non-linear variables and internal bootstrap validation was carried out with the R-package rms (version 5.1-2). In the multivariable analyses, backwards selection was carried out with the R-package pfsmi(23).

Missing data and evaluation of loss to follow-up

The relationship between complete predictors and predictors with missing values (>20%) was studied with univariate logistic regression analyses. In this analysis, significant relationships between predictors and the variable being either missing or not missing support the assumption that missing values are probably missing at random (MAR). When including all potential predictors in the model, the percentage of missing cases was around 40%, which required 40 multiple imputed datasets. Multivariate analyses were conducted in each dataset, and results were pooled according to Rubin's rules(24). To evaluate the influence of loss to follow-up, the group of patients with complete baseline questionnaires was compared with the group of patients answering at least one follow-up measurement. Differences on predictor variables between these two groups were studied with logistic regression. Multiple imputation and evaluation of missing data were carried out using SPSS 22.

Results

Study population

Data was collected from February 2014 until February 2016. A discrepancy was found between the number of patients classified with low back pain by the physician, and the self-classification by the patient through the web-based questionnaire. Frequently, patients classified themselves as other, but indicated complaints in text that would classify as low back pain. It was therefore decided to use the classification of the physician to select LBP patients. In the web-based registry MSK physicians recorded 2026 patients with LBP. Of these patients 1664 were recruited for our study. Our study population consisted of 1377 patients who answered the baseline questionnaire. A total of 1117 patients (81%) answered at least one of the follow up measures next to the baseline questionnaire and were included in the LCGA and prediction analyses. Although 19 practices participated in the study, the LBP patients were recruited by 16 practices, and the number of LBP patients recruited by the various practices varied from 1-285.

Missing data loss to follow-up

Table 1 shows the handling of predictor variables, including the percentages of missing values. Only one variable, Baseline ODI, showed a high percentage of missing values (25.6%), which could be explained by the tailored distribution of the ODI to patients who had self-classified as LBP patient. Because not all LBP patients classified themselves as such, not all LBP patients received the ODI. Because the percentage of missing Baseline ODI values was >20, it was decided to impute the baseline ODI.

Evaluation of loss to follow-up (Table 2) showed that baseline scores on the ODI, SF6D, and NRS did not differ significantly between patients who only answered the baseline questionnaire and the patients included in the analyses. Female patients (OR 1.33), older patients (OR 1.01), and patients treated effectively by a chiropractor (OR 2.14) were significantly more inclined to remain in our study. Patients treated effectively at a pain clinic were significantly less inclined to remain in our study (OR 0.19).

Table 1; Handling of predictor variables.

Predictor	Type	Missing %	Handling	Category
Gender	dichotomous	0	unchanged	Male/female
Age	continuous	0	unchanged	
Education	categorical	1.8	dichotomised	higher/lower education
Radiating pain into the leg	dichotomous	0	unchanged	Yes/ no
Radiating pins and needles	dichotomous	0	unchanged	Yes/ no
Time since start 1st complaints	continuous	0.6	unchanged	
Duration current pisode	continuous	3.9	unchanged	
Baseline SF-6D	continuous	0.4	unchanged	
Baseline ODI	continuous	25.6	imputed	
Previous specialist visit	dichotomous	0	unchanged	Yes/no
Previous visit neurologist	dichotomous	0	unchanged	Yes/no
Previous visit orthopedic	dichotomous	0	unchanged	Yes/no
Previous visit rehabilitaion	dichotomous	0	unchanged	Yes/no
Previous visit pain clinic	dichotomous	0	unchanged	Yes/no
Medication	categorical	0	unchanged	Four categories; none, rarely, regularly not daily, daily
Concomitant complaints	dichotomous	0	unchanged	Yes/no
Previous physiotherapy	dichotomous	0	categorised	Combined into one categorical value, treated without effect (reference), treated with effect, or not treated
Effect physiotherapy	ordinal			
Previous manual therapy	dichotomous	0	categorised	Combined into one categorical value, treated without effect (reference), treated with effect, or not treated
Effect manual therapy	ordinal			
Previous chiropractic treatment	dichotomous	0	categorised	Combined into one categorical value, treated without effect (reference), treated with effect, or not treated
Effect chiropractic treatment	ordinal			

Predictor	Type	Missing %	Handling	Category
Previous medication	dichotomous	0	categorised	Combined into one categorical value, treated without effect (reference), treated with effect, or not treated
Effect medication treatment	ordinal			
Previous pain clinic treatment	dichotomous	0	categorised	Combined into one categorical value, treated without effect (reference), treated with effect, or not treated
Effect pain clinic	ordinal			
Previous surgical treatment	dichotomous	0	categorised	Combined into one categorical value, treated without effect (reference), treated with effect, or not treated
Effect surgical treatment	ordinal			
Previous treatment other	dichotomous	0	categorised	Combined into one categorical value, treated without effect (reference), treated with effect, or not treated
Effect other treatment	ordinal			
Work status	various dich. variables	0	categorised	Combined into one categorical variable, no work (reference),not physical work and physical work
Type of work				
Pain avoidant	continuous	0	dichotomised	Pain avoidant yes/ no (no; FABQ <14)
Type of treatment	various dich. variables	7.9	categorised	Combined into one categorical value, MG treatment (reference), OMG treatment, both, or none
Number of treatment sessions	continuous	7.9	unchanged	
McKenzie therapy	dichotomous	7.9	unchanged	Yes/no
Treated differently	various dich. variables	7.9	dichotomised	Combined into one dichotomous variable, treated differently yes/ no

Table 2; Comparison of responders with non-responders

Predictor variabele	p-value	OR
Gender (female)	*0.042*	1.325
Age	*0.008*	1.014
Education (high)	0.261	1.171
Radiating pain	0.453	0.900
Radiating pins and needles	0.848	1.030
Time since start 1st complaints	0.792	1.002
Duration current episode	0.423	0.989
Concomitant complaint neck pain	0.881	0.976
Concomitant complaints other	0.757	1.044
Baseline ODI	0.265	0.994
Baseline SF-6D	0.691	0.764
Previous specialist visit	0.078	1.280
Previous visit neurologist	0.287	0.843
Previous specialist visit orthopedic	0.413	0.874
Previous specialist visit rehabilitation	0.317	0.745
Previous specialist visit pain clinic	0.522	1.192
Medication	0.477	–
Previous physiotherapy	0.987	1.022*
Previous manual therapy	0.899	1.152*
Previous chiropractic treatment	*0.034*	2.141*
Previous medication treatment	0.735	0.727*
Previous pain clinic treatment	*0.046*	0.192*
Previous surgical treatment	0.333	3.289*
Work status	0.060	0.753*
Pain avoidant	0.490	0.895

p-value and OR of responders versus non-responders
*Of categorical variables only the OR of category 1 versus category 0 is presented. In the treatment variables, category 1 represents patients who had been treated effectively, category 0 represents patients who had been treated not effectively. For the variable work, category 1 represents patients with non-physical work, versus category 0, patients without work.

Defining subgroups with distinct pain trajectories

Model fit characteristics are presented in Table 3. Although a four-class quadratic model without fixed variance showed slightly better fit (BIC 17836 versus 17842), it was decided to choose the three-class quadratic model without fixed variance (Figure 1) as the preferred model, because of its better interpretability and practicality. The four-class model included a small group of patients (7.0%) who showed a strong improvement in the first three months,

Table 3; LCGA models; Several models were evaluated with 1-5 classes, quatratic or linear, and with subgroup variance fixed to 0 or free. Model fit was evaluated with the Vuong-Lo-Mendell-Rubin likelihood ratio test (LMR-LRT) and the Bayesian Information Criterion (BIC). The Akaike information criterion (AIC) is presented as well. The choice of the best model was based upon model fit, clinical interpretability and practicality.

Model	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Classes	2	3	4	5	2	3	4	5	2	3	4	5	2	3	4	5
Quadratic	Yes	Yes	Yes	Yes	No	No	No	No	No	No	No	No	Yes	Yes	Yes	Yes
Variance	0	0	0	0	0	0	0	0	Free	Free	Free	Free	Free	Free	Free	Free
Likelihood	-8948	-8888	-8850	-8830	-9156	-9125	-9097	-9084	-9106	-9091	-9077	-9067	-8885	-8844	-8826	-8806
AIC	17923	17810	17743	17711	18333	18276	18226	18207	18239	18214	18193	18178	17807	17732	17705	17672
BIC	17988	17896	17849	17836	18383	18341	18306	18302	18304	18295	18288	18228	17898	17842	17836	17822
p-value*	0.000	0.002	0.000	0.000	0.000	0.147	0.000	0.056	0.000	0.065	0.173	0.018	0.000	0.000	0.029	0.074
Inter-pretability	+	+++	+	-	+	+++	+	-	+	+++	+	-	+	+++	+	-

* p-value based upon the LMR-LR

and a return to previous pain levels in the subsequent months. Compared to the three-class model, this for a large part only changed the proportion of patients in the group that started with high pain levels and showed no improvement, suggesting that this group consisted of patients who remained unchanged during the study combined with patients who improved strongly, only to deteriorate again. The course of the average NRS scores of the three groups is presented in Table 4, together with the mean changes in the ODI scores. In the three-class model, a group of 226 patients started with high NRS scores at baseline and showed hardly any change during the follow-up period (mean NRS-pain changed from 6.9 to 6.7). In this group the mean ODI score changed from 24.8 to 19.4. A group of 578 patients started with high baseline scores and showed considerable improvement (mean NRS changed from 7.0 to 1.8). In this group , the mean ODI score changed from 26.4 to 6.1. A group of 313 patients started with lower baseline scores and showed moderate, but clinically relevant improvement (mean NRS changed from 3.5 to 2.2). In this group the mean ODI score changed from 15.5 to 8.0. Demographic data of the study population and of the three groups are presented in Table 5.

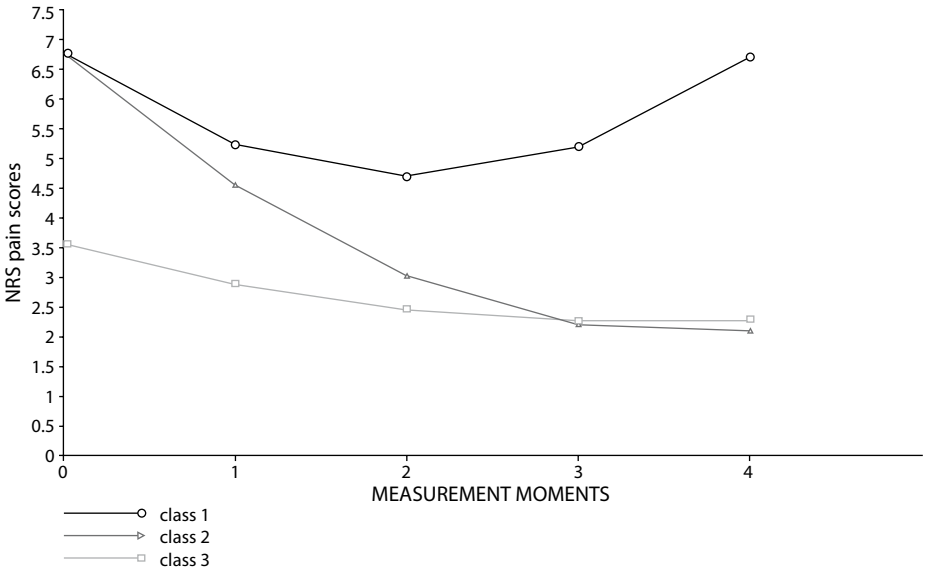

Figure 1; Three-class model of low back pain trajectories identified with LCGA. The Y-axis represents the NRS mean group scores, the X-axis represents the measurement moments (1=baseline, 2=6 weeks, 3=12 weeks, 4=18 weeks, and 5=24 weeks follow-up).

Table 4; Mean NRS and ODI scores of the three classes.

	Class 1 (N=226, 20%)		Class 2 (N=578, 52%)		Class 3 (N=313, 28%)	
Mean NDI and ODI	**NRS (SD)**	**ODI (SD)**	**NRS (SD)**	**ODI (SD)**	**NDI (SD)**	**ODI (SD)**
Baseline	6.9 (1.2)	24.8 (15.8)	7.0 (1.1)	26.4 (15.4)	3.5 (1.1)	15.5 (12.7)
6 weeks	5.2 (2.2)	17.6 (14.5)	4.0 (2.4)	15.0 (14.4)	2.5 (1.9)	8.6 (9.7)
12 weeks	5.7 (2.4)	19.7 (15.6)	2.9 (2.3)	11.0 (11.4)	2.6 (2.2)	8.7 (10.8)
18 weeks	5.7 (2.1)	18.9 (16.3)	2.6 (2.2)	8.1 (10.2)	2.5 (2.2)	8.6 (10.4)
26 weeks	6.7 (1.2)	19.4 (14.7)	1.8 (1.4)	6.1 (7.3)	2.2 (1.7)	8.0 (9.9)
Baseline-26 wk. change	0.2 (3%)	5.4 (22%)	5.2 (74%)	20.3 (77%)	1.3 (37%)	7.5 (48%)

Mean NRS and ODI scores (SD) of all three LCGA classes at baseline and at all follow-up moments, and the score change between baseline and 26 weeks follow-up.

Predictors

In Table 5 in the Supplement the predictor values are described for the whole study population, and separately for the three groups of patients with distinct pain trajectories identified with LCGA. Because the group of patients with low baseline pain scores could be identified by the baseline NRS scores, our main interest was to identify predictors that distinguish patients with high NRS scores who showed a favourable course from patients with high NRS scores who did not show a favourable outcome. Further analyses therefore focused on the two subgroups that started with high pain scores, i.e. the group of patients that was considered to be improved and the group of patients that was considered to be not improved. Baseline variables were evaluated as possible predictors of a favourable course. For all baseline variables, univariate odds ratio's for improvement are presented. The relationship between the continuous predictors and group membership was shown to be linear for all continuous variables except for the duration of the current episode, which was further analyzed as a spline variable. In the univariate analyses the odds of a favourable course decreased in the first four years of the current episode, but increased in the years thereafter. Male gender, previous specialist visit, previous surgical treatment, and having work were associated with a favourable course. Previous consultation with a neurologist or an orthopedic surgeon, no effect of previous treatments and concomitant complaints were associated with a non-favourable outcome. Other predictors did not show a significant association with the outcome.

Table 5; Univariate analyses of predictor variables.

	Whole Study Population	LCGA classes of 1117 patients classified			OR for a favourable outcome			
		Class 1	Class 2	Class 3	Univariate analyses Class 2 versus Class 1			
Predictor variables		Not impr.	Improved	Low NRS	p-value	OR	95% CI of OR	
N	1377	226	578	313			Lower	Upper
General baseline features								
Gender (female)	41.3	30.1	42.2	43.1	0.002	0.589	0.424	0.818
Age, mean (SD)	47.0 (13.5)	47.6 (14.4)	47.4 (13.8)	47.5 (12.2)	0.845	0.999	0.988	1.010
Education (high versus low)	60.4	60.1	57.8	67.9	0.448	0.883	0.641	1.217
Radiating pain into the leg	39.9	43.8	42.9	29.7	0.817	0.964	0.707	1.315
Radiating pins and needles	27.0	31.0	28.0	22.7	0.407	0.932	0.788	1.102
Time since complaints 1st started	11.4 (12.0)	11.6 (12.3)	11.4 (12.2)	11.4 (11.6)	0.855	0.999	0.986	1.011
Duration of current episode	2.1 (4.8)	3.2 (6.2)	1.9 (4.8)	1.7 (3.6)	0.003	0.958	0.932	0.986
SF-6D baseline score	0.74 (0.11)	0.73 (0.11)	0.73 (0.10)	0.78 (0.10)	0.523	0.613	0.137	2.746
ODI baseline score	23.4 (15.7)	24.8 (15.8)	26.4 (15.4)	15.5 (12.7)	0.525	1.004	0.992	1.015
Concomitant complaints present	53.8	58.8	51.4	54.3	0.048	0.731	0.535	0.998
Pain avoidant	70.9	73.5	73.6	62.1	0.938	0.986	0.697	1.397
Previous medical consumption								
Previous specialist visit	62.1	52.2	64.2	69.3	0.002	1.640	1.201	2.240
Previous visit neurologist	22.5	28.3	20.9	19.2	0.026	0.819	0.687	0.976
Previous visit orthopedic surgeon	21.2	28.8	19.7	16.9	0.006	0.847	0.753	0.953
Previous visit rehabilitation	4.9	7.1	5.0	2.2	0.255	0.912	0.779	1.068
Previous visit pain clinic	7.5	8.8	8.5	5.4	0.866	0.991	0.888	1.105
Medication (categorical)								
-None (ref.)	32.8	26.1	28.9	48.2	0.231			
-Rarely	24.5	23.9	24.6	24.3	0.738	0.929	0.603	1.430

-Regularly not daily	28.2	35.4	28.4	20.8	0.113	0.724	0.486	1.080
-Daily	14.5	14.6	18.2	6.7	0.640	1.124	0.688	1.837
Effect of previous treatments								
Previous treatment physiotherapy								
-Treated no effect (ref.)	64.6	72.6	63.5	60.4	0.037			
-Treated effect	6.6	3.1	5.9	10.5	0.069	2.170	0.943	4.998
-Not treated	28.8	24.3	30.6	29.1	0.044	1.438	1.009	2.049
Previous treatment manual therapy								
-Treated no effect (ref.)	28.8	35.4	27.3	26.5	0.066			
-Treated effect	6.4	4.9	6.7	7.3	0.112	1.795	0.873	3.692
-Not treated	64.8	59.7	65.9	66.1	0.036	1.429	1.024	1.994
Previous chiropractic treatment								
-Treated no effect (ref.)	13.6	20.4	11.1	9.3	0.003			
-Treated effect	2.6	2.7	2.9	2.6	0.165	2.036	0.746	5.563
-Not treated	83.8	77.0	86.0	88.2	0.001	2.053	1.354	3.113
Previous treatment medication								
-Treated no effect (ref.)	19.5	27.0	21.3	11.5	0.224			
-Treated effect	2.8	2.2	2.4	3.5	0.546	1.389	0.478	4.033
-Tot treated	77.7	70.8	76.3	85.0	0.086	1.367	0.957	1.952
Previous treatment pain clinic								
-Treated no effect (ref.)	6.0	10.2	6.2	4.5	0.152			
-Treated effect	0.9	0.9	0.7	0.3	0.787	1.278	0.216	7.548
-not treated	93.1	88.9	93.1	95.2	0.055	1.710	0.989	2.957
Previous treatment surgery								
-Treated no effect (ref.)	1.7	4.4	0.9	1.3	0.010			
-Treated effect	2.0	2.2	2.1	2.6	0.040	4.800	1.074	21.447

	Whole Study Population	LCGA classes of 1117 patients classified			OR for a favourable outcome			
		Class 1	Class 2	Class 3	Univariate analyses Class 2 versus Class 1			
Predictor variables		Not impr.	Improved	Low NRS	p-value	OR	95% CI of OR	
-Tot treated	96.3	93.4	97.1	96.2	0.003	5.318	1.797	15.739
Previous treatment other								
- Treated no effect (ref.)	16.6	19.5	15.6	17.3	0.369			
-Treated effect	4.4	4.0	3.5	7.0	0.851	1.086	0.457	2.581
-Not treated	78.9	76.5	81.0	75.7	0.171	1.323	0.886	1.974
Work status								
-No work (ref.)	22.3	31.4	21.1	17.9	0.005			
-Not physical work	51.6	48.2	51.4	58.1	0.013	1.586	1.100	2.286
-Physical work	26.1	20.4	27.5	24.0	0.002	2.012	1.296	3.122
Treatment characteristics								
Type of treatment								
-Manual Medicine (MM, ref.)	20.0	14.6	21.5	18.6	0.109			
-Orthomanual Medicine (OMM)	76.9	70.4	76.5	77.7	0.166	0.739	0.482	1.133
-Both MM and OMM	1.8	2.2	0.9	3.1	0.058	0.284	0.078	1.042
-Other treatment	1.3	0.4	1.1	0.7				
Number of treatment sessions	3.4 (1.6)	3.7 (1.5)	3.5 (1.6)	3.5 (1.6)	0.121	0.925	0.839	1.021
McKenzie	14.2	11.1	13.3	18.2	0.801	1.065	0.645	1.733
Treated.differently	4.4	6.1	4.6	2.1	0.008	1.000	1.000	1.000

Predictor values of the whole study population, and for the three groups of patients with different trajectories separately, and results of the univariate analyses of improved versus not improved patients (p-values, odds ratio's and 95% confidence intervals). For dichotomous variables percentages are presented, for continuous variables mean and (SD).

Prediction model

A multivariable prediction model was constructed , and the prediction model is presented in Table 6. In this model male gender, previous specialist visit, previous pain clinic treatment, having work, a shorter duration of the current episode, and a longer time since the complaints first started were predictive of a favourable course. No effect of previous chiropractic treatment was predictive of a non-favourable course, both compared with patients reporting a positive effect of previous chiropractic treatment, or patients not previously treated by a chiropractor. The fitted model showed an AUC of 0.677, with a non-significant Hosmer and Lemeshow test (0.734), supporting model fit, and an explained variance (R^2) of 0.10. Bootstrap validation resulted in a corrected R^2 of 0.09.

Table 6; Multivariable prediction model.

Predictor variables in multivariate model	Coefficient	OR	95% CI of OR	
			Lower	Upper
Gender (female)	-0.6089	0.5439	0.3691	0.8016
Time since complaints 1st started	0.0146	1.0147	0.9982	1.0314
Previous specialist visit	0.4642	1.5907	1.0658	2.3741
Previous visit pain clinic	0.1415	1.1520	1.0020	1.3244
Previous chiropractic treatment (treated without effect is reference)				
-Treated effective	0.2626	1.3003	0.4141	4.0831
-Not treated	0.5770	1.7806	1.0828	2.9282
Work status, no work is reference				
-Non-physical work	0.3575	1.4297	0.9245	2.2108
-Physical work	0.5327	1.7036	1.0042	2.8902
Duration of current episode (non-linear spline variable)				
-Duration of current episode	-0.5185	0.5954	0.4299	0.8247
-Spline variable duration of current episode	1.9662	7.1435	1.9694	25.9107
Intercept	0.8452	2.3285	1.0030	5.4060

Prediction model using baseline variables to predict a favourable outcome in patients presenting with high NRS for pain scores, including the Odds ratio's (OR) and the 95% Confidence Intervals of the OR.

Discussion

Studying the clinical course of low back pain in patients consulting MSK physicians in The Netherlands with Latent Class Growth Analyses, three distinct pain trajectories were identified. More than half of all the patients (52%) presented with high baseline pain scores and showed considerable improvement. A second group of patients with high baseline

pain scores (20% of all patients) showed no improvement during six months follow-up. A third group, with moderate baseline pain scores (28%) showed slight, but clinically relevant improvement. The prediction model presented showed a moderate AUC and a low explained variance, and one can question its usefulness in clinical practice. Apparently, with the baseline data collected in our study, it is hard to predict which patient might improve after consulting an MSK physician.

Previous studies almost invariably reported similar clusters of pain trajectories, generally including clusters with persistent high pain, clusters with more or less persistent moderate or low pain, clusters showing improvement, and clusters with a fluctuating pattern. The proportion of patients in each cluster, however, differed, possibly because of variations in the study designs. Patients were recruited in General Practice(6-8, 25), at chiropractic clinics(9, 26), combined in General Practice and Chiropractic clinics(27), combined in General Practice and physiotherapy practices(28), or in a population based survey(29). Also, studies varied in recruiting patients: i.e. only patients with chronic(28), only acute(25), or a mix of both chronic and acute LBP(6-9, 26, 27, 29). Moreover, follow-up periods varied from 12 weeks(25) to one year(9, 27-29), and follow-up measurements varied form weekly text messages(9, 26, 27) to monthly questionnaires(6-8). The population based study was the only study in which a cluster showing improvement was not reported(29). And the only study recruiting acute LBP patients showed high percentages of recovery(25). The clusters presented in our study are well comparable to those reported in other studies, with the exception of a cluster representing a fluctuating pattern. In our four class model a small cluster was added that would in other studies have qualified as fluctuating. We considered this cluster merely a subgroup of the consistent high pain cluster, eventually showing no improvement after six months follow-up, and therefor chose to use the three class model.

Most trajectory studies reported variables that were associated with group membership. Although varying variables were reported, only higher pain intensity(6-8, 25), longer duration(6, 7, 25, 27, 30), and more physical disability(8) were more or less consistently associated with a more severe trajectory. The same variables were reported in other studies to be associated with a worse prognosis in LBP patients, together with unemployment(31, 32). Similarly, in our study the duration of the current episode and unemployment were both associated with a lower probability of improvement, but baseline disability was not associated with the outcome. In our univariate analysis, ineffective previous treatments were consistently predictive of an unfavourable course. Of these previous treatments reported in our study, only chiropractic treatment ended up in our prediction model.

A challenging question remains to what extent the clinical course represented the natural recovery or the consequence of the treatment administered. Of all the treatment variables, only the number of treatment sessions ended up in the final model. Although this variable is unknown at the start of the treatment, retaining this variable in the model offers a correction for its influence on the outcome. Because this variable was not known at baseline it was not presented in Table 2.

Strengths and limitations

A strength of our study was the web-based data-collection, which enabled us to follow a large number of patients at a relatively low cost. In this way data could be collected from patients who consulted an MSK physician, with questionnaires that were tailored to their main complaint. A weakness was the difficulty to identify patients before consulting the physician. Our solution using web-based self-classification, aided with manikins, appeared to lead to a high proportion of miss-classification. Because we tailored the distribution of PROMs to the main complaint as reported by the patient, this miss-classification led to a high percentage of missing baseline ODI values. We therefore chose to use the physician's diagnosis to identify patients with LBP, and we imputed the baseline ODI. Another weakness of our study set-up was the high proportion of patients that discontinued their participation. The response rate gradually diminished during the follow-up period. Out of the 1117 patients included in the baseline population 93% responded after 6 weeks, 74% after 12 weeks, 58% after 18 weeks, and 43% at six months. We found that some baseline variables were related to loss to follow-up which made the MAR assumption more plausible, supporting multiple imputation of missing values.

Conclusion

In patients with low back pain, three different clinical courses were identified in the six months after consulting an MSK physician in the Netherlands. A large group of patients presented with high baseline pain scores, and showed improvement. In patients with a high pain score at baseline, a multivariable prediction model showed a number of predictors of a favourable course. In this model, male gender, longer time since the complaints first started, shorter duration of the current episode of pain, previous specialist visit, previous pain clinic visit, effective treatment by a chiropractor, or no previous chiropractic treatment, and having work were predictors of a favourable course. The prediction model, however, showed a low AUC and explained only 9% of the variance. It is a continuing challenge to identify predictors of a favourable outcome in LBP patients.

References

1. Lambeek LC, van Tulder MW, Swinkels IC, Koppes LL, Anema JR, van Mechelen W. The trend in total cost of back pain in The Netherlands in the period 2002 to 2007. Spine (Phila Pa 1976). 2011;36(13):1050-8.
2. Deyo RA. The Role of Spinal Manipulation in the Treatment of Low Back Pain. JAMA. 2017;317(14):1418-9.
3. Dunn KM, Croft PR. Epidemiology and natural history of low back pain. Eura Medicophys. 2004;40(1):9-13.
4. van den Hoogen HJ, Koes BW, van Eijk JT, Bouter LM, Deville W. On the course of low back pain in general practice: a one year follow up study. Ann Rheum Dis. 1998;57(1):13-9.
5. Itz CJ, Geurts JW, van Kleef M, Nelemans P. Clinical course of non-specific low back pain: a systematic review of prospective cohort studies set in primary care. Eur J Pain. 2013;17(1):5-15.
6. Dunn KM, Campbell P, Jordan KP. Long-term trajectories of back pain: cohort study with 7-year follow-up. BMJ Open. 2013;3(12):e003838.
7. Chen Y, Campbell P, Strauss VY, Foster NE, Jordan KP, Dunn KM. Trajectories and predictors of the long-term course of low back pain: cohort study with 5-year follow-up. Pain. 2017;159(2):252-260
8. Dunn KM, Jordan K, Croft PR. Characterizing the course of low back pain: a latent class analysis. Am J Epidemiol. 2006;163(8):754-61.
9. Ailliet L, Rubinstein SM, Hoekstra T, van Tulder MW, de Vet HCW. Long-term trajectories of patients with neck pain and low back pain presenting to chiropractic care: A latent class growth analysis. Eur J Pain. 2018;22(1):103-13.
10. Schuller W, Ostelo R, Rohrich DC, Apeldoorn AT, de Vet HCW. Physicians using spinal manipulative treatment in The Netherlands: a description of their characteristics and their patients. BMC Musculoskelet Disord. 2017;18(1):512.
11. Lamberts H, Wood, M. (Eds.). International Classification of Primary Care (ICPC). Oxford: Oxford University Press; 1987.
12. Brazier JE, Roberts J. The estimation of a preference-based measure of health from the SF-12. Med Care. 2004;42(9):851-9.
13. Vendrig A, Deutz P, Vink I. Nederlandse vertaling en bewerking van de Fear-Avoidance Beliefs Questionnaire. Nederlands tijdschrift voor pijn en pijnbestrijding. 1998;18(1):11-4.
14. Waddell G, Newton M, Henderson I, Somerville D, Main CJ. A Fear-Avoidance Beliefs Questionnaire (FABQ) and the role of fear-avoidance beliefs in chronic low back pain and disability. Pain. 1993;52(2):157-68.
15. Aaronson NK, Muller M, Cohen PD, Essink-Bot ML, Fekkes M, Sanderman R, et al. Translation, validation, and norming of the Dutch language version of the SF-36 Health Survey in community and chronic disease populations. J Clin Epidemiol. 1998;51(11):1055-68.
16. Beurskens AJ, de Vet HC, Koke AJ. Responsiveness of functional status in low back pain: a comparison of different instruments. Pain. 1996;65(1):71-6.

17. Berlin KS, Parra GR, Williams NA. An Introduction to Latent Variable Mixture Modeling (Part 2): Longitudinal Latent Class Growth Analysis and Growth Mixture Models. J Pediatr Psychol. 2014;39(2):188-203.
18. Nylund KL, Asparoutiov T, Muthen BO. Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. Struct Equ Modeling. 2007;14(4):535-69.
19. Muthén LK, Muthén, B. Mplus statistical analysis with latent variables user's guide Los Angeles, CA: Muthén & Muthén; [
20. Hosmer DW, Lemeshow S. Goodness of Fit Tests for the Multiple Logistic Regression-Model. Commun Stat a-Theor. 1980;9(10):1043-69.
21. Nagelkerke NJD. A Note on a General Definition of the Coefficient of Determination. Biometrika. 1991;78(3):691-2.
22. Harrell FE. Regression Modelling Strategies. 2nd ed. Switzerland: Springer International Publishing; 2015.
23. Heymans MW. R package psfmi: Predictor Selection Functions for Logistic and Cox regression models in multiply imputed datasets; 0.2.0. 2020 [Available from: https://cran.r-project.org/web/packages/psfmi/index.html.
24. Rubin DB. Multiple Imputation for Nonresponse in Surveys. New York: J. Wiley & Sons; 1987.
25. Downie AS, Hancock MJ, Rzewuska M, Williams CM, Lin CW, Maher CG. Trajectories of acute low back pain: a latent class growth analysis. Pain. 2016;157(1):225-34.
26. Axen I, Bodin L, Bergstrom G, Halasz L, Lange F, Lovgren PW, et al. Clustering patients on the basis of their individual course of low back pain over a six month period. BMC Musculoskelet Disord. 2011;12:99.
27. Kongsted A, Kent P, Hestbaek L, Vach W. Patients with low back pain had distinct clinical course patterns that were typically neither complete recovery nor constant pain. A latent class analysis of longitudinal data. Spine J. 2015;15(5):885-94.
28. Macedo LG, Maher CG, Latimer J, McAuley JH, Hodges PW, Rogers WT. Nature and determinants of the course of chronic low back pain over a 12-month period: a cluster analysis. Phys Ther. 2014;94(2):210-21.
29. Tamcan O, Mannion AF, Eisenring C, Horisberger B, Elfering A, Muller U. The course of chronic and recurrent low back pain in the general population. Pain. 2010;150(3):451-7.
30. Hill JC, Konstantinou K, Egbewale BE, Dunn KM, Lewis M, van der Windt D. Clinical outcomes among low back pain consulters with referred leg pain in primary care. Spine (Phila Pa 1976). 2011;36(25):2168-75.
31. Dunn KM, Jordan KP, Croft PR. Contributions of prognostic factors for poor outcome in primary care low back pain patients. Eur J Pain. 2011;15(3):313-9.
32. Mallen CD, Peat G, Thomas E, Dunn KM, Croft PR. Prognostic factors for musculoskeletal pain in primary care: a systematic review. Br J Gen Pract. 2007;57(541):655-61.

Chapter 4.

Adverse events after spinal manipulative treatment by musculoskeletal physicians in The Netherlands

Wouter Schuller, Raymond W. Ostelo, Daphne C. Rohrich, Henrica C.W. de Vet

Submitted

Abstract

Objectives: To evaluate adverse events after spinal manipulative treatment of low back pain (LBP) and neck pain (NP) by musculoskeletal (MSK) physicians in The Netherlands.

Methods: In an observational cohort study MSK physicians recorded various baseline and treatment variables of new patients. Patients were asked to answer questionnaires at baseline including information about previous medical consumption, together with PROMs measuring the level of pain and functional status. Three months after the start of the treatment, patients were invited to answer questionnaires enquiring after the type, the severity, and the duration of adverse events.

Results: Of 823 LBP and 315 NP patients answering the adverse events questionnaire, 362 patients (31.8%) reported a total of 683 adverse events. All patients except five were treated with a manipulative or mobilising technique, or both, in, on average, 3-6 sessions (range 1-12). The highest proportion of patients (15.8%) reported only one adverse event, and the adverse event most frequently reported was fatigue (10.9% of all patients). Patients with a main complaint of NP reported adverse events more frequently (38.4%) than patients with a main complaint of LBP (29.3%), and NP patients also displayed a different pattern of adverse events. Most adverse events were not severe and resolved within a week, but some patients reported adverse events to be more severe (6.9%) or lasting longer (7.1%).

Conclusion: Adverse events after spinal manipulative treatment by musculoskeletal physicians were common but mostly short-lived and mild to moderately severe. Neck pain patients displayed different adverse events than low back pain patients.

Introduction

Spinal Manipulative Treatment (SMT) is frequently used in the treatment of a variety of mainly musculoskeletal complaints, such as low back pain or neck pain(1-3). SMT is used among others by chiropractors, manual therapists, osteopaths, and physicians, and a wide range of SMT techniques are applied in a variety of clinical settings. While major adverse events after SMT are rare, minor adverse events appear to be common. A number of studies reported on minor adverse events after chiropractic treatment(4-6), manual therapy(7), or both(8). Minor adverse events were reported by 34-61% of patients, most frequently increased pain or stiffness, tiredness, headache, and radiating discomfort, but also dizziness, nausea, tinnitus and impaired vision(6). In clinical trials evaluating the effect of SMT, adverse events were described in control groups as well. In two trials evaluating adverse events after chiropractic treatment in patients with neck pain, manipulative techniques were associated with a higher proportion of adverse events than mobilizing techniques(9, 10). One cohort study reported more adverse events in neck pain patients after the use of manipulative techniques with a rotatory component(11).

It is not unlikely that the type and frequency of adverse events is influenced by the clinical setting, or by the practitioner delivering the treatment. In The Netherlands, there is a group of physicians who have specialised in musculoskeletal (MSK) medicine(12). These physicians frequently use a form of SMT in their treatment. A previous study described the characteristics of MSK physicians and their patients(12). The clinical setting, and part of the SMT techniques used by MSK physicians differs from the clinical setting and the techniques used by chiropractors and manual therapists(12, 13). We conducted this study to describe adverse events after treatment by MSK physicians, and to examine whether the frequency, type, and severity of reported adverse events differed between patients treated for low back or neck pain. When the neck was treated we assessed whether the reported adverse events differed between patients treated with a mobilising technique and patients treated with a manipulating technique. We also examined whether the report of adverse events was associated with the reported improvement.

Materials and methods

Study design and participants

We conducted a prospective, observational cohort study. All MSK physicians registered with the Dutch Association for Musculoskeletal Medicine were invited to participate in the study. Participating physicians were instructed to register all patients who presented for the first time in their MSK practice in a web-based registry, and to invite these patients to take

part in the study. Inclusion criteria were aged ≥ 18, and sufficient mastery of the Dutch language to answer questionnaires in Dutch. If patients gave informed consent, the treating physician entered email addresses of the recruited patients in the web-based registry. Thereafter, a specially designed computer program (Readmail) was used to automatically distribute invitations to patients by email to fill in web-based questionnaires. Our study procedures were approved by the Medical Ethical Committee (METc) of the VU Medical Center (2012/414).

Study procedure

Both the treating physicians and the individual patients provided data via web-based registries. Study procedures were explained to participating physicians at specially organised information sessions. Next to this, a research assistant visited all participating practices to explain the procedures. Practices that agreed to participate at a later stage were informed by telephone. Instructions were to ask all consecutive patients presented for a first consultation to participate in the study. Recruited patients received invitations to fill in web-based questionnaires within three weeks before the first consultation. After 3 months patients received another email, inviting them to fill in a questionnaire enquiring after the type, the severity, and the duration of adverse events. When patients did not respond, a maximum of three reminders were sent within a period of two weeks. Both the invitational email and the web-based questionnaires contained links to a folder with information about the study. In the invitational email patients could indicate that they wanted to discontinue their participation. These patients received a stop-questionnaire, in which they were asked why they wanted to discontinue, and received no further invitations.

Measurement

At baseline, physicians registered data about age, gender, type and duration of the main complaint, and the existence of concomitant complaints. Complaints were registered according to the International Classification of Primary Care (ICPC)(14). At the end of treatment data were registered about the number of treatment sessions, the type of treatment used, whether the neck had been treated, and if the neck had been treated whether a manipulative or mobilising technique had been used.

At baseline, patients were asked to indicate whether their main complaint was either low back pain, neck pain or any other complaint. This question was supported by text and manikins, explaining which area was considered to cover neck pain or low back pain. For other complaints, patients could explain these in text. Patients were asked to indicate whether their pain radiated to the legs or arms, and whether they had radiating neurologic complaints (numbness or pins and needles in their legs or arms). Patients were also asked

about the duration of the current episode, the time since the first episode, educational level, work status, previous specialist consultations, and the type of previous treatments. In addition, patients were asked to complete a set of Patient Reported Outcome Measures (PROMs), including a Numerical Rating Scale (NRS) for pain severity, the SF6D(15), the Fear Avoidance Beliefs Questionnaire (FABQ)(16, 17), and the Oswestry Disability Index (ODI)(18, 19) for patients who indicated a main complaint of low back pain, and the Neck Disability Index(20) for patients who indicated a main complaint of neck pain. The SF6D is a short version of the SF36, measuring health related quality of life(15). Scores range from 0-1, with higher scores indicating lower quality of life. The FABQ consists of 16 items, and measures pain related fear in LBP patients(16, 17). Higher scores indicate more pain related fear. The ODI consists of 10 items with scores ranging from 0-50, with higher scores indicating more disability because of LBP(18, 21). The NDI consists of 10 items with scores ranging from 0-50, with higher scores indicating more disability because of neck pain(20). All PROMs have been validated in patients with low back pain, and are frequently used in research. At follow-up patients were asked to answer the same PROMs, except for the FABQ, together with a question about the Global Perceived Effect (GPE) of treatment.

Adverse events

Three months after the baseline measurement patients were asked to complete a questionnaire about any adverse event experienced after MSK treatment. This questionnaire was layered as follows. A first question asked whether the patient had experienced any adverse event after any of the treatment sessions. If this question was answered affirmatively, a next question asked about the type of adverse event, with the following response options; dizziness, light-headedness, new headache, new radiating pain in the arm(s), new radiating pain in the leg(s), new pins and needles in the arm(s), new pins and needles in the leg(s), general malaise, fatigue, or other. Furthermore, patients with a main complaint of neck pain were asked whether they experienced any new low back pain, and patients with a main complaint of low back pain were asked whether they had experienced any new neck pain. For each type of adverse event reported by the patient, questions were added asking about the severity and the duration of the adverse event. The severity was questioned on a 5-point ordinal scale, with the following response options: (1) very mild complaints, very few limitations; (2) mild complaints, few limitations; (3) moderate complaints, some limitations; (4) substantial complaints, substantial limitations; (5) severe complaints, severe limitations. The duration was questioned on a 5-point ordinal scale as well, with the following response options: (1) only a few minutes; (2) longer than a few minutes, but shorter than a day; (3) longer than a day, but shorter than a week; (4) longer than a week, but shorter than a month, and (5) continuous. Two more questions were asked about whether the patient had visited the GP, or had taken up sick leave, due to any of the reported adverse events.

Treatment

MSK physicians use an array of treatment possibilities, almost invariably involving a form of SMT(12). Other treatment options used include McKenzie treatment, prescription medication, or injection treatment(12). MSK physicians were asked in the follow-up questionnaire at the end of the treatment to register the type of treatment administered, with the following response options: type of manipulative technique, McKenzie, medication, injections, or other. In addition, physicians were asked to register the number of treatment sessions used. If the neck was treated, physicians registered whether a manipulative or a mobilising technique was used.

Statistical analyses

Our study population consisted of all patients with low back or neck pain who answered the baseline questionnaire. To examine the possibility of selection bias, baseline characteristics of the group of patients who answered the follow-up questionnaire at three months were compared with the baseline characteristics of the group of patients who did not answer the adverse events questionnaire at three months. Selective loss to follow-up was further evaluated by assessing the proportion of patients who indicated in the stop-questionnaire that they discontinued their participation because their complaints had deteriorated or because other complaints had developed. Descriptive analyses of the reported adverse events are presented for the population as a whole, and for relevant subgroups: patients with a primary complaint of low back pain or neck pain, and patients in whom the neck was treated with a manipulative or with a mobilising technique. Furthermore descriptive analyses are presented of the severity and the duration of the adverse events, and whether the adverse event was the reason for patients to visit their general practitioner, or to take time off on sick leave. It was also evaluated how many patients showed a deterioration of their NRS for pain scores of more than 30%. The association of the report of any adverse event with the reported improvement at three months was tested with a X^2 test. For this purpose the GPE was dichotomised into a group of patients considered to be improved (GPE 1-3, completely recovered, strongly improved, and little improved), and a group of patients considered not to be improved (all other GPE scores).

Results

Study population

A total of 1391 LBP patients and 549 NP patients answered the baseline questionnaire. Of these, 823 (59%) LBP and 315 (57%) NP patients answered the questionnaire at three months follow-up. In Table 1 the baseline data of 1138 patients who answered the follow-up questionnaire are compared with the baseline data of 802 patients who only answered

the baseline questionnaire, showing only minor differences. Patients who did answer the follow-up questionnaire were, on average, slightly older (48.2 versus 44.2 years of age), of female gender (62.4% versus 58.4%), and higher educated (65.2% versus 56.6%). During the follow-up period of three months 439 patients answered the stop questionnaire. Only 2.3% of these patients indicated worsening of existing complaints, or new complaints as the reason to discontinue their participation.

Concomitant complaints were reported by 63.8% of the patients, most frequently involving concomitant upper extremity complaints (25.8%), neck pain (17.8%), headache (11.1%), lower extremity complaints (11.2%), or other (10.3%). Other complaints explained in text most generally consisted of local aches or tenderness. All patients except five were treated with some form of SMT in, on average, 3.6 treatment sessions (range 1-12). Of all other treatment options, only the McKenzie treatment was used frequently (12.7%), but invariably in addition to the SMT. In a proportion of patients (17.1%) without a main complaint of NP, and without concomitant NP, headache, or upper extremity complaints, the neck was still included in the treatment.

Reported adverse events

Of all 1138 patients 362 (31.8%) reported a total of 683 adverse events. Most frequently a single adverse event was reported (15.8%). Table 2 presents the type of adverse events reported for the population as a whole, and for relevant subgroups (LBP patients, NP patients, patients in whom the neck was treated with a manipulative technique, and patients in whom the neck was treated with a mobilising technique). The adverse event most frequently reported was fatigue (10.9% of all patients (N=1138), 8.5% of LBP patients (N=823), 17.1% of NP patients (N=315)). In general, patients treated for NP more frequently reported adverse events, except for radiating pain or pins and needles in the leg(s). NP patients displayed a different pattern of adverse events than LBP patients. Dizziness, feeling lightheaded, headache, pins and needles in the arm(s), and fatigue were reported more frequently by NP patients. Radiating pain in the leg(s) and pins and needles in the leg(s) were reported more frequently by LBP patients. In a proportion of patients with a main complaint of LBP the neck was treated as well (42.8% of LBP patients), even without a concomitant complaint of NP, headache, or upper extremity complaints (17.1% of LBP patients). There were differences in the frequency and type of the reported adverse events between patients in whom the neck was treated with a manipulative or mobilising technique. Patients in whom the neck was treated with a mobilising technique reported significantly more adverse events. After the use of a manipulative technique dizziness and headache were more frequently reported, while after the use of a mobilising technique light-headedness, radiating pain, feeling generally unwell and fatigue were more frequently reported. These differences were only statistically

significant for fatigue. The NRS at three months was deteriorated more than 30% in 79 patients (6.9%). Either adverse events or deterioration was reported by 414 patients (36.8%).

Table 3 presents the severity and the duration of the adverse events. Generally, patients reported the severity of the adverse events to be limited (first three categories of the ordinal scale, 1.9% very mild complaints, very few limitations, 3.2% mild complaints, few limitations, and 6.7% moderate complaints, some limitations), but 6.9% of all patients reported adverse events to be more severe (4.7% substantial complaints, substantial limitations, 2.2% severe complaints, severe limitations). Similarly, the duration was reported to be limited by most patients (1.1% lasting only a few minutes, 1.9% lasting longer than a few minutes, but shorter than a day, and 7.3% lasting longer than a day, but shorter than a week), but 7.2% of all patients reported adverse events to last longer (3.7% lasting longer than a week, but shorter than a month, and 3.5% continuous). There was little difference in severity between the various adverse events reported. The duration of adverse events was generally reported to be longer than a day, but shorter than a month. A small number of patients reported that the adverse events where still present at the time of answering the questionnaire (3.4%). A number of 23 patients (2.0%) reported to have visited their general practitioner due to the adverse events, and 41 patients (3.6%) reported to have taken time of on sick leave due to the adverse events.

Considering the relation between the occurrence of any adverse events and the improvement reported at three months, 82.0% of patients with adverse events showed improvement after three months compared to 82.6% of patient without adverse events. The X^2 test was not statistically significant (OR 0.96, p=0.818).

Table 1; Comparing baseline characteristics of patients who completed the follow-up questionnaire and patients who only answered the baseline questionnaire.

Variable	Only baseline N=802	Baseline and follow-up N=1138
General characteristics		
Age, *mean (SD)**	44.2 (13.6)	48.2 (13.1)
Gender *(male)*	41.6	37.6
Educational level *(high)*	56.6	65.2
Work *(having work)*	79.4	76.9
Physical work *(doing physical work)*	34.8	30.1
Complaints		
Neck pain	29.2	27.7
Low Back Pain	70.8	72.3

Variable	Only baseline N=802	Baseline and follow-up N=1138
Time since 1st complaints, *mean years (SD)**	10.3 (11.4)	10.6 (12.1)
Duration current episode, *mean years (SD)**	2.4 (5.2)	2.1 (4.8)
Radiating pain	40.1	35.8
Radiating neurologic complaints	28.2	25.4
Pain avoidant	65.1	65.6
Baseline SF-6D, *mean (SD)**	0.74 (0.11)	0.75 (0.10)
Baseline NDI (neck pain patients), *mean (SD)**	14.8 (7.4)	13.7 (6.9)
Baseline ODI (low back pain patients), *mean (SD)**	24.4 (16.5)	22.5 (15.1)
Concomitant complaints		
- Concomitant complaints reported	63.3	63.8
- Concomitant Neck Pain	17.6	17.8
- Concomitant Low Back Pain	8.7	8.7
- Concomitant Headache	11.0	11.1
- Concomitant upper extremity	25.4	25.8
- Concomitant lower extremity	10.1	11.2
- Concomitant other	8.5	10.3
- Concomitant pins and needles	4.2	2.6
Previous specialist visit		
Specialist visit	60.8	65.1
- Neurologist	23.5	21.3
- Orthopedic surgeon	21.1	18.5
- Rehabilitation	5.4	5.2
- Pain clinic	6.6	6.7
Medication		
Medication none	31.9	33.0
Medication seldom	23.6	23.4
Madication every now and then	30.5	29.4
Medication daily	14.0	14.1
Previous treatment		
Physiotherapy	69.9	69.8
Manual therapy	36.4	37.1
Chiropractic	17.6	16.5
Medication	21.5	22.0
Pain clinic	6.7	5.7
Surgery	2.9	3.3
Other	21.2	24.5

*For continuous variables the means and standard deviations (SD) are presented. For all other variables percentages are presented.

Table 2; Type of adverse events (in percentages) reported by the whole study population and by relevant subgroups (NP patients, LBP patients, patients in whom the neck had been treated with a manipulative technique, and patients in whom the neck had been treated with a mobilising technique). Within group percentages are presented.

Type of adverse event	All patients (N=1138)	LBP patients (N=823)	Neck patients (N=315)	Neck manipulated (N=85)	Neck mobilized (N=529)
Any adverse event	31.8	29.3	38.4	27.1	37.2
Including 30% worse NRS	36.4	34.4	41.6	29.4	41.6
Neck pain	3.3	3.5	nn	4.7	4.5
Low back pain	3.0	nn	10.2	1.2	4.2
Dizziness	2.5	1.7	4.8	5.9	3.6
Lightheaded	3.1	2.1	5.7	3.5	4.5
Headache	5.0	3.6	8.6	8.2	6.2
Radiating pain arm(s)	2.9	1.2	7.3	2.4	4.0
Radiating pain leg(s)	4.5	4.9	3.5	1.2	4.7
Pins and needles arm(s)	1.8	1.1	3.5	2.4	2.5
Pins and needles leg(s)	3.7	4.6	1.3	3.5	3.4
Generally unwell	5.6	4.7	7.9	5.9	7.2
Fatigue	10.9	8.5	17.1	7.1	15.9
Other	13.7	13.7	13.7	14.1	14.7
Sick leave	3.6	3.4	4.1	0.0	3.8
Consulted GP	2.0	2.2	1.6	1.2	1.7

Table 3; Percentages for severity and duration of adverse events, measured on a 5-point ordinal scale.

Type of adverse event	N	Severity					Duration				
		1	2	3	4	5	1	2	3	4	5
Total	1138	1.9	3.2	6.7	4.7	2.2	1.1	1.9	7.2	3.7	3.4
Neck pain	38	0.0	0.4	2.2	0.7	0.1	0.0	0.3	1.8	0.4	0.8
Back pain	34	0.1	0.1	1.2	1.3	0.3	0.1	0.2	1.1	1.1	0.5
Dizziness	29	0.1	0.5	1.5	0.4	0.1	0.5	0.8	0.7	0.3	0.3
Lightheaded	34	0.4	0.6	1.3	0.7	0.0	0.2	1.1	0.9	0.5	0.4
Headache	56	0.0	0.4	2.8	1.5	0.2	0.0	1.1	2.4	0.9	0.6
Radiating pain arm(s)	33	0.2	0.1	1.2	1.4	0.0	0.0	0.5	1.1	0.5	0.8
Radiating pain leg(s)	51	0.0	0.3	2.2	1.7	0.4	0.0	0.2	1.1	1.3	1.9
Pins and needles arm(s)	20	0.1	0.3	0.7	0.6	0.1	0.0	0.4	0.4	0.5	0.4
Pins and needles leg(s)	42	0.1	0.5	1.9	1.0	0.2	0.0	0.4	1.3	0.5	1.5
Generally unwell	63	0.1	0.6	2.6	1.6	0.6	0.0	0.7	2.8	1.4	0.6
Fatigue	123	0.3	1.5	4.8	3.5	0.8	0.0	2.7	4.4	2.0	1.7

Type of adverse event	N	Severity					Duration				
		1	2	3	4	5	1	2	3	4	5
Other	155	0.9	2.2	5.7	3.7	1.1	0.4	1.1	6.5	2.9	2.6
LBP patients	823	1.5	3.0	6.3	4.6	1.7	1.1	1.8	7.0	4.3	2.9
Neck pain patients	315	3.2	3.5	7.6	5.1	3.5	1.3	2.2	7.6	2.2	4.8

Severity: (1) Very mild complaints, very few limitations; 2. Mild complaints, few limitations; 3. Moderate complaints, some limitations; 4. Substantial complaints, substantial limitations; 5. Severe complaints, severe limitations.
Duration: (1) Only a few minutes; (2) Longer than a few minutes, but shorter than a day; (3) Longer than a day, but shorter than a week; (4) Longer than a week, but shorter than a month, and (5) Continuous.

Discussion

This was a first study of adverse events reported by patients after treatment by MSK physicians in The Netherlands. In general, adverse events were reported by 31.8% of the patients, and were mostly reported to be short-lived and mild to moderately severe. Neck pain patients reported adverse events more frequently than low back pain patients (38.4% versus 29.3%), especially fatigue (17.1%). Some adverse events were reported to be more severe (6.9%) and longer lasting (7.2%), and some patients visited their GP (2.0%), or reported taking time off on sick leave (3.6%) due to the adverse events. There was no relation between the occurrence of adverse events and the global perceived effect of the treatment.

Mild to moderately severe adverse events after SMT were reported by a number of observational studies, and by several RCT's studying the effectivity of SMT. A systematic review by Carnes et al. presented a pooled proportion of patients reporting adverse events that was higher for the observational studies (41%), than for the RCT's (22%)(7). It was suggested that in most RCT's adverse events may be underreported because adverse events were not the primary outcome. RCT's designed to evaluate adverse events reported proportions that were comparable to those reported by observational studies(6, 10). The proportion of adverse events in our study was low (31.8%) compared to other observational studies. However, comparing reports of adverse events is hampered by differences in study designs. First of all, different populations were recruited and different SMT techniques were applied. Patients were treated by chiropractors(8, 22-25), or by a mixed group of practitioners: chiropractors, physical therapists and osteopaths(26). Some studies recruited all patients consulting an SMT practitioner(6, 25), while other studies were limited to neck pain patients only(9, 10, 24). Several studies recruited patients from a population that returned for treatment(10, 24), instead of new patients, which may have selected patients with positive previous experience with this treatment. Secondly, there is no standardised way to measure adverse events, and studies varied in the way that adverse events were evaluated. In some studies patient

questionnaires were used(22, 24, 26) while in other studies questionnaires were filled in by the treating therapist(25). And the questionnaires used varied from single open end questions(25) to closed sets of questions, listing several response options(6, 24). Thirdly, the time frame and the recall period varied between studies. In the study by Rubinstein et al., e.g., the occurrence of adverse events was measured prior to the 2nd and 4th visit, asking about any changes following the 1st and the 2nd and 3rd treatment respectively. In the study by Senstad et.al. the chiropractor asked about adverse events experienced after the previous treatment at the next visit.

Although comparison of the reported prevalence of adverse events after SMT between studies is not straightforward, our study contributes to the awareness that minor adverse events after SMT are common, but generally short-lived and not severe. To put these figures into perspective, in RCTs adverse events of SMT could be compared to those reported in control arms. Control groups treated by physical therapy or general practitioners reported adverse events less frequently than the SMT intervention groups(27), control groups receiving exercise treatment or sham manipulations reported similar percentages of adverse events(6, 7, 28), and control groups receiving non-steroid anti-inflammatory medication reported adverse events more frequently(7). It is therefore not clear to what extent adverse events can be attributed to the SMT. Nevertheless, practitioners should be aware of the prevalence of non-serious adverse events.

Limitations

A limitation of our study is the substantial loss to follow-up. We found minor differences in baseline characteristics between patients who only answered the baseline questionnaire and patients who also answered the follow-up questionnaire. These differences were observed on variables that, in literature, seem to be not highly associated with adverse events. Furthermore, part of the patients who discontinued their participation answered a stop questionnaire, in which only a minority indicated that adverse events were the reason to discontinue their participation. Therefore we consider the risk of bias because of non-response in our study as low.

Conclusion

A proportion of 31.8% of patients with low back pain or neck pain treated by musculoskeletal physicians in The Netherlands reported adverse events. Adverse events typically were short-lived and not severe, but a small proportion of adverse events were reported to be more severe (6.9%) or longer lasting (7.2%). Neck pain patients displayed different patterns of adverse events compared to low back pain patients. Patients in whom the neck had been

treated with a mobilising technique more frequently reported adverse events, which was largely due to the frequent reporting of fatigue. There was no relation between the report of adverse events and the reported improvement after three months follow-up.

References

1. Gross A, Miller J, D'Sylva J, Burnie SJ, Goldsmith CH, Graham N, et al. Manipulation or mobilisation for neck pain. Cochrane Database Syst Rev. 2010(1):CD004249. doi: 10.1002/14651858.CD004249.pub3. PubMed PMID: 20091561.
2. Rubinstein SM, Terwee CB, Assendelft WJ, de Boer MR, van Tulder MW. Spinal manipulative therapy for acute low back pain: an update of the cochrane review. Spine (Phila Pa 1976). 2013;38(3):E158-77. doi: 10.1097/BRS.0b013e31827dd89d. PubMed PMID: 23169072.
3. Rubinstein SM, van Middelkoop M, Assendelft WJ, de Boer MR, van Tulder MW. Spinal manipulative therapy for chronic low-back pain: an update of a Cochrane review. Spine (Phila Pa 1976). 2011;36(13):E825-46. doi: 10.1097/BRS.0b013e3182197fe1. PubMed PMID: 21593658.
4. Hurwitz EL, Morgenstern H. Adverse reactions to chiropractic care in the UCLA Neck Pain Study. J Manipulative Physiol Ther. 2006;29(7):597-8; author reply 8-9. doi: 10.1016/j.jmpt.2006.07.002. PubMed PMID: 16949953.
5. Senstad O, Leboeuf-Yde C, Borchgrevink C. Predictors of side effects to spinal manipulative therapy. J Manipulative Physiol Ther. 1996;19(7):441-5. PubMed PMID: 8890024.
6. Walker BF, Hebert JJ, Stomski NJ, Clarke BR, Bowden RS, Losco B, et al. Outcomes of usual chiropractic. The OUCH randomised controlled trial of adverse events. Spine (Phila Pa 1976). 2013;38(20):1723-9. doi: 10.1097/BRS.0b013e31829fefe4. PubMed PMID: 23778372.
7. Carnes D, Mars TS, Mullinger B, Froud R, Underwood M. Adverse events and manual therapy: a systematic review. Man Ther. 2010;15(4):355-63. doi: 10.1016/j.math.2009.12.006. PubMed PMID: 20097115.
8. Rubinstein SM, de Zoete A, van Middelkoop M, Assendelft WJJ, de Boer MR, van Tulder MW. Benefits and harms of spinal manipulative therapy for the treatment of chronic low back pain: systematic review and meta-analysis of randomised controlled trials. BMJ. 2019;364:l689. doi: 10.1136/bmj.l689. PubMed PMID: 30867144.
9. Hurwitz EL, Morgenstern H, Vassilaki M, Chiang LM. Adverse reactions to chiropractic treatment and their effects on satisfaction and clinical outcomes among patients enrolled in the UCLA Neck Pain Study. J Manipulative Physiol Ther. 2004;27(1):16-25. doi: 10.1016/j.jmpt.2003.11.002. PubMed PMID: 14739870.
10. Hurwitz EL, Morgenstern H, Vassilaki M, Chiang LM. Frequency and clinical predictors of adverse reactions to chiropractic care in the UCLA neck pain study. Spine (Phila Pa 1976). 2005;30(13):1477-84. PubMed PMID: 15990659.

11. Rubinstein SM, Leboeuf-Yde C, Knol DL, de Koekkoek TE, Pfeifle CE, van Tulder MW. Predictors of adverse events following chiropractic care for patients with neck pain. J Manipulative Physiol Ther. 2008;31(2):94-103. doi: 10.1016/j.jmpt.2007.12.006. PubMed PMID: 18328935.
12. Schuller W, Ostelo R, Rohrich DC, Apeldoorn AT, de Vet HCW. Physicians using spinal manipulative treatment in The Netherlands: a description of their characteristics and their patients. BMC Musculoskelet Disord. 2017;18(1):512. doi: 10.1186/s12891-017-1863-z. PubMed PMID: 29207995; PubMed Central PMCID: PMCPMC5718083.
13. van de Veen EA, de Vet HC, Pool JJ, Schuller W, de Zoete A, Bouter LM. Variance in manual treatment of nonspecific low back pain between orthomanual physicians, manual therapists, and chiropractors. J Manipulative Physiol Ther. 2005;28(2):108-16. doi: 10.1016/j.jmpt.2005.01.008. PubMed PMID: 15800510.
14. Lamberts H, Wood, M. (Eds.). International Classification of Primary Care (ICPC). Oxford: Oxford University Press; 1987.
15. Brazier JE, Roberts J. The estimation of a preference-based measure of health from the SF-12. Med Care. 2004;42(9):851-9. PubMed PMID: 15319610.
16. Vendrig A, Deutz P, Vink I. Nederlandse vertaling en bewerking van de Fear-Avoidance Beliefs Questionnaire. Nederlands tijdschrift voor pijn en pijnbestrijding. 1998;18(1):11-4.
17. Waddell G, Newton M, Henderson I, Somerville D, Main CJ. A Fear-Avoidance Beliefs Questionnaire (FABQ) and the role of fear-avoidance beliefs in chronic low back pain and disability. Pain. 1993;52(2):157-68. PubMed PMID: 8455963.
18. Aaronson NK, Muller M, Cohen PD, Essink-Bot ML, Fekkes M, Sanderman R, et al. Translation, validation, and norming of the Dutch language version of the SF-36 Health Survey in community and chronic disease populations. J Clin Epidemiol. 1998;51(11):1055-68. PubMed PMID: 9817123.
19. Beurskens AJ, de Vet HC, Koke AJ. Responsiveness of functional status in low back pain: a comparison of different instruments. Pain. 1996;65(1):71-6.
20. Vernon H, Mior S. The Neck Disability Index: a study of reliability and validity. J Manipulative Physiol Ther. 1991;14(7):409-15.
21. Beurskens AJ, de Vet HC, Koke AJ, van der Heijden GJ, Knipschild PG. Measuring the functional status of patients with low back pain. Assessment of the quality of four disease-specific questionnaires. Spine (Phila Pa 1976). 1995;20(9):1017-28.
22. Barrett AJ, Breen AC. Adverse effects of spinal manipulation. J R Soc Med. 2000;93(5):258-9. doi: 10.1177/014107680009300511. PubMed PMID: 10884771; PubMed Central PMCID: PMCPMC1298004.
23. Leboeuf-Yde C, Hennius B, Rudberg E, Leufvenmark P, Thunman M. Side effects of chiropractic treatment: a prospective study. J Manipulative Physiol Ther. 1997;20(8):511-5. PubMed PMID: 9345679.
24. Rubinstein SM, Leboeuf-Yde C, Knol DL, de Koekkoek TE, Pfeifle CE, van Tulder MW. The benefits outweigh the risks for patients undergoing chiropractic care for neck pain: a prospective,

multicenter, cohort study. J Manipulative Physiol Ther. 2007;30(6):408-18. Epub 2007/08/19. doi: 10.1016/j.jmpt.2007.04.013. PubMed PMID: 17693331.

25. Senstad O, Leboeuf-Yde C, Borchgrevink C. Frequency and characteristics of side effects of spinal manipulative therapy. Spine (Phila Pa 1976). 1997;22(4):435-40; discussion 40-1. PubMed PMID: 9055373.
26. Cagnie B, Vinck E, Beernaert A, Cambier D. How common are side effects of spinal manipulation and can these side effects be predicted? Man Ther. 2004;9(3):151-6. doi: 10.1016/j.math.2004.03.001. PubMed PMID: 15245709.
27. Hoving JL, Koes BW, de Vet HC, van der Windt DA, Assendelft WJ, van Mameren H, et al. Manual therapy, physical therapy, or continued care by a general practitioner for patients with neck pain. A randomized, controlled trial. Ann Intern Med. 2002;136(10):713-22. PubMed PMID: 12020139.
28. Bronfort G, Evans R, Nelson B, Aker PD, Goldsmith CH, Vernon H. A randomized clinical trial of exercise and spinal manipulation for patients with chronic neck pain. Spine (Phila Pa 1976). 2001;26(7):788-97; discussion 98-9. Epub 2001/04/11. doi: 10.1097/00007632-200104010-00020. PubMed PMID: 11295901.

Chapter 5.

Smallest Detectable Change and Minimal Important Change of the Neck Disability Index were influenced by population characteristics

Wouter Schuller, MD, Raymond W.J.G. Ostelo PhD, Richard Janssen MD, Henrica C.W. de Vet PhD

Health and Quality of Life Outcomes (2014);12:53

Abstract

Objective: To assess the influence of the type of anchor, the definition of improvement, and population characteristics on the smallest detectable change (SDC) and the minimal important change (MIC) of the Neck Disability Index (NDI).

Study design and setting: A cohort study including 101 patients with chronic neck pain. SDC and MIC were calculated using two types of external anchors. For each anchor we applied two different definitions to dichotomise the population in a group of improved and a group of unimproved patients. The influence of patient characteristics was assessed in relevant subgroups: patients with or without radiating pain, patients with or without concomitant headache and patients with high or low baseline scores.

Results: Different anchors and different definitions of improvement hardly influenced estimates of the SDC and the MIC. The SDC and the MIC were similar for subgroups of patients with or without radiation, but differed strongly for subgroups of patients with or without concomitant headache and for patients with high or low baseline scores.

Conclusions: The SDC and the MIC are not an invariable characteristic of the NDI but are influenced by patient characteristics.

Introduction

The Neck Disability Index (NDI) was published by Vernon in 1991 as a patient reported outcome measure of disability in patients with neck pain [1]. It has been reported to be the most commonly used self-report instrument for evaluating functional status in neck pain clinical research [2, 3]. A review published in 2008 stated that the NDI had been used in approximately 300 publications, and translated into 22 languages [4]. Many studies have addressed the measurement properties of the NDI. A systematic review of these measurement properties by MacDermid et. al. identified 3 comprehensive review articles and 41 studies that addressed at least 1 psychometric property [2]. They concluded that the NDI is reliable, valid and responsive in numerous patient populations, including patients with acute and chronic conditions, as well as those suffering from neck pain associated from musculoskeletal dysfunction, whiplash-associated disorders, and cervical radiculopathy. The authors followed Vernon in suggesting an accepted minimal important change (MIC) of 5, but stated that the work on minimal detectable change and MIC is sparse and inconsistent [2]. In this paper we will use the term smallest detectable change (SDC) instead of minimal detectable change. A recently published systematic review by Schellingerhout et. al. used the COSMIN checklist [5] to select articles based upon their methodological quality. The authors concluded that the NDI shows positive results for internal consistency, content validity, structural validity, hypothesis testing and responsiveness, but a negative result for reliability. According to this study a value for the MIC cannot be provided yet, as the estimates for the MIC are too diverse [6]. More studies are needed that determine the SDC and the MIC in different subgroups of patients.

To enable a proper interpretation of change scores on a measurement instrument the SDC and the MIC are considered the most important benchmarks. Of all studies reporting measurement properties, only a few studies presented estimates of the SDC or the MIC [7-13]. An overview of these studies is given in Table 1. In these studies different patient populations were recruited, and different follow-up periods and different definitions of improvement on the anchor were used. Two studies recruited patients with cervical radiculopathy [7, 13], one study recruited patients with acute pain [14], and three studies recruited a mixed population of patients with acute and chronic pain [8-10]. One study did not describe the characteristics of their neck pain patients [12]. The follow-up period ranged from one week [10, 14] to seven weeks [9]. Clearly the studies were rather heterogeneous. The reported MIC values vary from 3.5 to 9.5, and the SDC values vary from 3.0 to 17.9, raising questions about the generalisability of these parameters. Apparently different study designs or different study populations lead to different estimates of the SDC and the MIC for the NDI, raising the challenging question whether it is at all possible to adopt one general value for the SDC

and the MIC. Or should we use different estimates of the SDC and the MIC for different populations?

In addition to different patient populations and different follow-up periods previous studies used different designs and methods, various anchors to define important change, and different definitions of improvement on the anchor. They report different estimates of the SDC and the MIC, but we do not know if this originates from sampling bias only, or is caused by the different designs and methods used. To assess which of these sources could have influenced the estimates of the SDC and the MIC one should compare different ways of calculating these parameters on different subgroups of patients in a single study. We studied this influence in a single population of patients with chronic neck pain, enabling us to assess and compare the influence of the type of anchor, the definition of improvement and of population characteristics on the SDC and the MIC of the NDI.

Methods

Design

From March to October 2009 patients with chronic neck pain were recruited in 4 practices for Musculoskeletal (MSK) Medicine in the Netherlands. Inclusion criteria were chronic neck pain, defined as neck pain existing at least 3 months, aged 18 years or older, and having no contraindications for manipulative treatment. Duration of complaints, age, radiation into the arm(s), and the presence of concomitant headache were recorded. After signing an informed consent form patients filled in the NDI (T0). Patients received treatments by one of six MSK physicians. After a follow-up period of 6 months patients were asked by email to fill in a NDI questionnaire again together with questions about their global perceived effect on pain and on function (T1).

Measurement instruments

The NDI is generally considered to be unidimensional, and contains 10 items. Seven items are related to activities of daily living, two are related to pain, and one item is related to concentration. Each item is scored on a 0-5 scale, adding up to an overall score ranging from 0 to 50, with higher scores corresponding to more severe disability. Global perceived effect (GPE) was used as the external criterion and was phrased to question either change of pain or change of function, with the following possible scores: completely recovered (6), much improved (5), slightly improved (4), no change (3), slightly worse (2), or much worse (1).

Table 1: Results from other studies

Study (N)	Population	NDI (T0)	GPE	Stable	SDC	MIC (sens/spec)
Cleland 2008[8]	N=137, NP acute + chronic	32.2-35.7	15 pt	-3 to +3	11.6*	9.5 (0.83/0.72)
Young I 2010[13]	N=165, CR av. 4 weeks	24-25	13/15 pt	-1 to +1	17.9*	8,5 (0.62/0.79)
Young B 2009[12]	N=91, NP unclear	15.8-18.0	15 pt	-2 to +2	12.1*	7.5 (n.r.)
Cleland 2008[8]	N=137, NP acute + chronic	32.2-35.7	15 pt	-2 to +2	8.1	7.0
Cleland 2006[7]	N=38, NP+CR, av. 2 weeks	21.9-20.7	15 pt	-3 to +3	12.0	7.0 (0.52/0.59)
Young I 2010[13]	N=165, CR av. 4 weeks	24-25	13/15 pt	-2 to +2	15.9*	6.5 (0.57/0.67)
Pool 2007[9]	N=183, NP acute + chronic	14.5	6 pt	2 to 4	10.5	3.5 (0.90/0.70)
Vos 2006[11]	N=187, NP acute	13.0-16.0	7 pt	3 to 5	7.62	n.r.
Trouli 2008[10]	N=65, NP acute + chronic	n.r.	15 pt	-3 to +3	3.03	n.r.

NP = neck pain, CR = cervical radiculopathy, n.r. = not reported
Vos et. al. reported a different SDC of 1.66 in a published paper [14]. The correct SDC reported in his dissertation [11] is presented;
Trouli et. al. reported a SDC of 1.78 but this value does not agree with the Bland & Altman plot presented. We present a SDC estimated from the Bland & Altman plot;
Cleland 2008 and Young 2010 reported two different analyses, these analyses are represented in the table as separate studies;
*reported SDC(90) was changed to SDC(95) (SDC(95) = 1.96 X SDC(90)/ 1.65)

Type of anchor, definition of improvement and choice of clinical subgroups

To assess the influence of the type of anchor we used two external anchors: one anchor phrased to question change of *pain* (GPE_{pain})and one anchor phrased to question change of *function* ($GPE_{function}$). In addition we used two different definitions of improvement. Firstly we defined both completely recovered and much improved (GPE 1-2) as improved, whilst defining slightly worse, no change and slightly better (GPE 3-5) as unchanged. In these calculations patients reporting much worse were excluded. Secondly we defined patients reporting completely recovered, much improved, and slightly improved (GPE 1-3) as improved, only categorising patients reporting no change (GPE 4) as unchanged. In these calculations patients reporting slightly worse or much worse were excluded. This enabled us to analyse four different situations:

a. GPE-pain, unchanged = GPE 3-5
b. GPE-pain, unchanged = GPE 4
c. GPE-function, unchanged = GPE 3-5
d. GPE-function, unchanged = GPE 4

We present means and standard deviations of the NDI scores at T0, of the NDI scores at T1, and of the changes in NDI score between T0 and T1 for both external anchors and for the groups of improved and stable patients. For each situation we calculated the SDC and the MIC.

To assess the influence of different clinical characteristics we considered the following subgroups of patients:

a. Patients with or without radiation
b. Patients with or without concomitant headache
c. Patients with baseline scores of above or below the median (NDI = 24)

Analyses

- The smallest detectable change (SDC) was based on the standard error of measurement (SEM) which was derived from the variance component in the formula for $ICC_{agreement}$ [15]. It was calculated on the group of patients who were considered to be unchanged by $1.96 \times \sqrt{2} \times SEM_{agreement}$ [16-18].
- MIC was calculated using a ROC curve to establish the optimal cut-of point distinguishing the groups of patients dichotomised according to the anchor. We report the MIC and the sensitivity and specificity at the cut-of point [19].

For each subgroup we calculated the SDC and the MIC. In the clinical subgroup analysis we also chose to use a GPE score of 3-5 to define stable patients, increasing the number of patients in the subgroups of unchanged patients. Patients reporting much worse were again excluded from these calculations.

Results

Overall results

A total of 101 patients were recruited and gave informed consent, of whom 99 patients completed the follow up measurement. Patients characteristics are presented in Table 2. The mean age at inclusion was 42 years (SD 12). Contrary to our inclusion criteria we included 1 patient who had neck pain of 2 months duration, but the average duration of complaints was 77 months (range 2-480, median 24 months). 54 Patients reported the pain to radiate into the arm(s), 78 patients reported concomitant headache. For both external anchors, and for each definition of improvement, the mean scores and the standard deviations at T0, at T1, and of the change scores are presented in Table 3. A clear trend can be seen of an increasing change of NDI score in accordance with GPE scores. Spearman correlation coefficients between the NDI change score and the GPE for pain and the GPE for function were 0.60 and 0.58 respectively.

Table 2: Patient characteristics at inclusion (N=101).

Mean age at inclusion (range)	42 (19-71)
Mean duration of complaints in months (range)	77 (2-480)
Mean NDI score at baseline (SD)	24.4 (6.2)
Radiating pain	54%
Headache	78%

Table 3: Mean scores and standard deviation of the NDI at baseline and after 6 months for different scores of the GPE for pain and the GPE for function respectively (N=99).

GPE pain	NDI:T0, mean (sd)	NDI:T1, mean (sd)	NDI:T1-T0, mean (sd)
1 Completely recovered(N=14)	22.1 (3.1)	11.3 (1.9)	-10.9 (3.6)
2 Much improved (N=41)	24.0 (6.8)	16.5 (3.3)	-7.5 (6.0)
3 Slightly improved (N=17)	25.1 (5.4)	22.0 (4.8)	-3.1 (4.0)
4 Unchanged (N=23)	25.6 (7.1)	24.6 (9.1)	-1.0 (5.9)
5 Slightly worse (N=2)	28.0 (7.1)	32.0 (1.5)	4.0 (8.5)
6 Much worse (N=2)	27.0 (0.0)	27.0 (9.9)	0.0 (9.9)
GPE 1-3 (N=72)	23.9 (6.0)	16.8 (5.0)	-7.1 (5.8)
GPE 4 (N=23)	25.6 (7.1)	24.6 (9.1)	-1.0 (5.9)
GPE 5-6 (N=4)	27.5 (4.1)	29.5 (6.5)	2.0 (7.8)
GPE 1-2 (N=55)	23.5 (6.1)	15.6 (3.8)	-8.4 (5.7)
GPE 3-5 (N=42)	25.5 (6.3)	23.9 (7.6)	-1.6 (5.4)
GPE 6 (N=2)	27.0 (0.0)	27.0 (9.9)	0.0 (9.9)
GPE function	**NDI:T0, mean (sd)**	**NDI:T1, mean (sd)**	**NDI:T1-T0, mean (sd)**
1. Completely recovered (N=19)	22.7 (6.5)	12.3 (2.6)	-10.5 (6.9)
2. Much improved (N=34)	24.3 (6.2)	16.9 (3.4)	-7.5 (4.7)
3. Slightly improved (N=17)	25.0 (5.5)	22.1 (4.5)	-3.0 (3.4)
4. Unchanged (N=26)	25.0 (6.8)	23.5 (9.1)	-1.6 (5.9)
5. Slightly worse (N=2)	25.0 (2.8)	33.5 (0.7)	8.5 (2.1)
6. Much worse (N=1)	32.0	32.0	
GPE 1-3 (N=70)	24.1 (6.0)	16.9 (4.9)	-7.2 (5.7)
GPE 4 (N=26)	25.0 (6.8)	23.5 (9.1)	-1.6 (5.9)
GPE 5-6 (N=3)	27.3 (4.5)	33.0 (1.0)	5.7 (5.1)
GPE 1-2 (N=53)	23.8 (6.3)	15.2 (3.8)	-8.6 (5.7)
GPE 3-5 (N=45)	25.0 (6.1)	23.4 (7.8)	-1.6 (5.4)
GPE 6 (N=1)	32.0	32.0′	0.0

Influence of the type of anchor and of the definition of improvement

Table 4 presents the SDC and the MIC (including its sensitivity and specificity). The estimates in the four different situations reveal no large differences. The SDC ranges from 10.6 to 11.4, the MIC is 2.5, independent whether the anchor was based on pain or function, and independent of the way in which important change was defined.

Table 4: SDC and MIC with sensitivity and specificity for the NDI using separate anchors for pain and for function and using two different ways to dichotomise GPE scores (N=99).

GPE	SDC	MIC	sensitivity	specificity
Pain:				
Unchanged= GPE 4 (N=23)	11.5	2.5	0.739	0.806
Unchanged= GPE 3-5 (N=42)	11.0	2.5	0.690	0.927
Function:				
Unchanged= GPE 4 (N=26)	11.8	2.5	0.731	0.829
Unchanged= GPE 3-5 (N=45)	11.0	2.5	0.644	0.925

Influence of clinical characteristics

Analyses for subgroups of patients are presented in Table 5. We chose to carry out these analyses using the external anchor questioning change of pain only, because we found hardly any difference in our estimates between the differently phrased external anchors, and the GPE questioning improvement of pain is frequently used in other studies. For patients with or without radiation the SDC and the MIC were similar (11.0 and 2.5 respectively). For patients with or without concomitant headache the SDC and the MIC were different. Without concomitant headache the SDC was 3.4, with a MIC of 3.5. With concomitant headache the SDC was 11.6, with a MIC of 2.5. The SDC and the MIC were also different for patients with baseline NDI scores above or below 24. With a baseline score < 24 the SDC was 5.1, with a MIC of 2.5. With a baseline score ≥ 24 the SDC was 13.0, with a MIC of 4.0.

Table 5: Clinical subgroup analysis of SDC and MIC (with sensitivity and specificity) for the NDI. GPE on pain, 3-5 = unchanged.

Subgroups	SDC	MIC	sensitivity	specificity
Radiation				
No radiation (N=47)	11.0	2.5	0.750	0.889
Radiation (N=54)	11.0	2.5	0.654	0.963
Headache				
No headache (N=22)	3.1	3.5	1.000	0.812
Headache (N=79)	11.8	2.5	0.667	0.921
Baseline score				
Baseline < 24 (N=49)	6.8	2.5	0.882	0.871
Baseline ≥24 (N=52)	13.0	4.0	0.600	1.000

Discussion

Our results show that using different types of anchors or applying different definitions of improvement hardly influenced estimates of the SDC and the MIC. Estimates of the SDC and the MIC were similar for patients with or without radiation, but differed for patients with or without concomitant headache and for patients with different baseline scores. Especially patients with higher baseline scores have a higher MIC, where the different MIC's in subgroups of patients with or without headache is less outspoken. One could postulate that the different estimates in the subgroup of patients with headache could also be caused by higher baseline scores, but while the subgroup of patients with headache does have higher baseline scores (25.5 versus 20.4) the estimated MIC is in fact higher in the subgroup without headache. On the other hand the number of patients without headache is rather limited, and a small difference in MIC could be due to chance. The SDC is similar in almost all analyses but varies strongly in patients with or without headache and in patients with higher or lower baseline scores. It can be concluded that the different estimates of SDC and MIC in our study are predominantly explained by patient characteristics.

Could differences in population characteristics alone explain the different estimates in previous studies? The results of these studies are presented in Table 1, ranked according to the estimated MIC. Clearly the study with the highest baseline NDI does have the highest MIC, but in this study the MIC was reduced to 7 with a narrower definition of unchanged patients. Studies recruiting patients with cervical radiculopathy regardless of the coexistence of neck pain have the highest estimates of the SDC, but patients with cervical radiculopathy not necessarily have neck pain too. The perceived effect in these patients could be related to arm symptoms, while the NDI specifically measures neck symptoms. This could reduce the correlation between the NDI score and the GPE, increasing the variance of the change scores in the group of stable patients, subsequently leading to higher estimates of the SDC while decreasing the sensitivity and the specificity of the MIC. In our study we did not observe any difference in SDC between the subgroup of patients with or without radiating pain, but this could be explained by the fact that we recruited a population of patients with neck pain, and we did not specifically screen for radiculopathy. Judged by the high SDC's and the low sensitivity and specificity of the MIC it does seem that the NDI is less useful in populations with cervical radiculopathy.

Due to the different populations recruited and the different methods used it remains very difficult to compare the estimates from previous studies. In our view different patient populations indeed seem to lead to different estimates of the SDC and of the MIC, but it is unclear whether patient characteristics are the only source explaining these differences. It

is clear that the SDC is much higher than the MIC in most studies, ranging even to 17.9 in a population with cervical radiculopathy. Given this high SDC one needs a change score much higher than the MIC to reliably label a patient as improved. This raises questions about the usefulness of the NDI to assess change in individual patients [15].

The different estimates of the MIC for patients with high baseline scores could be explained by our methods. We calculate the MIC by comparing the change score with the global perceived effect. This global perceived effect has been reported to correlate stronger with present status than with change in status [20]. A patient with severe disability needs a large improvement to arrive at a better present status after treatment and could still end up in the group of patients reporting to be unchanged even with a strong improvement of the NDI score. A patient with a low baseline score but no real change after treatment will still have a good present status at follow-up and could end up in the group of improved patients while the NDI change score is small. This could explain higher estimates of the MIC for patients with higher baseline scores. This does not necessarily reflect a real need for a larger improvement, but could be a shortcoming of our way of calculating the MIC using a global perceived effect as external anchor.

Strengths and limitations

The strength of our study lies in the use of a single study to assess the influence of different anchors, different definitions of improvement on the anchor and of population characteristics on estimates of the SDC and the MIC. Using differently phrased anchors for pain and for function gave us the opportunity to study the influence of the phrasing of the anchor in estimating the SDC and the MIC for the NDI in the same population, thereby excluding the possibility of sampling bias. Although interesting we did not study the influence of follow-up time. A possible weakness lies in the number of recruited patients. This number of patients was enough for the main analyses, but may have been relatively small for our subgroup analyses. Especially the subgroup of patients without headache was relatively small. Criticisms regarding the use of a global perceived effect are described above, but in the absence of a better alternative this still seems to be the best available method.

Further study

In terms of methodology future studies may focus on alternatives for the GPE. It is quite understandable that different patients have different perceptions of what magnitude of effect they consider an important change, perhaps also depending upon the treatment administered. A treatment that is costly, painful, or strenuous might need a larger effect to be considered worthwhile, and a patient who has experienced severe side effects might even consider a large improvement not worthwhile. The development of other methods to

estimate sufficient important change could lead to new perspectives. We could still make progress in defining clinical relevance [21, 22].

Conclusions

In our study we have shown that estimation of the SDC and the MIC of the NDI can be influenced by population characteristics. This means that one cannot label a single change score of the NDI as an important change for patients. A serious drawback of the NDI is the high SDC. One needs quite a large change score to reliably label a patient as improved.

Acknowledgements and financial support
This study was originally conducted as a scientific paper in the specialist training for MSK physician. We thank the MSK physicians Brouwer, Cossee, Cuppen, Jonquiere, and Savelkouls for recruiting patients for this study. The first authors position at the EMGO+ Institute for Health and Care Research is funded by the Dutch Association for Musculoskeletal Medicine (Nederlandse Vereniging voor Artsen Musculoskeletale Geneeskunde, NVAMG).

References

1. Venon H, Mior S. The Neck Disability Index: a study of reliability and validity. J Manipulative Physiol Ther. 1991;14(7):409-15.
2. MacDermid JC, Walton DM, Avery S, Blanchard A, Etruw E, McAlpine C, et al. Measurement properties of the neck disability index: a systematic review. J Orthop Sports Phys Ther. 2009;39(5):400-17.
3. Pietrobon R, Coeytaux RR, Carey TS, Richardson WJ, DeVellis RF. Standard scales for measurement of functional outcome for cervical pain or dysfunction: a systematic review. Spine (Phila Pa 1976). 2002;27(5):515-22.
4. Vernon H. The Neck Disability Index: state-of-the-art, 1991-2008. J Manipulative Physiol Ther. 2008;31(7):491-502.
5. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Qual Life Res. 2010;19(4):539-49.
6. Schellingerhout JM, Verhagen AP, Heymans MW, Koes BW, de Vet HC, Terwee CB. Measurement properties of disease-specific questionnaires in patients with neck pain: a systematic review. Qual Life Res. 2012;21(4):659-70.
7. Cleland JA, Fritz JM, Whitman JM, Palmer JA. The reliability and construct validity of the Neck Disability Index and patient specific functional scale in patients with cervical radiculopathy. Spine (Phila Pa 1976). 2006;31(5):598-602.

8. Cleland JA, Childs JD, Whitman JM. Psychometric properties of the Neck Disability Index and Numeric Pain Rating Scale in patients with mechanical neck pain. Arch Phys Med Rehabil. 2008;89(1):69-74.
9. Pool JJ, Ostelo RW, Hoving JL, Bouter LM, de Vet HC. Minimal clinically important change of the Neck Disability Index and the Numerical Rating Scale for patients with neck pain. Spine (Phila Pa 1976). 2007;32(26):3047-51.
10. Trouli MN, Vernon HT, Kakavelakis KN, Antonopoulou MD, Paganas AN, Lionis CD. Translation of the Neck Disability Index and validation of the Greek version in a sample of neck pain patients. BMC Musculoskelet Disord. 2008;9:106.
11. Vos CJ. Acute Neck Pain in General Practice: Erasmus University Rotterdam; 2006.
12. Young BA, Walker MJ, Strunce JB, Boyles RE, Whitman JM, Childs JD. Responsiveness of the Neck Disability Index in patients with mechanical neck disorders. Spine J. 2009;9(10):802-8.
13. Young IA, Cleland JA, Michener LA, Brown C. Reliability, construct validity, and responsiveness of the neck disability index, patient-specific functional scale, and numeric pain rating scale in patients with cervical radiculopathy. Am J Phys Med Rehabil. 2010;89(10):831-9.
14. Vos CJ, Verhagen AP, Koes BW. Reliability and responsiveness of the Dutch version of the Neck Disability Index in patients with acute neck pain in general practice. Eur Spine J. 2006;15(11):1729-36.
15. de Vet HC, Terwee CB, Mokkink LB, Knol D. Measurement in Medicine. Cambridge: Cambridge University Press; 2011.
16. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986;1(8476):307-10.
17. de Vet HC, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. J Clin Epidemiol. 2006;59(10):1033-9.
18. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol. 2007;60(1):34-42.
19. van der Roer N, Ostelo RW, Bekkering GE, van Tulder MW, de Vet HC. Minimal clinically important change for pain intensity, functional status, and general health status in patients with nonspecific low back pain. Spine. 2006;31(5):578-82.
20. Kamper SJ, Ostelo RW, Knol DL, Maher CG, de Vet HC, Hancock MJ. Global Perceived Effect scales provided reliable assessments of health transition in people with musculoskeletal disorders, but ratings are strongly influenced by current status. J Clin Epidemiol. 2010;63(7):760-6.
21. Barrett B, Brown D, Mundt M, Brown R. Sufficiently important difference: expanding the framework of clinical significance. Med Decis Making. 2005;25(3):250-61.
22. Ferreira ML, Herbert RD, Ferreira PH, Latimer J, Ostelo RW, Nascimento DP, et al. A critical review of methods used to determine the smallest worthwhile effect of interventions for low back pain. J Clin Epidemiol. 2012;65(3):253-61.

Chapter 6.

Measurement properties of the Dutch-Flemish PROMIS Pain Behaviour item bank in patients with musculoskeletal complaints

Wouter Schuller, Caroline B. Terwee, Thomas Klausch, Leo D. Roorda, Daphne C. Rohrich, Raymond W. Ostelo, Berend Terluin, Henrica C.W. de Vet

Journal of Pain. 2019;20(11):1328-1337

Abstract

We studied the measurement properties of the 39 item v1.1 Dutch-Flemish PROMIS Pain Behaviour item bank in a sample of 1602 patients with musculoskeletal complaints. We evaluated the assumptions of the underlying Item Response Theory (IRT) model (unidimensionality and local dependency with Confirmatory Factor Analyses (CFA), and monotonicity with scalability coefficients). We studied IRT-model fit of all items, and estimated the item parameters of the IRT model. Differential Item Functioning (DIF) was studied for age and gender, and DIF for language was studied as a measure of cross-cultural validity. CFA showed suboptimal fit of a unidimensional model, but a bi-factor model showed low risk of bias when a unidimensional model was assumed (Omega-H 0.92, Explained Common Variance (ECV) 0.70). Fifteen item pairs (2%) were locally dependent. Five items showed poor scalability. All items fitted the IRT model; slope parameters ranged from 0.60 to 2.00, and threshold parameters from -2.05 to 6.80. One item showed DIF for age, one item DIF for gender, and five items showed DIF for language, but the impact on total scores was low. Our study supports the psychometric properties of the Dutch-Flemish PROMIS Pain Behaviour item bank, although problems with dimensionality and monotonicity need further evaluation.

Introduction

Pain behaviours are external manifestations of experiencing pain, such as sighing or crying, verbal reports of pain, and also include pain severity behaviours such as resting, guarding, facial expressions, and asking for help. Pain behaviours have been associated with pain intensity(1), disability(1, 2), depression(3), and with the development of chronic pain(2). Interest in how to measure pain behaviours in patients with pain has been growing in order to identify subgroups of patients that might benefit from tailored interventions(4, 5), or as a possible target for treatment in itself(6). Traditionally pain behaviours have been evaluated with a patient diary, or by an external observer(4); however, both methods are complicated and time consuming. A valid and reliable self-report tool would make the evaluation of pain behaviour more feasible for routine clinical practice and for clinical research.

A self-report tool of pain behaviour has been developed by the Patient-Reported Outcomes Measurement Information System (PROMIS) initiative(7). The PROMIS initiative uses Item Response Theory (IRT) to construct item banks consisting of a large collection of questions (i.e. items) covering a wide range of a given trait (i.e. construct), such as pain behaviour. All items in the item bank are ordered (i.e. calibrated) on a single scale in a large population representing a wide range of the pain behaviour trait. After calibration subsets of items can be used, either in short form, or in Computerised Adaptive Testing (CAT). In CAT, a computer algorithm administers items one by one. After each answer the computer decides on the basis of previous answers which next item would be most informative to ask. Items are thus tailored to the individual patient and only a small number of items is needed to obtain a reliable score(8-12).

The PROMIS Pain Behaviour item bank v1.0 (later updated to v1.1) was developed by Revicki et al.(7), and calibrated in a US population including a community sample and a clinical sample. The item bank contains 39 items, demonstrating coverage of a wide range of the pain behaviour construct. Further validation studies are needed to evaluate how this item bank functions in other samples, by studying different patient populations, and translated versions in other languages. The validity of the Dutch-Flemish translation of the PROMIS v1.1 Pain Behaviour item bank has previously been studied in a population (N=1140) of chronic pain patients in an outpatient rehabilitation setting(13), showing high reliability, and sufficient cross-cultural validity and construct validity in this population. The aim of the present study was to further evaluate the psychometric properties of the Dutch-Flemish translation of the PROMIS Pain Behaviour item bank and to evaluate cross-cultural validity in a new population of patients presenting with musculoskeletal complaints in primary care practices.

Methods

Study design and procedure

We conducted a cross-sectional study using an existing web-based registry of patients presenting for the first time in musculoskeletal practice. A group of 31 musculoskeletal (MSK) physicians in The Netherlands was recruited to register patient characteristics in a web-based registry. Most MSK practices are primary care facilities, focused on patients with musculoskeletal complaints. During the first visit the physician entered the following patient characteristics in the registry: age, gender, type and duration of the main complaint and the existence of concomitant complaints. Complaints were coded by the treating physician according to the International Classification of Primary Care (ICPC). Physicians were instructed to ask all consecutive patients who were presented for a first consultation to participate in the study. Following an informed consent procedure, the physician entered email addresses of the recruited patients in the registry. Thereafter, a specially designed computer program (Readmail) automatically distributed invitations to patients by email to fill in web-based questionnaires. From October 2013 until February 2014 this registry was used for the present study. To evaluate cross cultural validity we used data from part of the sample included in the original US calibration study. This part of the US calibration sample consisted of 967 patients who were recruited through the website of the American Chronic Pain Association (ACPA), and who had at least one chronic pain condition for at least three months prior to participating in the survey(7).

Measures

Translation of the PROMIS Pain Behaviour item bank into Dutch-Flemish was carried out by FACITtrans according to standard PROMIS methodology and approved by the PROMIS Statistical Center(14). Our study population responded to the full PROMIS Pain Behaviour item bank v1.1, containing 39 items. For each item patients rated how frequently they expressed the given pain behaviours in the past 7 days, using a six point Likert scale with the following categories: 1=Had no pain, 2=Never, 3=Rarely, 4=Sometimes, 5=Often and 6=Always. PROMIS scores were expressed as T-scores, where a T-score of 50 represents the average score of the general US population, with a standard deviation (SD) of 10. Higher scores represent higher trait levels, i.e. more pain behaviour in the case of the Pain Behaviour item bank.

Ethics

This study was approved by the Medical Ethical Committee of the VU Medical Center (2013/20). This Medical Ethical Committee decided that our observational study did not require the strict procedure for written and signed informed consent based on the law for Scientific Medical Research (WMO). Nonetheless, verbal informed consent was obtained from all patients in this study, which was recorded by the treating physician.

Statistical analyses

Demographic analyses were carried out using SPSS statistics, version 22.

Psychometric properties

The psychometric properties of the Dutch-Flemish PROMIS Pain Behaviour item bank were studied in accordance with the PROMIS analyses plan(15). These analyses were similar to the ones used in the development of the PROMIS Pain Behaviour item bank(7). The following psychometric properties were studied: IRT-model assumptions and fit, measurement invariance, and cross-cultural validity. Following the analyses by Revicki et al.(7) patients reporting no pain on one or more of the Pain Behaviour items were excluded from the analysis, and response categories with less than five respondents were collapsed. Table 1 provides a detailed overview of the psychometric properties that were studied, the analyses, statistical parameters, criteria for acceptable values, and software packages used.

IRT-model assumptions and fit

PROMIS item banks have been developed using IRT methods, and the estimates of patient scores are based upon the underlying IRT-model. The IRT-model estimates the IRT-parameters, which are used to examine the quality of the items, the coverage of the item banks and to calculate the patient scores. Consequently, meeting the assumptions of the IRT-model and adequate fit of this model supports the validity of the estimated patient scores. IRT assumptions are unidimensionality and monotonicity. Unidimensionality means that only one construct is measured. If the assumption of unidimensionality is not met, then it is questionable whether items can be calibrated on a single scale. Monotonicity is a measure of scalability, which relies on the assumption that the probability for patients to select response categories corresponds with their level of impairment. We tested unidimensionality using a one-factor and a bi-factor model with Confirmatory Factor Analyses (CFA), and by assessing local dependency. We tested monotonicity using non-parametric Mokken scaling. We calculated T-scores using the US calibration parameters.

Table 1; Psychometric properties studied; type of analyses; outcomes; criteria; software or R-packages used.

IRT-model assumptions and fit					
Property	*Analysis*	*Description of analysis*	*Indices*	*Criteria*	*Software*
Dimensionality	One factor model	A Graded Response Model (GRM) is used for item banks that include items that have several response categories. For the CFA, a Probit-link GRM with a Weighted Least Squares Mean and Variance adjusted estimator (WLSMV)(17) was used. Slope parameters represent the discriminative ability of each item. Threshold parameters locate each response category on the scale of the measured trait.	Scaled Comparative Fit Index (CFI) Scaled Tucker Lewis Index (TLI) Scaled Root Mean Square Error of Approximation (RMSEA)(18)	>0.95 >0.95 <0.06	R-package Lavaan Version 0.5-23.1097
	Bi-factor model(19)	In a bi-factor model, all items are considered to load on one general factor and one or more secondary factors. These secondary factors are groups of items that explain part of the residual variance. In an iterative process, item pairs with the highest Standardised Expected Parameter Change (SEPC) were selected for group factors until all SEPC's were below 0.3, or until no further improvement of model fit could be achieved(20).			
	Local Dependency	Local independency implies that most variance between the items is explained by the underlying construct. If item pairs test positive for local dependence this implies that part of the variance is caused by other factors. To test for local dependence the matrix of residual correlations of both the one factor and the bi-factor CFA is studied.	Correlation coefficient	>0.2	
	Risk of bias	Indices indicate the relative strength of the general factor(21) in the bi-factor model.	Explained Common Variance (ECV)(22)	>0.60	
			Omega-H(22)	>0.80	

IRT-model assumptions and fit					
Property	*Analysis*	*Description of analysis*	*Indices*	*Criteria*	*Software*
Monotonicity	Mokken scaling(23)	Monotonicity implies that the probability of choosing a higher response category for an item increases with increasing pain behaviour. The strength of the relationship of each item with the latent trait is tested in a non-parametric IRT model(23, 24).	Scale - Scalability coefficient H Items - Scalability coefficient H_i	Unscalable <0.3 Weak 0.3-0.4 Moderate 0.4-0.5 Strong >0.5	R-package Mokken
Item fit	IRT modelling	A logistic GRM using the Bock-Aitkin maximal likelihood method(16) is used for individual item fit. The S-X^2 the statistic compares the observed and expected response frequencies under the estimated IRT model.	S-X^2 and p-value	p> 0.001	IRTPRO®
Measurement invariance					
Comparing subgroups with different age or gender	Differential Item Functioning	Ordinal Logistic Regression, comparing item functioning between groups of patients(25, 26).	Change in Mcfadden R^2	>0.02	R-package Lordif Version 0.3-3
Cross cultural validity					
Comparing subgroups with different language	Differential Item Functioning	Ordinal Logistic Regression, comparing item functioning between groups of patients(25, 26)	Change in Mcfadden R^2	>0.02	R-package Lordif Version 0.3-3

Measurement invariance

Measurement invariance addresses the question of whether or not the scores between subgroups can be compared. We studied measurement invariance by evaluating Differential Item Functioning (DIF), analysing whether the IRT parameters are equivalent for subgroups of patients. If the IRT parameters are equivalent in different subgroups of patients, e.g. subgroups of patients with musculoskeletal complaints who differ with respect to age or gender, then their respective scores are comparable. If IRT parameters are not equivalent, differences in subgroup scores may result from differences in the interpretation of the items, and may not reflect "real" differences.There are two types of DIF: uniform or non-uniform. DIF is considered to be uniform when the magnitude of the DIF is similar for all trait levels. DIF is considered to be non-uniform when the magnitude varies for different trait levels. We evaluated DIF for several age groups (median split, below or over 30, below or over 60), and for gender (male vs. female).

Cross-cultural validity

Cross-cultural validity addresses the question whether or not scores between cultural or language groups can be compared meaningfully. Within IRT, cross-cultural validity for language can be studied by assessing DIF between comparable populations from different language groups(16). We evaluated DIF for language by comparing our data with the data available from the US-ACPA sample used by Revicki et al.(7). In the US-ACPA sample, patients responded to 31 of the 39 items from the PROMIS pain behaviour item bank. These patients did not respond to the items PAINBE32, PAINBE34, PAINBE38, PAINBE40, PAINBE41, PAINBE46, PAINBE47, and PAINBE48, DIF for language was thus evaluated for 31 out of the 39 items.

Results

Demographic characteristics

A total of 2610 patients were asked to participate in the study and 2171 consented. Of these 2171 patients 1745 (67%) completed the questionnaires. Because only the year of birth was reported we excluded patients who could have been under the age of 18 at inclusion. After removal of patients reporting no pain, patients under 18, and patients who had failed to complete the whole item bank a sample of 1602 patients remained for the analyses. Another 7 patients had a small number of missing items. Model analyses were conducted on the whole population of 1602 patients. T-scores were calculated for the 1595 patients who had answered all items. Demographic data of our sample are presented in Table 2. Half of the patients presented with a primary complaint of low back pain (51.2%), with or without sciatica, followed by neck or shoulder pain (20.7%).

Table 2; Demographics of patient sample: sex, age, type and duration of main complaints, and T-scores. Comparison with US-ACPA sample.

	SMT sample (N=1602)	US ACPA sample (N=967)[a]
Demographic data		
Age, average (range)	47 (19-91)	48 (21-86)
Gender (% female)	59	81
Duration of complaints %		
< 3 months	19	
3 months- 1 year	25	6
> 1 year	56	91
Type of complaints (%)		
Low back pain	813 (51.2)	533 (55)
Neck or shoulder pain	332 (20.7)	447 (46)
Other back pain	130 (8.1)	
Lower extremity	154 (9.6)	
Upper extremity	35 (2.2)	
Headache	51 (3.2)	209 (22)
Rheumatoid arthritis		59 (6)
Osteoarthritis		195 (20)
Pain related to cancer		8 (0.8)
Fibromyalgia		338 (35)
Chronic widespread pain		
Other neuropathic pain		370 (38)
Other	84 (5.1)	298 (31)
T-scores		
T-score (SD)	50.2 (10.4)	63.7 (3.5)
T-score range	21.6-79.5	54.0-78.6

[a] In our study only the main complaint could be scored, while in the US-ACPA study multiple complaints could be indicated.

Psychometric properties

IRT-model assumptions and fit

The results of the psychometric analyses are presented in Table 3. The one-factor model did not achieve the predefined fit for unidimensionality, and some local dependency was reported, suggesting multidimensionality. The bi-factor model had much better fit, but still marginally below the set criteria. The high ECV (0.70) and Omega-H (0.92) suggested a low risk of biased scores when unidimensionality is assumed. The scalability of the whole scale was weak according to Mokken's rules of thumb (Mokken H 0.34).

Table 3; Results of the psychometric analyses.

Analyses	Outcome	Result
IRT assumptions and model fit		
CFA of one-factor model	Scaled CFI	0.816
	Scaled TLI	0.806
	Scaled RMSEA	0.093
Local Dependency, one-factor model	Residual correlation	15 item pairs locally dependent (2%)
CFA of bi-factor model	Scaled CFI	0.922
	Scaled TLI	0.915
	Scaled RMSEA	0.062
Local Dependency of bi-factor model	Residual correlation	3 item pairs locally dependent (0.4%)
Risk of biased scores	ECV	0.70
	Omega-H	0.92
Monotonicity	Scalability coefficient H	Weak 0.34

Table 4 shows the scalability coefficients, fit statistics, the slope and threshold parameters, and measurement invariance of all items. In nine items less than 5 respondents chose the highest response category (category 6, always): PAINBE17, PAINBE23, PAINBE27, PAINBE34, PAINBE37, PAINBE39, PAINBE40, PAINBE41, and PAINBE45. For these nine items, we collapsed the two highest response categories into one. The scalability coefficient H_i for individual items ranged from 0.14 (PAINBE40) to 0.41 (PAINBE21 and PAINBE43). We found five items with a scalability coefficient lower than the required 0.3: PAINBE29, PAINBE46, PAINBE50, PAINBE38, and PAINBE40. The lowest value of $S\text{-}X^2$ was 0.03, well above the limit of 0.001, indicating good item fit for all items. Item slope parameters ranged from 0.60 (PAINBE40) to 2.00 (PAINBE26). Item threshold parameters ranged from -2.05 (PAINBE2 and PAINBE24) to 6.80 (PAINBE38). Our study population showed a wide coverage of the Pain Behaviour trait, with T-scores ranging from 21.6 to 79.5, with a mean of 50.2 (SD 10.4).

Measurement invariance

One item showed uniform DIF for gender: PAINBE27, *"I had pain so bad it made me cry"* was more likely to be reported by women at similar levels of pain behaviour. One item showed non-uniform DIF between the age groups below and above 60: PAINBE29, *"When I was in pain I used a cane or something else for support"* was more likely to be reported at lower levels of pain behaviour by patients aged >60 than in the upper range of the pain behaviour scale. No DIF was shown for the other age groups.

Table 4; Mokken H_i fit statistics, GRM calibration parameters and measurement invariance of the Dutch-Flemish Pain Behaviour item bank.

Item		Monotonicity	Item fit statistics[b]		Slope[c]	Category threshold [d]				Measurement invariance		
ID	Item phrasing	Mokken H_i [a]	S-X^2	Prob X^2	a	b1	b2	b3	b4	Age>60	Gender	Language
PAINBE2	When I was in pain I became irritable	0.32	231.42	0.4242	1.08	**-2.05**	-0.78	1.43	4.01			
PAINBE3	When I was in pain I grimaced	0.38	222.16	0.3020	1.40	-1.88	-0.83	0.99	3.27			UD
PAINBE6	When I was in pain I would lie down	0.30	261.33	0.1535	1.13	-0.65	0.33	1.91	4.22			
PAINBE8	When I was in pain I moved extremely slowly	0.38	268.67	0.1149	1.42	-1.30	-0.28	1.05	2.61			
PAINBE9	When I was in pain I became angry	0.33	150.40	0.9740	1.45	-0.03	0.97	2.67	4.48			
PAINBE11	When I was in pain I clenched my teeth	0.34	187.50	0.9598	1.48	-0.12	0.66	1.86	3.54			
PAINBE13	When I was in pain I tried to stay very still	0.31	210.35	0.9014	1.17	-0.88	0.35	1.81	4.02			
PAINBE16	When I was in pain I appeared upset or sad	0.38	167.95	0.8506	1.81	0.07	0.91	2.18	3.83			
PAINBE17	When I was in pain I gasped	0.34	200.07	0.0255	1.59	0.80	1.51	2.87	Col			
PAINBE18	When I was in pain I asked for help doing things that needed to be done	0.31	228.56	0.4962	1.20	-0.52	0.50	2.23	4.34			
PAINBE21	When I was in pain it showed on my face (squinting eyes, opening eyes wide, frowning)	0.41	191.74	0.7214	1.84	-0.63	0.13	1.38	2.95			
PAINBE22	Pain caused me to bend over while walking	0.33	222.32	0.8254	1.37	-0.17	0.46	1.50	3.05			UD

Item		Monotonicity	Item fit statistics[b]		Slope[c]	Category threshold [d]				Measurement invariance		
PAINBE23	When I was in pain I asked one or more people to leave me alone	0.34	208.28	0.1258	1.45	0.08	1.03	2.76	Col			
PAINBE24	When I was in pain I moved stiffly	0.35	240.33	0.6078	1.17	**-2.05**	-1.01	0.72	2.98			
PAINBE25	When I was in pain I called out for someone to help me	0.32	198.82	0.3524	1.32	-0.10	1.05	2.89	4.79			UD
PAINBE26	Pain caused me to curl up in a ball	0.40	206.05	0.3882	**2.00**	0.09	0.70	1.62	2.76			UD
PAINBE27	I had pain so bad it made me cry	0.37	166.30	0.2352	1.91	0.72	1.30	2.62	Col		UD	
PAINBE28	When I was in pain I squirmed	0.39	146.70	0.8570	1.98	0.44	1.14	2.27	3.49			
PAINBE29	When I was in pain I used a cane or something else for support	0.29	159.72	0.1606	1.29	1.76	2.34	3.63	5.01	NUD		
PAINBE31	I limped because of pain	0.31	241.35	0.7505	1.22	-0.08	0.50	1.69	3.13			
PAINBE32	When I was in pain I became quiet and withdrawn	0.38	207.29	0.5402	1.57	-0.61	0.26	1.67	3.66			
PAINBE33	When I was in pain I frowned	0.32	229.40	0.3521	1.23	-0.53	0.38	2.09	4.50			
PAINBE34	When I was in pain I asked for help when walking or changing positions	0.30	161.52	0.6054	1.35	1.23	2.01	3.14	Col			
PAINBE35	When I was in pain I groaned	0.34	196.71	0.5529	1.37	-0.29	0.70	2.55	4.83			
PAINBE37	When I was in pain I isolated myself from others	0.36	196.42	0.4377	1.67	0.04	0.82	2.03	Col			
PAINBE38	When I was in pain I drew my knees up	0.23	239.19	0.2921	0.82	0.45	1.18	3.15	**6.80**			
PAINBE39	When I was in pain I moaned, whined or whimpered	0.32	165.40	0.7589	1.34	0.22	1.38	2.96	Col			

Item		Monotonicity	Item fit statistics[b]		Slope[c]	Category threshold [d]				Measurement invariance		
PAINBE40	When I was in pain I flung my arms or limbs around	0.14	197.58	0.1753	**0.60**	2.43	3.59	6.07	Col			
PAINBE41	When I was in pain I screamed	0.36	111.68	0.5965	1.90	1.33	2.23	3.43	Col			
PAINBE42	When I was in pain my upper body would tense up	0.34	226.45	0.4044	1.46	0.02	0.64	1.78	3.40			
PAINBE43	When I was in pain I walked carefully	0.41	262.78	0.2919	1.50	-1.22	-0.52	0.77	2.23			
PAINBE44	When I was in pain I bit or pursed my lips	0.35	163.66	0.7882	1.62	0.60	1.28	2.62	4.22			
PAINBE45	When I was in pain I thrashed	0.33	80.15	0.6579	1.75	1.92	2.69	3.82	Col			
PAINBE46	When I was in pain I protected the part of my body that hurt	0.29	261.50	0.1856	1.10	-0.31	0.44	1,87	3.94			
PAINBE47	When I was I pain my body became stiff	0.36	222.87	0.5467	1.53	-0.39	0.34	1.55	3.07			
PAINBE48	When I was in pain I clenched my jaw or gritted my teeth	0.33	221.52	0.1648	1.41	0.42	1.13	2.46	4.43			
PAINBE49	When I was in pain I winced	0.39	181.40	0.8097	1.68	-0.57	0.14	1.64	3.70			
PAINBE50	When I was in pain I moved my limbs protectively	0.28	233.62	0.0758	1.17	0.67	1.52	2.83	4.84			UD
PAINBE51	When I was in pain I avoided physical contact with others	0.36	230.58	0.2510	1.61	-0.12	0.55	1.69	3.11			

[a] Mokken H_i indicates the scalability of each individual item

[b] Item fit indicates whether the item fits the IRT model, p-value < 0.001 indicates poor fit

[c] Slope indicates the discriminative ability of each item

[d] Threshold indicates the location of this item category on the scale of the measured construct

Highest and lowest slope and threshold parameters are printed in bold. UD uniform DIF, NUD non-uniform DIF

Cross-cultural validity

Five items showed uniform DIF for language. Items PAINBE3 "When I was in pain I grimaced", PAINBE22, "Pain caused me to bend over while walking", PAINBE25, "When I was in pain I called out for someone to help me", and PAINBE26, "Pain caused me to curl up in a ball" were more likely to be reported by the Dutch patients at similar levels of pain behaviour. Item PAINBE50, "When I was in pain I moved my limbs protectively" was more likely to be reported by the US population at similar levels of pain behaviour.

The combined influence of these items is depicted in Figure 1. It shows that there is a difference of less than ten points in T-scores between US and NL patients with a similar level of pain behaviour when using DIF items only, but the difference is negligible when using the item bank as a whole.

Discussion

We studied the validity of the Dutch-Flemish version of the PROMIS v1.1 Pain Behaviour item bank in a large population of patients presenting with predominant complaints of musculoskeletal pain. Our results indicated that all items fitted the IRT model, and that the item bank covered a wide range of the pain behaviour construct. IRT-model assumptions concerning unidimensionality and monotonicity were not met.

CFA fit indices and the presence of local dependence showed suboptimal fit of a one-factor model, suggestive of multidimensionality. An earlier validation study by Crins et al.(13) reported better fit indices. However, in their study unscaled indices were presented. We suggest that scaled indices should be reported, because the distribution of the data is non-normal, and scaled indices correct for overestimation due to non-normality. The original US calibration study of Revicki et al.(7) only reported fit indices after modelling the original 52 candidate items (CFI 0.902, TLI 0.991, RMSEA 0.156), however it was not reported whether these were scaled or unscaled. Multidimensionality could hinder the possibility to calibrate an item bank on a single scale, and forcing a multidimensional item bank into a unidimensional model could lead to biased T-scores. We studied the risk of biased scores with a bi-factor model. This bi-factor model showed a better, but still suboptimal fit of the data. Although the high Omega-H and the ECV indicated a low risk of biased scores when regarding the item bank as unidimensional(21, 22), the reliability of these coefficients depends on the proper fit of the bi-factor model. Several studies have suggested that optimal unidimensionality according to the CFA criteria might be unachievable for item banks developed for clinical measurement(12, 22, 27), and it has been suggested that in using IRT modelling for PROMs practicality is more important than obtaining optimal unidimensionality(28, 29). Overall,

there seems to be enough support to use the item bank as a unidimensional instrument, but it will be interesting to see the results of other studies in different populations.

The monotonicity assumption was not sufficiently met and scalability of the whole item bank was weak, as indicated by a Mokken H of 0.34. Three items showed scalability indices that were marginally lower than the required 0.3, and two items showed scalability indices that clearly deviated from the required 0.3 (PAINBE38, H=0.23, and PAINBE40, H=0.14). These two items showed relatively high threshold parameters, which was not reported in other studies. The high threshold parameters could be related to the low scalability. Based on our study, one could consider removing these items from the item bank. However, as other studies did not report similar results, such a decision would be premature.

Our analyses of measurement invariance showed one item displaying DIF for age, and one item displaying DIF for gender. The impact of DIF on the total scores was low, which

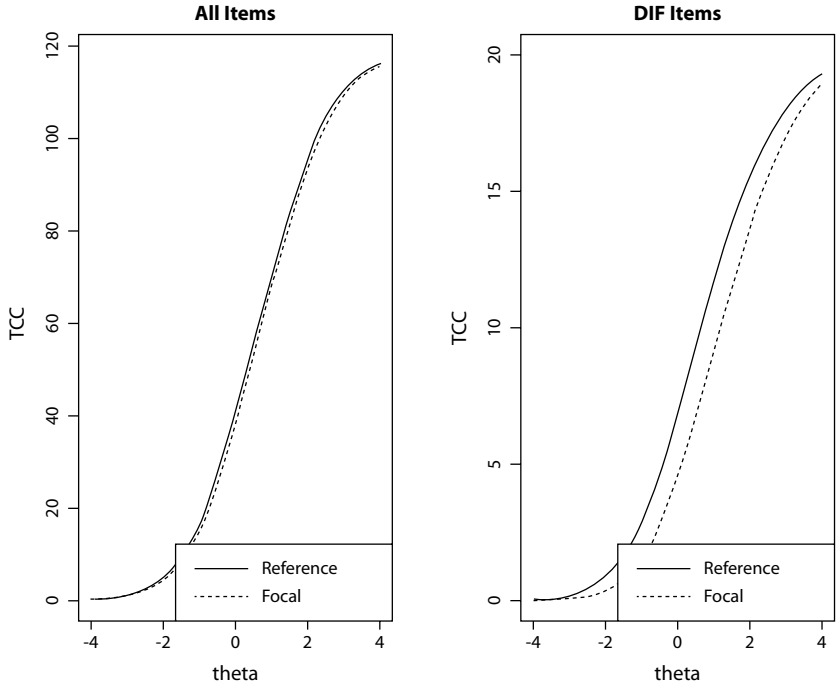

Figure 1; Test Characteristics Curves showing the influence of the DIF for language on theta estimates.

The figure on the left shows the impact of DIF for language on the Test Characteristics Curve when using the item bank as a whole. The figure on the right shows the impact when using DIF items only. The reference line represents the Dutch sample. It shows that, when only the DIF items were used, our study population was more likely to report these items at lower level of the pain behaviour construct.

means that the item bank can be used to compare subgroups of patients who differ on these characteristics with respect to age and gender.

Our analyses of cross-cultural validity showed five items with DIF for language. These five items were among the six items that showed DIF for language in the study by Crins et.al. (13). The impact of DIF on the total score of the item bank was negligible. However, the difference in T-scores can be considerable when only DIF items are used, which decreases the comparability between groups of patients for which DIF has been shown (Fig. 1). Two of the items showing DIF for language (PAINBE3 and PAINBE25) are part of the standard 7-item short form. Although the influence of DIF is likely to be small when the whole short form is used some care must be taken comparing the outcomes of these short forms between Dutch and US patients.

Strengths and weaknesses

The strength of our study lies in the large number of patients (N=1602) answering the 39 item PROMIS Pain Behaviour item bank in a primary care population of patients with a range of musculoskeletal complaints. Previous studies included a general population sample, supplemented with patients recruited from the ACPA website(7), or a population with longstanding pain in a rehabilitation setting, probably displaying more activity limitations. By combining our data with a part of the US calibration sample we could study Differential Item Functioning for language to examine cross-cultural validity. A weakness of our study could be the relatively low number of patients with high levels of pain behaviour. This caused sparsity of data in the extreme ends of the scale which could have influenced CFA fit statistics.

Further study

The results of our study, and of previous studies indicate that this item bank can be used as a basis for short forms and for CAT assessment in clinical research and in clinical practice. Further IRT analyses are recommended in a Dutch general population and on a combined set of clinical and general population data to determine the optimal IRT item parameters for use of CAT in The Netherlands and Flanders (Dutch or Flemish speaking part of Belgium). We also recommend more validation studies of translated versions in other languages and in populations with other types of pain, especially to study dimensionality and monotonicity.

Conclusion

The Dutch-Flemish PROMIS Pain Behaviour item bank showed that all items fitted the IRT model, and our results supported cross-cultural validity. However, the assumptions of

unidimensionality and monotonicity were not met. Bi-factor analysis indicated a low risk of biased scores when assuming unidimensionality, although the fit of the bi-factor model was still suboptimal. We conclude that, the DF-PROMIS-Pain Behaviour item bank can be used in clinical research and in clinical practice, although further research should examine whether problems concerning dimensionality and monotonicity occur in other populations.

Acknowledgements

We would like to thank all members of the Dutch Association for Musculoskeletal Medicine who cooperated in this study. We also would like to thank K. Uegaki for reviewing the manuscript.

References

1. McCahon S, Strong J, Sharry R, Cramond T. Self-report and pain behaviour among patients with chronic pain. Clin J Pain. 2005;21(3):223-31.
2. Prkachin KM, Schultz IZ, Hughes E. Pain behaviour and the development of pain-related disability: the importance of guarding. Clin J Pain. 2007;23(3):270-7.
3. Krause SJ, Wiener RL, Tait RC. Depression and pain behaviour in patients with chronic pain. Clin J Pain. 1994;10(2):122-7.
4. Keefe FJ, Smith S. The assessment of pain behaviour: implications for applied psychophysiology and future research directions. Appl Psychophysiol Biofeedback. 2002;27(2):117-27.
5. Waters SJ, Riordan PA, Keefe FJ, Lefebvre JC. Pain behaviour in rheumatoid arthritis patients: identification of pain behaviour subgroups. J Pain Symptom Manage. 2008;36(1):69-78.
6. Fordyce WE, Brockway JA, Bergman JA, Spengler D. Acute back pain: a control-group comparison of behavioural vs traditional management methods. J Behav Med. 1986;9(2):127-40.
7. Revicki DA, Chen WH, Harnam N, Cook KF, Amtmann D, Callahan LF, et al. Development and psychometric analysis of the PROMIS pain behaviour item bank. Pain. 2009;146(1-2):158-69.
8. Chakravarty EF, Bjorner JB, Fries JF. Improving patient reported outcomes using item response theory and computerized adaptive testing. J Rheumatol. 2007;34(6):1426-31.
9. de Vet HC, Terwee CB, Mokkink LB, Knol D. Measurement in Medicine. Cambridge: Cambridge University Press; 2011.
10. Fayers PM. Applying item response theory and computer adaptive testing: the challenges for health outcomes assessment. Qual Life Res. 2007;16 Suppl 1:187-94.
11. Haley SM, Ni P, Hambleton RK, Slavin MD, Jette AM. Computer adaptive testing improved accuracy and precision of scores over random item selection in a physical functioning item bank. J Clin Epidemiol. 2006;59(11):1174-82.
12. Reise SP, Waller NG. Item response theory and clinical measurement. Annu Rev Clin Psychol. 2009;5:27-48.
13. Crins MH, Roorda LD, Smits N, de Vet HC, Westhovens R, Cella D, et al. Calibration of the Dutch-Flemish PROMIS Pain Behaviour item bank in patients with chronic pain. Eur J Pain. 2016;20(2):284-96.
14. Terwee CB, Roorda LD, de Vet HC, Dekker J, Westhovens R, van Leuwen J, et al. Dutch-Flemish translation of 17 item banks from the patient-reported outcomes measurement information system (PROMIS). Qual Life Res. 2014;23(6):1733-41.
15. Reeve BB, Hays RD, Bjorner JB, Cook KF, Crane PK, Teresi JA, et al. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). Med Care. 2007;45(5 Suppl 1):S22-S31.
16. Bock RD, Aitkin M. Marginal Maximum-Likelihood Estimation of Item Parameters - Application of an Em Algorithm. Psychometrika. 1981;46(4):443-59.

17. Muthen B, du Toit, S.H.C.; Spisic, D. Robust Inference using Weighted Least Squares and Quardratic Estimating Equations in Latent Variable Modeling with Categorical and Continuous Outcomes. 1997.
18. Hu LT, Bentler P. Cutoff criteria for fit indices in covariance structure analysis: conventional criteria versus new alternatives. Structural equation modelling. 1999 1999:1-55.
19. Reise SP. Invited Paper: The Rediscovery of Bifactor Measurement Models. Multivariate Behav Res. 2012;47(5):667-96.
20. Whittaker TA. Using the Modification Index and Standardized Expected Parameter Change for Model Modification. J Exp Educ. 2012;80(1):26-44.
21. Rodriguez A, Reise SP, Haviland MG. Applying Bifactor Statistical Indices in the Evaluation of Psychological Measures. J Pers Assess. 2016;98(3):223-37.
22. Reise SP, Scheines R, Widaman KF, Haviland MG. Multidimensionality and Structural Coefficient Bias in Structural Equation Modeling: A Bifactor Perspective. Educ Psychol Meas. 2013;73(1):5-26.
23. Van der Ark LA. Mokken scale analysis in R. J Stat Softw. 2007;20(11):1-19.
24. van der Ark LA, Croon MA, Sijtsma K. Mokken Scale Analysis for Dichotomous Items Using Marginal Models. Psychometrika. 2008;73(2):183-208.
25. Choi SW, Gubbons LE, Crane PK. lordif: an R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/ item response theory and monte carlo simulations. J Stat Softw. 2011 2011:1-30.
26. Crane PK, Gibbons LE, Jolley L, van Belle G. Differential Item Functioning Analysis with Ordinal Logistic Regression Techniques. Medical Care2006. p. S115-S23.
27. Cook KF, Kallen MA, Amtmann D. Having a fit: impact of number of items and distribution of data on traditional criteria for assessing IRT's unidimensionality assumption. Qual Life Res. 2009;18(4):447-60.
28. Browne MW, Cudeck R. Alternative Ways of Assessing Model Fit. Sociol Method Res. 1992;21(2):230-58.
29. Iacobucci D. Structural equations modeling: Fit Indices, sample size, and advanced topics. J Consum Psychol. 2010;20(1):90-8.

Chapter 7.

Validation of the Dutch-Flemish PROMIS Pain Interference item bank in patients with musculoskeletal complaints

Wouter Schuller M.D, Caroline B. Terwee PhD, Thomas Klausch PhD, Leo D. Roorda M.D., PhD, D.C. Rohrich MsC, Raymond W. Ostelo PhD, Berend Terluin M.D., PhD, Henrica C.W. de Vet PhD

Spine (Philadelphia 1976). 2019;44(6): 411-419

Abstract

Study design: Cross sectional study.

Objective: To validate the Dutch-Flemish PROMIS Pain Interference item bank in patients with musculoskeletal complaints.

Summary of background data: PROMIS item banks have been developed and validated in the US. They need to be further validated in various patient populations and in different languages.

Methods: 1677 patients answered the full item bank. A Graded Response Model (GRM) was used to study dimensionality with confirmatory factor analyses and by assessing local independency. Monotonicity was evaluated with Mokken scaling. An IRT model was used to study item fit, and to estimate slope and threshold parameters. Differential Item Functioning (DIF) for language, age and gender was assessed using ordinal logistic regression analyses. DIF for language was evaluated by comparing our data with a similar US sample. Hypotheses concerning construct validity were tested by correlating item bank-scores with scores on several legacy instruments.

Results: The GRM showed suboptimal evidence of unidimensionality in confirmatory factor analysis (CFI: 0.903, TLI: 0.897, RSMEA: 0.144), and 99 item pairs with local dependence. A bi-factor model showed good fit (CFI: 0.964, TLI: 0.961, RSMEA: 0.089), with a high Omega-H (0.97), a high Explained Common Variance (ECV: 0.81), and no local dependence. Sufficient monotonicity was shown for all items (Mokken $H_{(i)}$: 0.367-0.686). The unidimensional IRT model showed good fit (Only two items with S-X^2 < 0.001), with slope parameters ranging from 1.00 to 4.27, and threshold parameters ranging from -1.77 to 3.66. None of the items showed DIF for age or gender. One Item showed DIF for language. Correlations with legacy instruments were high (Pearson's R: 0.53-0.75), supporting construct validity.

Conclusion: The high omega-H and the high ECV indicate that the item bank could be considered essentially unidimensional. The item bank showed good item fit, good coverage of the pain interference trait, and good construct validity.

Introduction

In 2004 a US National Institutes of Health (NIH) initiative set out to develop new PROMs for clinical research and health care delivery settings, based upon Item Response Theory (IRT); the Patient Reported Outcome Measurement Information System (PROMIS ®)(1-3). Under IRT item banks are constructed consisting of a large collection of questions (i.e. items) covering a wide range of a trait. These item banks are calibrated by modelling the relationship between a person's level of the construct and the likelihood of choosing a response on each item. After calibration item banks can give comparable scores on a standardised scale, even when subsets of items are used, whilst retaining reliability(4-7). Item banks can be used in Computer Adaptive Testing (CAT). In CAT, a computer algorithm decides on the basis of previous answers which next item would be most informative. The questions are thus tailored to the individual patient and only a small number of questions (on average 5 to 7) are needed to obtain a reliable score that can be compared to a score obtained from administering all items on the same scale(3, 8-11). IRT outcome measures are expected to play a major role in clinical measurement(12). The recently suggested research standards from the NIH taskforce for measurement of chronic low back pain, for example, already contain several items from the PROMIS Pain Interference item bank(13).

The Pain Interference item bank was developed as a unidimensional instrument, measuring the self-reported consequences of pain on relevant aspects of one's life. This includes the extent to which pain hinders engagement with social, cognitive, emotional, physical, and recreational activities. A large number of PROMIS item banks have been translated into Dutch-Flemish by the Dutch-Flemish PROMIS group(14), among others the v1.1 Pain Interference item bank. A previous study showed good cross-cultural and construct validity, good reliability, and good coverage of the pain interference continuum for the Dutch-Flemish translation of the v1.1 Pain Interference item bank (DF-PROMIS-PI) in a population of rehabilitation patients(15). For patients with musculoskeletal complaints in The Netherlands there is a possibility to consult physicians who are trained in musculoskeletal (MSK) medicine(16). Most MSK practices are primary care facilities primarily focused on patients with musculoskeletal pain. Patients generally consult MSK physicians with complaints of low back pain, with or without sciatica, neck pain, headache, and pain in the upper or lower extremities(16). Before using item banks in patients with various conditions it is necessary to validate them in different patient populations. For international use it is necessary to validate item banks in different languages. The aim of our present study was to validate the v1.1 DF-PROMIS-PI item bank in a large sample of patients presented in musculoskeletal practice.

Materials and Methods

Study design and procedure

We conducted a cross-sectional study using an existing web-based registry of patients presenting for the first time in MSK practice. The data documented in this registry were collected by a group of 31 MSK physicians in The Netherlands who agreed to participate in the establishment of the patient registry. At the first visit the treating physician entered the following patient characteristics via computer: age, gender, type and duration of the main complaint and the existence of concomitant complaints. Complaints were recorded by the treating physician according to the International Classification of Primary Care (ICPC) (17). Treating physicians asked patients if they were interested in participating in the study. Following an informed consent procedure, the treating physician entered email addresses of the recruited patients in the registry. Thereafter, a specially designed computer program (Readmail) automatically distributed invitations to patients by email to fill in web-based questionnaires. Data used for this present study were collected in the registry from October 2013 until February 2014. Our study procedures were approved by the Medical Ethical Committee of the VU Medical Center (2013/20).

Participants

MSK physicians were instructed to invite all consecutive patients who presented for the first time in MSK practice to participate. To evaluate cross-cultural validity we used a part of the sample that was used in the original US calibration study. This sample consisted of 967 patients who were recruited through the website of the American Chronic Pain Association (ACPA), and who had at least one chronic pain condition for at least three months prior to participating in the survey(18).

Measures

The PROMIS-PI item bank was developed as part of the NIH PROMIS project, and contains 40 items. The temporal context for all items is 7 days. Response categories are divided into three sets fitting the specific items: (1) not at all, a little bit, somewhat, quite a bit, very much, (2) never, rarely, sometimes, often, always, and (3) never, once a week or less, once every few days, once a day, every hour. The item bank was calibrated in a large US study on a population including a community sample, and clinical samples of cancer patients, and of patients with chronic pain recruited through the American Chronic Pain Association (US-ACPA sample). Translation of the item bank into Dutch-Flemish was carried out by FACITtrans according to standard PROMIS methodology and approved by the PROMIS Statistical Center(14). Our study population completed the full 40 item v1.1 Dutch-Flemish PROMIS Pain Interference item bank.

In addition to completing the PROMIS Pain Interference item bank, our study participants were asked to complete one of five condition-specific (legacy) instruments, according to their respective main complaint: the Roland Disability Questionnaire (RDQ)(19), the Neck Disability Index (NDI)(20), the Lower Extremity Function Scale (LEFS)(21), the Disabilities of the Arm, Shoulder and Hand (DASH)(22), and the Headache Impact Test (HIT-6)(23), for patients with low back pain, neck pain, lower extremity pain, upper extremity pain, or headache, respectively. The number of items, and the range of scores are indicated in Table 4. All legacy instruments are frequently used in research and have been validated in Dutch populations(24-30).

Statistical analyses

Statistical analyses were carried out according to the PROMIS plan for psychometric evaluation and calibration of health-related quality of life item banks(31). Descriptive analyses were carried out using SPSS statistics, version 22.

We evaluated dimensionality by confirmatory factor analyses (CFA) and by assessing local independency in a Graded Response Model. Model fit was evaluated by the following indices: the Comparative Fit Index (CFI, >0.95 for good fit), Tucker-Lewis Index (TLI, >0.95 for good fit), and the Root Mean Square Error of Approximation (RMSEA, <0.06 for good fit)(32). To evaluate the influence of multidimensionality a bi-factor model was fitted, and omega-H and Explained Common Variance (ECV) were calculated. A high coefficient omega (> 0.80)(33) and a high ECV (> 0.60)(34) indicate that the risk of biased parameters when fitting multidimensional data into a unidimensional model is low. CFA was carried out with the R-package Lavaan (version 0.5-23.1097).

We assessed monotonicity as a measure of scalability with the R-package Mokken(35). Mokken H was interpreted according to the following rules of thumb: unscalable if $H_{(i)} < 0.3$, weak if $0.3 \leq H_{(i)} < 0.4$, moderate if $0.4 \leq H_{(i)} < 0.5$, and strong if $H_{(i)} \geq 0.5$(35, 36).

An IRT model was used to study item fit, and to calculate slope and threshold parameters. Items were considered to misfit if the p-value is < 0.001. T-scores were calculated based on US calibration parameters with the expected a priori method, using the R-package Mirt (version 1.24)(37). A T-score of 50 represents the mean score of the general population, with a SD of 10.

DIF was assessed for several age groups, for gender and for language. We evaluated DIF for language (English vs Dutch) by comparing our data with the data available from the US-ACPA sample used by Amtmann et.al. (N=967)(38). DIF was analysed using ordinal

logistic regression models with the R-package Lordif (version 0.3-3)(39, 40), with theta as an estimation of the trait level. The change in McFadden's R^2 was used as an indicator of DIF, with a value of >0.02 serving as the critical value for rejecting the null hypothesis of no DIF(39).

Construct validity was studied by testing hypotheses about the correlation of T-scores with the scores on several legacy instruments using SPSS statistics, version 22. Our hypothesis was that for the condition-specific subgroups of patients the T-scores would correlate with the corresponding functional legacy instruments (R>0.50).

Results

Demographic characteristics

2610 patients were asked to participate in our study; 2171 consented. Of these 2171 patients 1745 (67%) answered the questionnaires. Because only the year of birth was reported we excluded patients who could have been under the age of 18 at inclusion. A small number of patients failed to answer any item at all. After removal of patients under 18, and patients who had failed to complete the whole item bank, a sample of 1677 patients (64%) remained. Another 27 patients had a number of missing items. Model analyses were conducted on the sample of 1677 patients. T-scores were calculated for the 1650 patients who had answered all items. Demographic data of our sample are presented in Table 1, together with the demographic data of the US-ACPA sample used in the DIF analyses. Half of the patients presented with a primary complaint of low back pain (50.6%), with or without sciatica, followed by neck or shoulder pain (20.8%).

Table 1; Demographics of patient sample: age, gender, duration of main complaints, primary complaints, T-scores, and scores on legacy instruments. Comparison with US-ACPA sample.

	MSK sample (N=1677)	US-ACPA sample (N=967)
General background		
Age mean (SD)	47 (14)	48 (11)
Gender (% female)	59	81
Duration of complaints number (%)		
< 3 months	259 (15.9)	
3 months- 1 year	330 (20.2)	53 (6)
> 1 year	1041 (63.9)	876 (94)
Type of complaints number (%) [a]		
Low back pain	849 (50.6)	533 (55)
Neck or shoulder pain	349 (20.8)	447 (46)
Other back pain	134 (8.0)	
Lower extremity	163 (9.7)	
Headache	56 (3.3)	290 (22)
Upper extremity	37 (2.2)	
Other	85 (5.4)	
Pain Interference scores		
T-score mean (SD)	58.1 (6.7)	68.6 (4.9)
T-score range	37.4 – 76.1	53.0 – 90.0
Legacy scores [b] **mean (range)**		
RDQ (N=827)	8.9 (0-23)	
NDI (N=269)	13.1 (0-33)	
LEFS (N=159)	55.0 (11-80)	
DASH (N=102)	31.6 (2.5-69.2)	
HIT-6 (N=54)	60.2 (36-73)	

[a] In our study only the main complaint could be scored, while in the US-ACPA study multiple complaints could be indicated.

[b] Roland Disability Questionnaire; 24 items, range 0-24; Neck Disability Index; 10 items, range 0-50; Lower Extremity Function Scale; 20 items, range 0-80; Disabilities of the Arm, Shoulder and Hand; 30 items, range 0-100; Headache Impact Test-6; 6 items, range 36-78

Dimensionality

The fit of a one-factor model in the CFA resulted in a CFI of 0.903 (unscaled 0.978), a TLI of 0.897 (unscaled 0.978), and a RSMEA of 0.145 (unscaled 0.185). CFA fit indices indicated suboptimal fit of a one-factor model. Evaluation of the residual correlation matrix showed local dependence for 99 of the possible 780 (1/2 X 40 X 39) item pairs (12%), with residual correlations greater than 0.2.

The bi-factor model contained one general factor, and five group factors. The group factor items are presented in Table 2. For the bi-factor model the fit indices were higher than for the one-factor model. CFI was 0.964 (unscaled 0.996), the TLI was 0.961 (unscaled 0.996), and the RMSEA was 0.089 (unscaled 0.083). Omega-H was 0.97, and ECV was 0.81. In the bi-factor model no item pairs showed residual correlations greater than 0.2.

Table 2; Group factors from the bi-factor analyses.

Factor	Item code [a]	Item
1	PAININ40	How often did pain prevent you from walking more than 1 mile?
	PAININ42	How often did pain prevent you from standing for more than one hour?
	PAININ47	How often did pain prevent you from standing for more than 30 minutes?
2	PAININ50	How often did pain prevent you from sitting for more than 30 minutes?
	PAININ51	How often did pain prevent you from sitting for more than 10 minutes?
	PAININ54	How often did pain keep you from getting into a standing position?
	PAININ55	How often did pain prevent you from sitting for more than one hour?
3	PAININ11	How often did you feel emotionally tense because of your pain?
	PAININ16	How often did pain make you feel depressed?
	PAININ24	How often was pain distressing to you?
	PAININ29	How often was your pain so severe you could think of nothing else?
	PAININ32	How often did pain make you feel discouraged?
	PAININ37	How often did pain make you feel anxious?
4	PAININ1	How difficult was it for you to take in new information because of pain?
	PAININ8	How much did pain interfere with your ability to concentrate?
	PAININ49	How much did pain interfere with your ability to remember things?
	PAININ56	How irritable did you feel because of pain?
5	PAININ9	How much did pain interfere with your day to day activities?
	PAININ18	How much did pain interfere with your ability to work (include work at home)?
	PAININ22	How much did pain interfere with work around the home?
	PAININ34	How much did pain interfere with your household chores?
	PAININ48	How much did pain interfere with your ability to do household chores?

[a] PAININ: Pain Interference

Monotonicity

Scalability coefficients are shown in Table 3. All items had a scalability coefficient higher than the required 0.3, ranging from 0.367 (PAININ54; "How often did pain keep you from getting into a standing position?") to 0.686 (PAININ46; "How often did pain make it difficult for you to plan social activities?"). The scalability of the whole scale was H=0.596, which is strong according to Mokken's rules of thumb.

Table 3; Scalability, GRM item parameters, and fit statistics.

	Item	Mokken's	Slope	Category threshold				Item statistics [a]	
		H_i	a	B1	B2	B3	B4	S-X2	Prob X2
PAININ1	How difficult was it for you to take in new information because of pain?	0.574	2.21	-0.07	0.71	1.61	3.00	261.46	0.3276
PAININ3	How much did pain interfere with you enjoyment of life?	0.655	2.98	-0.88	-0.01	0.76	1.95	299.87	0.0052
PAININ5	How much did pain interfere with your ability to participate in leisure activities?	0.623	2.59	-1.07	-0.17	0.47	1.67	343.81	0.0016
PAININ6	How much did pain interfere with your close personal relationships?	0.650	3.38	-0.13	0.53	1.29	2.51	233.87	0.0887
PAININ8	How much did pain interfere with your ability to concentrate?	0.585	2.15	-0.62	0.19	1.07	2.28	344.66	0.0110
PAININ9	How much did pain interfere with your day to day activities?	0.641	2.67	-1.43	-0.27	0.59	1.91	263.09	0.1568
PAININ10	How much did pain interfere with your enjoyment of recreational activities?	0.635	2.78	-1.12	-0.13	0.50	1.64	292.91	0.0612
PAININ11	How often did you feel emotionally tense because of your pain?	0.588	2.11	-0.72	0.06	1.19	3.07	294.79	0.0934
PAININ12	How much did pain interfere with the things you usually do for fun?	0.637	2.83	-1.13	-0.12	0.50	1.63	319.29	0.0034
PAININ13	How much did pain interfere with your family life?	0.632	2.79	-0.55	0.23	1.06	2.23	263.00	0.1924
PAININ14	How much did pain interfere with doing your tasks away from home (e.g. getting groceries, running errands)?	0.642	3.13	-0.16	0.52	1.15	2.06	247.34	0.3088
PAININ16	How often did pain make you feel depressed?	0.545	1.91	-0.18	0.65	1.87	3.64	273.65	0.3128
PAININ17	How much did pain interfere with your relationships with other people?	0.637	3.13	-0.16	0.52	1.40	2.50	268.72	0.0066

	Item	Mokken's	Slope	Category threshold				Item statistics [a]	
		H_i	a	B1	B2	B3	B4	S-X2	Prob X2
PAININ18	How much did pain interfere with your ability to work (include work at home)?	0.644	2.88	-0.84	-0.02	0.67	1.76	238.53	0.7197
PAININ19	How much did pain make it difficult to fall asleep?	0.452	1.30	-0.84	0.28	1.23	2.64	369.47	0.3677
PAININ20	How much did pain feel like a burden to you?	0.637	2.43	-1.77	-0.64	0.12	1.55	241.42	0.7049
PAININ22	How much did pain interfere with work around the home?	0.649	2.79	-1.32	-0.31	0.45	1.68	280.03	0.1084
PAININ24	How often was pain distressing to you?	0.507	1.50	-1.11	-0.21	1.36	3.26	366.90	0.0086
PAININ26	How often did pain keep you from socialising with others?	0.672	3.56	-0.46	0.21	1.11	2.34	206.22	0.3477
PAININ29	How often was your pain so severe you could think of nothing else?	0.598	2.34	-0.22	0.56	1.63	3.28	269.34	0.0937
PAININ31	How much did pain interfere with your ability to participate in social activities?	0.685	4.16	-0.44	0.23	0.89	1.87	307.30	0.0001
PAININ32	How often did pain make you feel discouraged?	0.613	2.34	-0.69	0.07	1.14	2.76	271.46	0.2028
PAININ34	How much did pain interfere with your household chores?	0.645	2.73	-1.24	-0.28	0.49	1.76	263.77	0.2923
PAININ35	How much did pain interfere with your ability to make trips from home that kept you gone for more than 2 hours?	0.658	3.52	0.07	0.60	1.17	1.89	274.31	0.0122
PAININ36	How much did pain interfere with your enjoyment of social activities?	0.662	3.23	-0.74	0.01	0.67	1.75	246.63	0.3039
PAININ37	How often did pain make you feel anxious?	0.520	1.62	-0.53	0.34	1.74	3.54	333.69	0.0509
PAININ38	How often did you avoid social activities because it might make you hurt more?	0.639	3.13	-0.10	0.50	1.22	2.32	239.70	0.2535
PAININ40	How often did pain prevent you from walking more than 1 mile?	0.531	1.83	-0.21	0.34	0.96	1.84	384.26	0.0490

	Item	Mokken's	Slope	Category threshold				Item statistics [a]	
		H_i	a	B1	B2	B3	B4	S-X2	Prob X2
PAININ42	How often did pain prevent you from standing for more than one hour?	0.511	1.53	-0.79	-0.22	0.62	1.82	396.04	0.0925
PAININ46	How often did pain make it difficult for you to plan social activities?	0.686	4.27	-0.25	0.39	1.12	2.16	176.08	0.5898
PAININ47	How often did pain prevent you from standing for more than 30 minutes?	0.517	1.62	-0.39	0.20	1.08	2.13	364.26	0.3419
PAININ48	How much did pain interfere with your ability to do household chores?	0.659	3.05	-0.91	0.09	0.73	1.88	308.91	0.0009
PAININ49	How much did pain interfere with your ability to remember things?	0.562	2.10	0.29	0.96	1.82	2.94	296.11	0.0294
PAININ50	How often did pain prevent you from sitting for more than 30 minutes?	0.512	1.63	-0.13	0.58	1.47	2.71	303.43	0.7259
PAININ51	How often did pain prevent you from sitting for more than 10 minutes?	0.504	1.66	0.44	1.27	2.27	3.66	264.51	0.2667
PAININ52	How often was it hard to plan social activities because you didn't know if you would be in pain?	0.644	3.29	0.06	0.62	1.29	2.16	239.88	0.2509
PAININ53	How often did pain restrict your social life to your home?	0.655	3.39	-0.03	0.58	1.36	2.61	206.16	0.4443
PAININ54	How often did pain keep you from getting into a standing position?	0.367	1.00	0.99	1.73	2.32	3.18	310.84	0.3505
PAININ55	How often did pain prevent you from sitting for more than one hour?	0.487	1.48	-0.17	0.48	1.33	2.50	366.36	0.3970
PAININ56	How irritable did you feel because of pain?	0.571	2.03	-0.97	0.16	1.11	2.48	294.74	0.3330

[a] Statistical significance indicates poor item fit.

Item fit, item parameters and T-scores

After fitting an IRT model to our data we studied item fit and the range of theta's covered. Table 3 shows the fit statistics of all items, and the slope and threshold parameters. There were two items with a S-X^2 below the threshold of 0.001 (PAININ31; "How much did pain interfere with your ability to participate in social activities?" and PAININ48; "How much did pain interfere with your ability to do household chores?"). Slope parameters ranged from 1.00 to 4.27, and threshold parameters ranged from -1.77 to 3.66. The average T-score of our study population was 58.1 (range 37.4-76.1, SD 6.7).

Differential Item Functioning

None of the items showed DIF for any of the age groups or for gender. Uniform DIF for language was demonstrated for one item: PAININ24 ("How often was pain distressing to you?") showed lower threshold parameters for the Dutch population. The influence of this item is depicted in Figure 1. It shows that, in theory, for patients with a similar trait level there would be a difference of less than 0.6 points in expected score when using this DIF item only. However, this difference was negligible when using the item bank as a whole.

Construct validity

Table 4 shows that the T-scores correlated highly (all R> 0.50) with the scores of the legacy instruments.

Table 4; Mean scores and ranges of legacy instruments and correlations between PROMIS Pain Interference T-scores and legacy instruments. Correlation with the LEFS is negative because higher disability is depicted in lower scores.

Instrument [a]			Measurement				Correlation	
	Items	Range	N	mean	min	max	Expected R	Observed R
RDQ	24	0-24	827	8.9	0.0	23.0	>0.50	0.700
NDI	10	0-50	269	13.1	0.0	33.0	>0.50	0.687
LEFS	20	0-100	159	55.0	11.0	80.0	<-0.50	-0.754
DASH	30	0-80	102	31.6	2.5	69.2	>0.50	0.731
HIT-6	6	36-78	54	60.2	36.0	73.0	>0.50	0.527

[a] RDQ: Roland Disability Questionnaire, NDI: Neck Disability Index, LEFS: Lower Extremity Function Scale, DASH: Disabilities of the Arm Shoulder and Hand, HIT-6: Headache Impact Test.

Discussion

We studied the validity of the Dutch-Flemish version of the PROMIS Pain Interference item bank in a large population of patients presenting with predominant complaints of

musculoskeletal pain. The DF-PROMIS-PI item bank showed suboptimal fit to a one-factor model in CFA and some local dependence. None of the items violated the monotonicity assumption. A bi-factor model showed good fit, a high coefficient omega-H and ECV, and no local dependence. The item bank showed good IRT item fit, good coverage of the pain interference construct, and good construct validity.

CFA fit indices and the presence of local dependence suggested suboptimal unidimensionality. In a previous study validating the Dutch-Flemish PROMIS Pain Interference item bank in an outpatient rehabilitation population with chronic pain, Crins et al. reported a better fit (CFI 0.986, TLI 0.986, RMSEA 0.159)(15). In the study of Crins et al., however, unscaled indices were reported, where it is now thought that, due to non-normality of the data, scaled indices should be used. The unscaled indices in our study would suggest better evidence of unidimensionality as well (CFI 0.978, TLI 0.978, RSMEA 0.185). The US calibration study reported good fit (CFI 0.974, TLI 0.997, RMSEA 0.175)(38), but did not state whether scaled or unscaled indices were reported. A secondary analysis on part of the US calibration sample reported suboptimal fit as well (CFI 0.90, TLI 0.90, RMSEA 0.135)(41). A cross-cultural validation study in a Spanish speaking population showed good fit (CFI 0.97, TLI 0.97, RSMEA 0.10) with no local dependency, without reporting whether scaled or unscaled indices were used(42). Some authors have mentioned that unidimensionality could be hard to achieve when developing item banks for clinical measurement(11, 43), and it has been suggested that fit indices should not be regarded as measures of usefulness of a model(44, 45). In our study the bi-factor model, however, showed good fit (CFI 0.964, TLI 0.961, RMSEA 0.089), and the Omega-H coefficient and the ECV were high (0.97 and 0.81 respectively), indicating a low risk of biased parameters when treating the item bank as unidimensional(33, 34).

Item slope parameters ranged from 1.00 to 4.27 and item threshold parameters ranged from -1.77 to 3.66. Considering that under a normal distribution 99,99% of the theta's will be in the range of -4 to +4, this range of threshold parameters represented a good coverage for a population of patients with pain. The item with the lowest slope parameter and both items with the lowest and the highest threshold parameters were the same as those reported by Crins et al(15). In our study two items showed poor fit, as opposed to one different item with poor fit reported by Crins et al.(15), and again one different item reported by Paz et al.(42).

DIF analyses did not show any DIF for age or gender, however, one item with DIF for language was found. The influence of this DIF for language on theta-scores was very limited. Crins et al. reported DIF for language for the same item (PAININ24) and also for item PAININ32, with minimal impact on the Test Characteristics Curve as well(15). As the influence of DIF for language is very limited we suggest that these items can be retained.

The strong correlations with several legacy instruments supports construct validity of the DF-PROMIS-PI item bank. It is interesting to note that the five legacy instruments were developed to measure functional limitations for specific conditions. The correlation of one single item bank with several condition specific legacy instruments supports the generic use of the Pain Interference item bank. The good fit of the bi-factor model, together with the high omega-H and the high ECV indicate that the PROMIS Pain Interference item bank could be considered essentially unidimensional. The limited influence of DIF for language and the strong correlations with legacy instruments supports the validity of the Dutch-Flemish translation. Because of these properties the PROMIS Pain Interference item bank can be considered suitable for use in both clinical research and practice, and can be used as a basis for short forms and computer adaptive testing.

Conclusion

The Dutch-Flemish v1.1 PROMIS Pain Interference item bank showed good IRT item fit, good coverage of the pain interference trait, and good construct validity. None of the items showed DIF for age or gender. One item showed minimal DIF for language. CFA and analyses of local independence showed evidence of multidimensionality, but omega-H and ECV were high, indicating a low risk of biased parameters when assuming unidimensionality. We conclude that our results support the validity of the DF-PROMIS-Pain Interference item bank, and that the item bank can be used as a basis for short forms and computer adaptive testing in clinical research and in clinical practice.

References

1. Cella D, Gershon R, Lai JS, Choi S. The future of outcomes measurement: item banking, tailored short-forms, and computerized adaptive assessment. Qual Life Res. 2007;16 Suppl 1:133-41.
2. Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. J Clin Epidemiol. 2010;63(11):1179-94.
3. de Vet HC, Terwee CB, Mokkink LB, Knol D. Measurement in Medicine. Cambridge: Cambridge University Press; 2011.
4. Fries J, Rose M, Krishnan E. The PROMIS of better outcome assessment: responsiveness, floor and ceiling effects, and Internet administration. J Rheumatol. 2011;38(8):1759-64.
5. Fries JF, Krishnan E, Rose M, Lingala B, Bruce B. Improved responsiveness and reduced sample size requirements of PROMIS physical function scales with item response theory. Arthritis Res Ther. 2011;13(5):R147.
6. Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. Med Care. 2000;38(9 Suppl):II28-II42.
7. Tugwell P, Knottnerus JA, Idzerda L. Tailoring patient reported outcome measurement. J Clin Epidemiol. 2010;63(11):1165-6.
8. Chakravarty EF, Bjorner JB, Fries JF. Improving patient reported outcomes using item response theory and computerized adaptive testing. J Rheumatol. 2007;34(6):1426-31.
9. Fayers PM. Applying item response theory and computer adaptive testing: the challenges for health outcomes assessment. Qual Life Res. 2007;16 Suppl 1:187-94.
10. Haley SM, Ni P, Hambleton RK, Slavin MD, Jette AM. Computer adaptive testing improved accuracy and precision of scores over random item selection in a physical functioning item bank. J Clin Epidemiol. 2006;59(11):1174-82.
11. Reise SP, Waller NG. Item response theory and clinical measurement. Annu Rev Clin Psychol. 2009;5:27-48.
12. Fries JF, Krishnan E. What constitutes progress in assessing patient outcomes? J Clin Epidemiol. 2009;62(8):779-80.
13. Deyo RA, Dworkin SF, Amtmann D, Andersson G, Borenstein D, Carragee E, et al. Focus article: report of the NIH task force on research standards for chronic low back pain. Eur Spine J. 2014;23(10):2028-45.
14. Terwee CB, Roorda LD, de Vet HC, Dekker J, Westhovens R, van Leuwen J, et al. Dutch-Flemish translation of 17 item banks from the patient-reported outcomes measurement information system (PROMIS). Qual Life Res. 2014;23(6):1733-41.
15. Crins MH, Roorda LD, Smits N, de Vet HC, Westhovens R, Cella D, et al. Calibration and Validation of the Dutch-Flemish PROMIS Pain Interference Item Bank in Patients with Chronic Pain. PLoS One. 2015;10(7):e0134094.

16. Schuller W, Ostelo R, Rohrich DC, Apeldoorn AT, de Vet HCW. Physicians using spinal manipulative treatment in The Netherlands: a description of their characteristics and their patients. BMC Musculoskelet Disord. 2017;18(1):512.
17. Lamberts H, Wood, M. (Eds.). International Classification of Primary Care (ICPC). Oxford: Oxford University Press; 1987.
18. Revicki DA, Chen WH, Harnam N, Cook KF, Amtmann D, Callahan LF, et al. Development and psychometric analysis of the PROMIS pain behaviour item bank. Pain. 2009;146(1-2):158-69.
19. Roland M, Morris R. A study of the natural history of back pain. Part I: development of a reliable and sensitive measure of disability in low-back pain. Spine (Phila Pa 1976). 1983;8(2):141-4.
20. Vernon H, Mior S. The Neck Disability Index: a study of reliability and validity. J Manipulative Physiol Ther. 1991;14(7):409-15.
21. Binkley JM, Stratford PW, Lott SA, Riddle DL. The Lower Extremity Functional Scale (LEFS): scale development, measurement properties, and clinical application. North American Orthopaedic Rehabilitation Research Network. Phys Ther. 1999;79(4):371-83.
22. Hudak PL, Amadio PC, Bombardier C. Development of an upper extremity outcome measure: the DASH (disabilities of the arm, shoulder and hand) [corrected]. The Upper Extremity Collaborative Group (UECG). Am J Ind Med. 1996;29(6):602-8.
23. Kosinski M, Bayliss MS, Bjorner JB, Ware JE, Jr., Garber WH, Batenhorst A, et al. A six-item short-form survey for measuring headache impact: the HIT-6. Qual Life Res. 2003;12(8):963-74.
24. Beurskens AJ, de Vet HC, Koke AJ. Responsiveness of functional status in low back pain: a comparison of different instruments. Pain. 1996;65(1):71-6.
25. Beurskens AJ, de Vet HC, Koke AJ, van der Heijden GJ, Knipschild PG. Measuring the functional status of patients with low back pain. Assessment of the quality of four disease-specific questionnaires. Spine (Phila Pa 1976). 1995;20(9):1017-28.
26. Hoogeboom TJ, de Bie RA, den Broeder AA, van den Ende CH. The Dutch Lower Extremity Functional Scale was highly reliable, valid and responsive in individuals with hip/knee osteoarthritis: a validation study. BMC Musculoskelet Disord. 2012;13:117.
27. Jorritsma W, de Vries GE, Geertzen JH, Dijkstra PU, Reneman MF. Neck Pain and Disability Scale and the Neck Disability Index: reproducibility of the Dutch Language Versions. Eur Spine J. 2010;19(10):1695-701.
28. Jorritsma W, Dijkstra PU, de Vries GE, Geertzen JH, Reneman MF. Detecting relevant changes and responsiveness of Neck Pain and Disability Scale and Neck Disability Index. Eur Spine J. 2012;21(12):2550-7.
29. Martin M, Blaisdell B, Kwong JW, Bjorner JB. The Short-Form Headache Impact Test (HIT-6) was psychometrically equivalent in nine languages. J Clin Epidemiol. 2004;57(12):1271-8.
30. Veehof MM, Sleegers EJ, van Veldhoven NH, Schuurman AH, van Meeteren NL. Psychometric qualities of the Dutch language version of the Disabilities of the Arm, Shoulder, and Hand questionnaire (DASH-DLV). J Hand Ther. 2002;15(4):347-54.

31. Reeve BB, Hays RD, Bjorner JB, Cook KF, Crane PK, Teresi JA, et al. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). Med Care. 2007;45(5 Suppl 1):S22-S31.
32. Hu LT, Bentler P. Cutoff criteria for fit indices in covariance structure analysis: conventional criteria versus new alternatives. Structural equation modelling. 1999 1999:1-55.
33. Rodriguez A, Reise SP, Haviland MG. Applying Bifactor Statistical Indices in the Evaluation of Psychological Measures. J Pers Assess. 2016;98(3):223-37.
34. Reise SP, Scheines R, Widaman KF, Haviland MG. Multidimensionality and Structural Coefficient Bias in Structural Equation Modeling: A Bifactor Perspective. Educ Psychol Meas. 2013;73(1):5-26.
35. van der Ark LA. Mokken scale analysis in R. Journal of Statistical Software. 2007 2007.
36. van der Ark LA, Croon MA, Sijtsma K. Mokken Scale Analysis for Dichotomous Items Using Marginal Models. Psychometrika. 2008;73(2):183-208.
37. Chalmers RP. mirt: A Multidimensional Item Response Theory Package for the R Environment. J Stat Softw. 2012;48(6):1-29.
38. Amtmann D, Cook KF, Jensen MP, Chen WH, Choi S, Revicki D, et al. Development of a PROMIS item bank to measure pain interference. Pain. 2010;150(1):173-82.
39. Choi SW, Gibbons LE, Crane PK. lordif: An R Package for Detecting Differential Item Functioning Using Iterative Hybrid Ordinal Logistic Regression/Item Response Theory and Monte Carlo Simulations. J Stat Softw. 2011;39(8):1-30.
40. Crane PK, Gibbons LE, Jolley L, van Belle G. Differential Item Functioning Analysis with Ordinal Logistic Regression Techniques. Medical Care2006. p. S115-S23.
41. Kim J, Chung H, Amtmann D, Revicki DA, Cook KF. Measurement invariance of the PROMIS pain interference item bank across community and clinical samples. Qual Life Res. 2013;22(3):501-7.
42. Paz SH, Spritzer KL, Reise SP, Hays RD. Differential item functioning of the patient-reported outcomes information system (PROMIS((R))) pain interference item bank by language (Spanish versus English). Qual Life Res. 2017;26(6):1451-62.
43. Cook KF, Kallen MA, Amtmann D. Having a fit: impact of number of items and distribution of data on traditional criteria for assessing IRT's unidimensionality assumption. Qual Life Res. 2009;18(4):447-60.
44. Browne MW, Cudeck R. Alternative Ways of Assessing Model Fit. Sociol Method Res. 1992;21(2):230-58.
45. Iacobucci D. Structural equations modeling: Fit Indices, sample size, and advanced topics. J Consum Psychol. 2010;20(1):90-8.

Chapter 8.

General discussion

Introduction

This thesis was to a large extent based upon an observational cohort study, in which patient data collection was automated by the use of a custom-built, web-based register. This data was combined with data from a survey among all MSK physicians registered with their professional organisation (NVAMG), and used to evaluate two main research topics. The first topic was to characterise MSK physicians and their patient population and to measure the course of patients' conditions after MSK treatment. Baseline variables were evaluated as possible predictors of a favourable course, and the occurrence of adverse events was monitored. The second topic was to evaluate the psychometric properties of a frequently used questionnaire, the Neck Disability Index, and of the recently developed PROMIS Pain Behaviour and Pain Interference item banks.

Research TOPIC 1: Characteristics of patients and physicians in MSK medicine

Summary of findings

The results of the survey among MSK physicians showed, as expected, that they most frequently used Spinal Manipulative Treatment (SMT) techniques. This treatment was embedded in a wide array of other treatment options, part of which were legally restricted to medical doctors, such as prescription of medication or injections in the spine. MSK physicians generally had a background in other medical specialties, such as General Practice, General Surgery and Orthopaedics. Patients consulted MSK physicians with, on average, longstanding pain of moderate severity, and with moderate functional impairment. Patients frequently reported previous medical consumption in the form of specialist consultations, and treatment by physical therapists, manual therapists, chiropractors, pain clinics and surgery. Our study demonstrated that a considerable group of patients with low back pain (80%) improved in the six months after a first consultation with an MSK physician. A prediction model, based upon baseline variables from the patients' history showed an explained variance of only 9%. Adverse events were common, but generally of short duration and not severe.

MSK medicine in an international context

This thesis portrays a distinct group of medical doctors who have followed medical training to diagnose and treat complaints of the locomotor system, and who have been extensively trained to use some form of SMT. Internationally, there is a lot of variation in the position of medical doctors using SMT techniques. In the Unites States there are medical colleges where osteopathy is a part of the medical curriculum. Students graduating from these colleges are called doctor of osteopathy (D.O.), and have similar career possibilities as medical doctors. In Europe, SMT has been accepted by the European Union of Medical Specialists (UEMS) as an additional competence for medical specialists, and in many countries there are courses for medical specialists to learn SMT techniques. The UEMS refers to Manual (and Musculoskeletal) Medicine as it was defined by the Federation for Manual and Musculoskeletal Medicine (FIMM): 'Manual (and Musculoskeletal) Medicine is the medical discipline of enhanced knowledge and skills in the diagnosis, therapy and prevention of functional reversible disorders of the locomotor system. Diagnostic skills build on conventional medical techniques with manual assessment of individual tissues and functional assessment of the whole system, based on scientific biomechanical and neurophysiologic principles. Therapeutic skills add manual/manipulative techniques and advanced interventional techniques to conventional treatments for the reduction of pain or other therapeutic outcome. The therapeutic regime includes pharmacological prescription and/or manual therapy as well as rehabilitation prescription and advice. The specialist with

additional MM competence represents an appropriately trained specialist with a broad skill set otherwise only available through a multidisciplinary approach.' While the UEMS accepted MSK medicine as an additional competence for medical specialists, it is only in The Netherlands that MSK medicine is under consideration as a medical (profile) specialty in its own right.

Organisation of musculoskeletal health care

The role of MSK physicians and how they relate to other professionals in the field of musculoskeletal health care in The Netherlands is still unclear. Several professions are involved in musculoskeletal health care, ranging from allied health professions such as postural therapists, physiotherapists, manual therapists or chiropractors, to medical specialists such as orthopaedic surgeons, neurologists, neurosurgeons and rehabilitation specialists. Musculoskeletal conditions generally consist of pain or dysfunction of locomotor structures. Some conditions clearly warrant referral to an orthopaedic surgeon, such as fractures or severe osteoarthritis of peripheral joints, or referral to a neurosurgeon in the case of severe spinal stenosis. In many conditions, however, the cause of the condition is not known, and guidelines mention a variety of treatment options. This is especially true for patients with low back pain or patients with neck pain, who form a major part of the patients presented in musculoskeletal practice. Only a minority of patients present with symptoms that may warrant further diagnostic tests or referral to an orthopaedic surgeon or a neurologist. The prevalence of spinal complaints, however, is high(1), and many patients do not consult their GPs(1-3). The study by Picavet et al. reported that subjects with low back pain consulted their GPs in 31.6%, visited a medical specialist in 19.8%, and visited a physiotherapist in 26.3%. Subjects with neck pain reported GP consultation in 40.8%, visiting a medical specialist in 29.9%, and visiting a physiotherapist in 32.8%(1). In chapter 2 and 3 of this thesis it was shown that most patients consulting MSK physicians were self-referred and had previously been treated without effect by physiotherapists. Only 17% of all patients presenting in MSK practices were referred by their GPs. In primary care guidelines, SMT is generally advocated as part of a number of possible interventions. Other treatment options are reassurance and the advice to stay active, exercise treatment, clinical massage or postural corrections, but also muscle relaxants and non-steroid anti-inflammatory drugs. If non-invasive treatments were not effective, epidural steroid injections, facet denervation or surgery may be indicated(4-8). The guidelines of the Dutch Federation of Medical Specialists suggest pain intervention as a treatment option for patients with low back pain and or lumbar radiculopathies(9, 10). With this range of possible interventions, specialist knowledge of, and experience with specific diagnostic possibilities, such as imaging and electromyography, and specific treatment options such as surgery or injections is of added value in advising patients with spinal complaints. This type of knowledge is typically included in the medical

curriculum, and on a more specialist level represented in the training programme for MSK physicians. Our survey showed that most MSK physicians have previous experience in relevant specialities, and the training programme to become registered as MSK physician includes relevant specialist medical knowledge. On the one hand, combining their array of diagnostic and treatment possibilities distinguishes MSK physicians from manual therapists and chiropractors, while on the other hand, combining specialist medical knowledge with SMT distinguishes MSK physicians from neurologists, orthopaedic surgeons and pain clinic anaesthetists. MSK physicians are well suited to have an important role in health care for patients with musculoskeletal complaints, in close cooperation with (extended scope) physical therapists, manual therapists and chiropractors, but also with neurologists, orthopaedic surgeons and rehabilitation specialists.

Patient perspective

With the GP as the first-line consultant, healthcare for patients with musculoskeletal complaints in The Netherlands is still organised as patchwork, and patients are frequently left to find their way through self-referral(11-13). My personal experience is that patients who consult MSK physicians have long-standing complaints and have, through the years, consulted a multitude of healthcare professionals. Many low back pain patients, though, have given up consulting their GPs about their complaint(14). Typically, these patients were referred to neurologists, who have made MRIs and stated that the patient had a disk problem, but not severe enough to warrant surgery; they have consulted orthopaedic surgeons, who stated that there were degenerative changes; they were treated by physiotherapists, manual therapists and chiropractors, who provided temporary relief but whose treatments became less effective over the years. Patients have experienced these consultations as conflicting information(14), which never led to an effective treatment strategy. Patients generally appreciate it when they are included in complex clinical reasoning, even if this brings about uncertainties about the cause of their pain. This clinical reasoning may include the interpretation of MRIs, potential pain interventions, the possibilities of surgery, the problem of sensitisation or hypothetical mechanisms that may explain the workings of SMT.

Thus, the fact that MSK physicians combine a full-fledged medical education with more specialist knowledge of neurology, orthopaedics and pain medicine as well as thorough training in SMT enables them to contextualise the different findings patients have accumulated over the years, discuss various treatment possibilities and present a step-by-step treatment plan. Such a patient-centred approach, I find, is much appreciated. Perhaps this is one of the reasons why patient satisfaction with MSK medicine is high. In the second phase of our data collection we evaluated patient satisfaction with MSK physicians. These data have not been published to date. Of 1505 patients who answered a baseline questionnaire, 1138

(76%) answered a patient satisfaction questionnaire three months after consulting an MSK physician. Of these 1138 patients, 91.9% scored at least 7 on a 0-10 NRS (average 8.4) measuring patient satisfaction with the treatment, and 91.7% of the 1138 patients would advise other patients with similar complaints to consult an MSK physician as well. A recent CBS evaluation reported comparable patient satisfaction for physiotherapists and postural therapists (8.1), and somewhat lower satisfaction with GP care (7.8)(15). Our results could be biased by selective loss to follow-up, but analyses of our cohort in chapter five and six showed that most patients who did not respond reported other reasons than dissatisfaction to discontinue their participation. In fact, most patients were not interested to participate in the study because their complaints were resolved. Besides, even if most non-responders would be dissatisfied, overall patient satisfaction would remain high.

A patient centred approach

A stratified approach within specialised centres where multidisciplinary treatment possibilities can be combined may improve quality of care whilst reducing costs(16). Dutch GP standards and primary care guidelines mention possible multidisciplinary treatment in a rehabilitation setting for patients who do not benefit from primary care interventions, while specialist guidelines mention pain interventions as possible treatment options. MSK physicians may have the tools to serve as specialist consultant for musculoskeletal complaints, and they are well-equipped to guide multidisciplinary treatment. There are already initiatives, such as the Spineclinic and the Rugpoli, where MSK physicians treat patients in a multidisciplinary setting, in close corporation with, for example, trainers, physical therapists, manual therapists or McKenzie therapists. In the 'Rugpoli' (back treatment centre) patient centred care for spinal complaints is organised around MSK physicians, in a protocol that encompasses Mechanical Diagnosis and Treatment (MDT)(17-19), SMT, and injections in the spine under X-ray guidance(20-22). Within the Rugpoli protocol, only patients in whom all conservative treatment options have been unsuccessful, including SMT and McKenzie, may be treated with injections in the spine, using the protocol of the Society for Interventions in the Spine (SIS)(23). Such a selection of patients may well enhance the efficacy of injection treatment. A study in which patients who were on a waiting list for disc surgery were treated through the Rugpoli protocol showed that in 78% of these patients the complaints improved to such an extent that an operation was no longer necessary(22). A trial to evaluate the Rugpoli protocol is currently being carried out(16).

Strengths and weaknesses of studies on TOPIC 1

In this thesis a clear description was given of the characteristics of MSK physicians in The Netherlands. A wide range of baseline variables was measured, and the course of patients' complaints in the six months after consulting MSK physicians was used to define groups

of patients with distinct pain trajectories. A weakness of this thesis is the non-response of both MSK physicians and patients. In a number of patients physicians failed to answer the treatment variables at follow-up, and patients frequently failed to answer some of the follow-up questionnaires. The chance of bias due to loss to follow-up could be estimated because baseline variables were collected from both patients and physicians, and because patients who discontinued their participation were asked about their motivation. Another weakness is the absence of more clinical variables, which could have been included in the prediction model. Clinical measurements were not included in our study because it would increase the burden for the participating physicians and this was expected to reduce the recruitment rates. Besides, as the mechanism underlying SMT is unknown, it is not evident which clinical measurements could be of value. Biomechanical studies, for example, have shown decreased motor control in patients with low back pain, but it is not known yet whether measuring motor control would help in detecting patients for specific interventions(24, 25). In addition, such biomechanical studies are complicated, need high-tech equipment, and have no standardised protocols.

Further development and research

Although increasingly accepted as a possible treatment option(4, 5, 7, 26), there is still a need for a more thorough scientific basis to support the use of SMT. While it is thought that some patients may benefit strongly from SMT treatment(27), is appears difficult to find variables that may help to select patients with a favourable prognosis. In the past decades, a lot of research has focussed on identifying patients who may benefit from SMT, without promising results. In our study we constructed a prediction model for a favourable prognosis after treatment by MSK physicians, which almost invariably included the use of SMT techniques. In chapter 3 of our study a variety of baseline characteristics were evaluated as possible predictors of a favourable course, the prediction model based on these data only showed an explained variance of 9%. The clinical value of this prediction model mainly lies in the realisation that none of these baseline characteristics can be used to identify patients with a favourable prognosis. I would suggest that there are two possible ways to move forward. The first possibility is to design escalating treatment protocols. Start with simple (and cheap) treatments, and move forward with more complex and costly interventions when there is no effect. The Dutch CBO guideline is an example of a step-by-step approach for patients with low back pain. Rather than a patchwork of often coincidental treatment options, patients are guided through consecutive interventions. Instead of studying separate treatment options with randomised trials, these protocols could be evaluated for their cost-effectiveness. The STarT back screening tool, for example, has been developed for such purposes in general practice(28, 29), and its cost-effectiveness has been studied in a large trial(28). The second possibility is to study clinical measurements as predictors of a favourable course.

Clinical measurements are also needed to gain insight into the mechanism explaining SMT, which is still unknown. Current hypotheses include biochemical, neurophysiological and biomechanical processes, or a combination of these(25, 30, 31).

Hypothesis

A biomechanical hypothesis derived from the SMT techniques used by Dutch MSK physicians was presented at international conferences in Oslo, Antwerp and Liverpool, with an inventory into possibilities for further study(32). In this hypothesis, the spine and the pelvic girdle are viewed as a construction composed of structural bony shaped elements and tensile ligamentous structures. Disturbances in this tensile construction will lead to local disturbances in spinal load, and eventually to overuse and pain. An important component of this hypothesis is the assessment of pelvic unleveling, which is considered to be a function of the iliac bones, the SI joints, and the L4 and L5 lumbar vertebrae. In an unlevel pelvis, the iliac bones rotate around a transverse axis. The iliac bones rotate in a ventro-cranial direction on the higher side and a dorso-caudal direction on the lower side. Due to the anatomical shape of the sacro-iliac joints, the sacrum shifts laterally from the ventro-cranially rotated ilium, combined with an anterior rotation on the lower side. Due to stretching forces of the iliolumbar ligaments, L4 and L5 show a similar rotation around the AP axis towards the lower side.

In practice, it seems as though an unlevel pelvis with L5-S1 pain leads to activation of muscles in support of the spinal joints by tensing both the erector muscles and the iliopsoas. The spine becomes stiffer, and the fine coordination of movement is disturbed, eventually leading to hypotrophy of the multifidus muscle. Similarly, dysfunction of the SI joint leads to reflectory tension in the glutaeus medius, minimus and tensor fasciae latae (the abductor chain), and the piriformis muscle. This may well mimic hip-joint pathology, or even radicular pain. If this situation exists for longer periods of time, further changes occur in the quality of the muscle tissue and in the motor control, with subsequent changes in the neurophysiologic mechanisms concerning motor signals and pain. With the SMT technique used by Dutch MSK physicians an unlevel pelvis is corrected in a strict sequence of specific mobilising techniques. To further study such hypothetic mechanisms, fundamental research is necessary. For example by assessing the level of the iliac crest with more objective measures. Measurements developed in these fundamental studies may eventually help to select patients suitable for SMT treatment, and support the development of multidisciplinary protocols within MSK medicine.

Research TOPIC 2: Clinimetric studies on measurement instruments

Summary of findings

In this thesis psychometric properties of the Neck Disability Index (NDI) were evaluated. Different methods to calculate the Minimal Important Change (MIC) and the Smallest Detectable Change (SDC) were compared. The MIC and the SDC were not influenced by the method used to calculate these properties, but they were influenced by population characteristics. Furthermore, the validity of the PROMIS Pain Behaviour and Pain Interference item banks was evaluated. Although the fit of individual items to the IRT model was generally good, it was shown that both item banks did not strictly measure a single construct. Further evaluation for the Pain Interference item bank showed that this did not seem to lead to biased scores, and it was concluded that this item bank appeared to be valid and can be used in clinical studies. Further evaluation for the Pain Behaviour item bank, however, showed a risk of biased scores due to multidimensionality, which would render its validity questionable.

Discussion of psychometric properties of the NDI

Chapter five concerned the method in which clinically important change can be estimated. Defining clinically important change is still a much debated issue. Currently there are initiatives to standardise the methods used to evaluate these properties. While it is appealing to adopt a single value which can distinguish between patients improved or not, defining such a value appears to be difficult. Various studies report a wide range of estimated values for the MIC and SDC. Our study contributes to this discussion by showing that estimated values of the MIC and SDC were not influenced by the methods used, but rather by the population studied. This is not completely surprising. To estimate the MIC, anchor-based approaches are used in which, after a follow-up period, patients are asked whether they consider their condition to be improved. This global perceived effect (GPE) is then compared to the change measured with the questionnaire at issue. Previous studies have shown that this GPE is correlated more with the present state than with the change in the condition of the patient(33). In other words: when patients are asked whether they have improved after a treatment their answer is influenced by their momentary condition. This would mean that patients who start the treatment with minor complaints will have minor complaints at follow-up, even when the improvement has been minimal, and may judge their condition as much improved. Patients who start with more severe complaints will need a stronger improvement to arrive at a present state with little complaints. Patients who had more severe complaints before the treatment thus need more improvement before they would consider themselves to be improved. Other patient characteristics may influence MIC and SDC estimates as well; and are listed in chapter five of this thesis. The influence of patient

characteristics on estimates of the MIC and the SDC will not be solved by standardising the method of calculating these estimations.

Discussion of the validity of PROMIS item banks

Chapters 6 and 7 concerned the validity of the recently developed PROMIS Pain Behaviour and Pain Interference item banks. One of the advantages of IRT based item banks is the fact that they score constructs on a single metric. Assumptions underlying IRT are unidimensionality, monotonicity and local independence. To scale item banks on a single metric it is necessary that they measure a single construct. Unidimensionality, therefore, is an important assumption of item banks, and problems with the dimensionality were the main issue in the discussions in chapters 6 and 7. For the PROMIS Pain Interference and Pain Behaviour item banks unidimensionality was assessed with the scaled CFI, the scaled TLI and the scaled RMSEA fit indices. Both item banks showed fit indices that were below the level defined to assume unidimensionality, which was not in line with previous reports about both item banks. The studies in which both newly developed item banks were first presented reported sufficient evidence to assume unidimensionality(34, 35), and two Dutch studies reporting psychometric properties of both item banks presented fit-indices supporting unidimensionality as well(36, 37). There are several possible explanations for this discrepancy. First of all, in the Dutch studies unscaled fit-indices were used which do not offer a correction for overestimation due to the non-normal distribution of data. Because normality cannot be assumed we used scaled indices, which presented lower values of the CFI and TLI fit indices, and higher values for the RMSEA index, not supportive of unidimensionality. In both studies in which the item banks were first presented, it was not mentioned whether the indices used were scaled or not. Only the study presenting the Pain Interference item bank mentioned the statistical software used (Multilog), which, to our knowledge, only presents scaled indices(34). The study in which the original Pain Behaviour item bank was presented reported indices that only partially supported unidimensionality, without mentioning the statistical software used. Furthermore, this study evaluated the fit only in a model including 52 candidate items, of which later in the process 13 items were omitted due to item fit issues(35). Combining the results of these studies we would consider unidimensionality questionable. When unidimensionality cannot be established it is important to know whether this will have an influence on the scoring. Bi-factor analyses can be used for this purpose(38). From a bi-factor analyses the Explained Common Variance (ECV) and Omega-H can be calculated, which indicate the risk of biased scores due to multidimensionality(38, 39). Bi-factor analysis improved the fit-indices for the Pain Interference item bank, and ECV and Omega-H were high, indicating that the risk of biased scores due to multidimensionality was low. For the Pain Behaviour item bank bi-factor analyses did not improve the fit sufficiently, but the ECV and Omega-H were high. Although the high ECV and Omega-H would indicate a low

risk for biased scores due to multidimensionality the suboptimal fit of the bi-factor model makes this conclusion questionable. It was concluded that the Pain Interference item bank appeared to be valid and can be used in clinical studies, but the validity of the Dutch-Flemish Pain Behaviour item bank needs further evaluation in other populations.

Another issue in IRT-based item banks is the evaluation of Differential Item Functioning (DIF). DIF analyses reveal whether different groups of patients answer individual items differently. This would mean that the answers of patients on these items, and thus the scores of these patients cannot be compared to each other. DIF analyses revealed only a few items with DIF for language, with minor influence on scores when the whole item bank was used. Eventually, however, the whole item bank will rarely be used. Most item banks will be used in short forms or Computer Adaptive Testing (CAT). In CAT only a few items are necessary to obtain a reliable score. If DIF items are included in these short forms or in CAT, considerable differences could be the result, and comparing groups of patients who have presented DIF may end up being invalid. Besides DIF for language as a measure of cross-cultural validity we tested DIF for various age groups and for sex. For the Pain Behaviour item bank one item showed DIF for gender: PAINBE27 (I had pain so bad it made me cry) was answered affirmatively by female patients at lower levels of Pain Behaviour, and one item showed DIF for age: PAINBE29 (When I was in pain I used a cane or something else for support). Older patients were inclined to answer this item affirmatively at lower levels of Pain Behaviour. For the Pain Interference item bank there were no DIF items besides one item showing DIF for language: PAININ24 (How often was pain distressing to you?) was answered affirmatively at lower levels of pain interference by the Dutch population.

Strengths and weaknesses of studies on TOPIC 2

A strength of this thesis was the large population studied, with a wide range of musculoskeletal complaints. It was one of the first studies in which scaled fit-indices were presented, and in which dimensionality was further evaluated with a bi-factor model. A minor weakness could be the fact that the patient population recruited may have included a low number of patients with high levels of pain. Therefore there were more patients with lower scores. Answer categories with a small number of patients were collapsed with the adjacent category and therefore lack a threshold parameter.

Further development and study

Although the PROMIS initiative has presented a solid bases for the use of IRT item banks in health care(40), further validation is necessary, with special attention for the aforementioned issues concerning dimensionality. Other issues include the choice of the calibration

parameters, the feasibility of calibrating measurements of clinical constructs on a general population, and the influence of DIF on measurement with short forms or CAT.

There has been a lot of discussion about the choice of the item parameters used for scoring PROMIS item banks. The PROMIS item banks were calibrated in a US population, supplemented with patients with specific conditions and the scale was subsequently centred on the mean and SD of the US general population to create a metric. The US parameters (metric) have been put forward as the standard set of parameters to use when calculating T-scores. The question has been raised, though, whether it is necessary to have a standard set of item parameters, and if so, how this standard set should be decided upon(41). While the US studies were conducted with the purpose to create sets of calibration parameters, the US calibration is put forward as the standard set of parameters. It is not yet clear, though, if this standard set is optimally valid in countries outside of the US. To have an international common metric may be important when the results of studies in other countries are compared to the results of US studies, but this may be seldom the case, and it can be argued that local calibration may offer better measurement in the local country. More research is recommended on the effects of the choice of item parameters on T-scores and whether this has any influence on clinical decisions.

Another issue, not much debated, is whether calibrating typical clinical constructs, like pain, in a general population is useful to begin with. Doing so has already lead to problems in IRT analyses of the PROMIS Pain Behaviour item bank. Each item of the pain behaviour item bank has six response categories, the first of which is the option: 'had no pain'. In the General Response Model (GRM), used for item banks with ordinal response categories, two parameters are estimated for each item: the slope and the threshold parameter. The threshold parameter indicates the level of the construct that optimally fits the transition of one item category to the next. The slope parameter indicates the strength of the relationship between the item and the construct. The threshold parameters are estimated for each item category separately. The slope parameter is calculated for each item, but not for each transition between item categories. In the development of the Pain Behaviour item bank a large number of subjects were included who had no pain. These subjects would all answer 'had no pain' on most of items. It would thus seem in the analyses as if there was a very strong relation between these items and the construct, with very high slope parameters. Slope parameters were 'inflated' because one item category was predominant in a large part of the population. In fact, these slope parameters did not depict a strong relation between the items and the construct, but rather the fact that most subjects did not demonstrate the construct tested(35). To correct these analyses, the authors decided to remove subjects without pain from the analyses. While this is understandable, it does not calibrate the item

bank on a general population. Because of the problems shown by the Pain Behaviour item bank, a new version was presented in which the calibration was limited to patients with pain, rather than calibrating the item bank on a general population(42). Further study should reveal whether it is feasible to calibrate clinical constructs on a general population.

Issues with DIF were already mentioned in the discussion. DIF is an important psychometric property for IRT-based instruments, i.e. measurement invariance. Especially in instruments that are meant to be generic, and that are expected to be applicable in a variety of conditions, it is perhaps to be expected that patients with different conditions may interpret items differently. Low back pain patients, for example, may interpret a question about a certain limitation different than neck pain patients. As an example we studied DIF between patients with upper body pain (neck or arms), and patients with lower body pain (lower back or legs) - these analyses were not included in chapter 5 or 6. This revealed DIF for PAINBE31 ("I limped because of the pain"), PAINBE38 ("When I was in pain I drew my knees up") and PAINBE43 ("When I was in pain I walked carefully"). Intuitively, this is very understandable, one can easily imagine patients limping because of lower body pain complaints, while patients with upper body pain would express different pain behaviour issues. Most studies report some DIF, but state that the influence of DIF on scores is limited when the whole item bank is used. This statement may not be very relevant, however, because part of the appeal of item banks is the use of short forms or CAT. Further study will have to reveal the influence of DIF when using short forms or CAT. It may be necessary to select items without DIF, or to replace DIF items with similar items without DIF. Offering a correction of scores due to DIF would be complex, because different groups of patients will probably display different DIF items. Although these DIF issues should be solved, one must bear in mind that problems only arise when the scores of different groups of patients are compared and not when specific groups of patients are followed over time. Apart from this, adding items for specific populations to existing item banks will make issues with multidimensionality and DIF more complex. All these solutions will make IRT measurement more complex and less generic. For the time being, there is sufficient support to use IRT based measurement tools and CAT, but I would expect further developments in their analyses and their use.

Conclusion

This thesis presented the results of a large research project in which two main topics were evaluated. Both topics were exiting in the fact that they concerned recent developments that may well become more widely used in the near future. MSK medicine may well develop towards a more established profession in the Dutch health care system, and could fuel more research studying the usefulness and the working mechanism of SMT, especially when

combined in multidisciplinary protocols or combined with pain clinic interventions. IRT-based item banks and CAT are the future of patient reported outcome measurement, and will be routinely used in many health care settings. Routine measurement will be much easier when only a few questions need to be answered to arrive at a reliable score. The PROMIS initiative is presently taking the lead in the development and use of IRT-based measurement. It will be interesting to see the further development of IRT based measurement instruments.

References

1. Picavet HS, Schouten JS. Musculoskeletal pain in the Netherlands: prevalences, consequences and risk groups, the DMC(3)-study. Pain. 2003;102(1-2):167-78.
2. Foster NE, Hartvigsen J, Croft PR. Taking responsibility for the early assessment and treatment of patients with musculoskeletal pain: a review and critical analysis. Arthritis Res Ther. 2012;14(1):205.
3. Nederlands Huisartsen Genootschap. NHG Standaard Aspecifieke Lagerugpijn 2017
4. Cohen SP, Hooten WM. Advances in the diagnosis and management of neck pain. BMJ. 2017;358:j3221.
5. Cote P, Yu H, Shearer HM, Randhawa K, Wong JJ, Mior S, et al. Non-pharmacological management of persistent headaches associated with neck pain: A clinical practice guideline from the Ontario protocol for traffic injury management (OPTIMa) collaboration. Eur J Pain. 2019;23(6):1051-70.
6. Reid SA, Callister R, Snodgrass SJ, Katekar MG, Rivett DA. Manual therapy for cervicogenic dizziness: Long-term outcomes of a randomised trial. Man Ther. 2015;20(1):148-56.
7. Wong JJ, Cote P, Sutton DA, Randhawa K, Yu H, Varatharajan S, et al. Clinical practice guidelines for the noninvasive management of low back pain: A systematic review by the Ontario Protocol for Traffic Injury Management (OPTIMa) Collaboration. Eur J Pain. 2017;21(2):201-16.
8. Hurwitz EL, Carragee EJ, van der Velde G, Carroll LJ, Nordin M, Guzman J, et al. 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 (Phila Pa 1976). 2008;33(4 Suppl):S123-52.
9. Federatie Medisch Specialisten. Richtlijn Lumbo-Sacraal Radiculair Syndroom. 2008.
10. Federatie Medisch Specialisten. Wervelkolom gerelateerde pijnklachten van de lage rug. 2012.
11. Rubinstein S, Pfeifle CE, van Tulder MW, Assendelft WJ. Chiropractic patients in the Netherlands: a descriptive study. J Manipulative Physiol Ther. 2000;23(8):557-63.
12. Schuller W, Ostelo R, Rohrich DC, Apeldoorn AT, de Vet HCW. Physicians using spinal manipulative treatment in The Netherlands: a description of their characteristics and their patients. BMC Musculoskelet Disord. 2017;18(1):512.
13. Swinkels IC, Kooijman MK, Spreeuwenberg PM, Bossen D, Leemrijse CJ, van Dijk CE, et al. An overview of 5 years of patient self-referral for physical therapy in the Netherlands. Phys Ther. 2014;94(12):1785-95.
14. Lim YZ, Chou L, Au RT, Seneviwickrama KMD, Cicuttini FM, Briggs AM, et al. People with low back pain want clear, consistent and personalised information on prognosis, treatment options and self-management strategies: a systematic review. J Physiother. 2019;65(3):124-35.
15. Centraal Bureau voor de Statistiek. Tevredenheid zorgverlener 2018 2019 [Available from: https://www.cbs.nl/nl-nl/nieuws/2019/48/nederlanders-tevreden-over-zorgverleners.
16. Mutubuki EN, van Helvoirt H, van Dongen JM, Vleggeert-Lankamp CLA, Huygen F, van Tulder MW, et al. Cost-effectiveness of combination therapy (Mechanical Diagnosis and Treatment and

Transforaminal Epidural Steroid Injections) among patients with an indication for a Lumbar Herniated Disc surgery: Protocol of a randomized controlled trial. Physiother Res Int. 2020;25(1):e1796.
17. Berthelot JM, Delecrin J, Maugars Y, Passuti N. Contribution of centralization phenomenon to the diagnosis, prognosis, and treatment of diskogenic low back pain. Joint Bone Spine. 2007;74(4):319-23.
18. Dunsford A, Kumar S, Clarke S. Integrating evidence into practice: use of McKenzie-based treatment for mechanical low back pain. J Multidiscip Healthc. 2011;4:393-402.
19. Machado LA, de SM, Ferreira PH, Ferreira ML. The McKenzie method for low back pain: a systematic review of the literature with a meta-analysis approach. Spine (Phila Pa 1976). 2006;31(9):E254-E62.
20. Baker RM. International Spine Intervention Society (ISIS) presidential address: 20th Annual Scientific Meeting. Wednesday, July 18, 2012. Pain Med. 2012;13(9):1108-9.
21. Bogduk N. International Spinal Injection Society guidelines for the performance of spinal injection procedures. Part 1: Zygapophysial joint blocks. Clin J Pain. 1997;13(4):285-302.
22. van Helvoirt H, Apeldoorn AT, Ostelo RW, Knol DL, Arts MP, Kamper SJ, et al. Transforaminal epidural steroid injections followed by mechanical diagnosis and therapy to prevent surgery for lumbar disc herniation. Pain Med. 2014;15(7):1100-8.
23. Bogduk N. Practice Guidelines for Spinal Diagnostic and Treatment Procedures. Hinsdal, IL: International Spine Intervention Society; 2004.
24. Srinivasan D, Mathiassen SE. Motor variability--an important issue in occupational life. Work. 2012;41 Suppl 1:2527-34.
25. Tong MH, Mousavi SJ, Kiers H, Ferreira P, Refshauge K, van Dieen J. Is There a Relationship Between Lumbar Proprioception and Low Back Pain? A Systematic Review With Meta-Analysis. Arch Phys Med Rehabil. 2017;98(1):120-36 e2.
26. Koes BW. [Manual therapy for neck pain: increasing evidence for effectiveness]. Ned Tijdschr Geneeskd. 2012;156(15):A4599.
27. Deyo RA. The Role of Spinal Manipulation in the Treatment of Low Back Pain. JAMA. 2017;317(14):1418-9.
28. Hill JC, Whitehurst DG, Lewis M, Bryan S, Dunn KM, Foster NE, et al. Comparison of stratified primary care management for low back pain with current best practice (STarT Back): a randomised controlled trial. Lancet. 2011;378(9802):1560-71.
29. Hill JC, Dunn KM, Lewis M, Mullis R, Main CJ, Foster NE, et al. A primary care back pain screening tool: identifying patient subgroups for initial treatment. Arthritis Rheum. 2008;59(5):632-41.
30. Freeman MD, Woodham MA, Woodham AW. The role of the lumbar multifidus in chronic low back pain: a review. PM R. 2010;2(2):142-6; quiz 1 p following 67.
31. Salzberg L. The physiology of low back pain. Prim Care. 2012;39(3):487-98.
32. Schuller W, Noordzij, J., Huetink, K., Hoogland, P. A Biomechanical Model of Pelvic Displacement. Back and Neck Forum; Oslo2019.

33. Kamper SJ, Ostelo RW, Knol DL, Maher CG, de Vet HC, Hancock MJ. Global Perceived Effect scales provided reliable assessments of health transition in people with musculoskeletal disorders, but ratings are strongly influenced by current status. J Clin Epidemiol. 2010;63(7):760-6.
34. Amtmann D, Cook KF, Jensen MP, Chen WH, Choi S, Revicki D, et al. Development of a PROMIS item bank to measure pain interference. Pain. 2010;150(1):173-82.
35. Revicki DA, Chen WH, Harnam N, Cook KF, Amtmann D, Callahan LF, et al. Development and psychometric analysis of the PROMIS pain behaviour item bank. Pain. 2009;146(1-2):158-69.
36. Crins MH, Roorda LD, Smits N, de Vet HC, Westhovens R, Cella D, et al. Calibration and Validation of the Dutch-Flemish PROMIS Pain Interference Item Bank in Patients with Chronic Pain. PLoS One. 2015;10(7):e0134094.
37. Crins MH, Roorda LD, Smits N, de Vet HC, Westhovens R, Cella D, et al. Calibration of the Dutch-Flemish PROMIS Pain Behaviour item bank in patients with chronic pain. Eur J Pain. 2016;20(2):284-96.
38. Rodriguez A, Reise SP, Haviland MG. Applying Bifactor Statistical Indices in the Evaluation of Psychological Measures. J Pers Assess. 2016;98(3):223-37.
39. Reise SP, Scheines R, Widaman KF, Haviland MG. Multidimensionality and Structural Coefficient Bias in Structural Equation Modeling: A Bifactor Perspective. Educ Psychol Meas. 2013;73(1):5-26.
40. Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. J Clin Epidemiol. 2010;63(11):1179-94.
41. Crins MHP. Promising PROMIS. Amsterdam: Amsterdam UMC; 2020.
42. Cook KF, Keefe F, Jensen MP, Roddey TS, Callahan LF, Revicki D, et al. Development and validation of a new self-report measure of pain behaviours. Pain. 2013;154(12):2867-76.

Chapter 9.

General summary

Introduction

This thesis represents a large research effort that was for most part funded by the Dutch Association for Musculoskeletal (MSK) Medicine. The background of this research effort was explained in the introduction (chapter 1). Main reason for this project was that there was no research yet in which the characteristics of MSK physicians and of their patient population were assessed. The backbone of the research project was an automated data collection system that enabled a large observational cohort study at limited costs. Data from this data collection system were used in all studies, except the study in which the psychometric properties of the neck Disability Index (NDI) were assessed. For the data collection system a custom build program was designed in which patients' email addresses were used to invite patients to answer questionnaires. Physicians would enter baseline characteristics, such as the main complaint, concomitant complaints, duration of the main complaint and age of all patients after a first consultation in a web-based register. At the end of the completed treatment they would enter more details about the treatment used, such as the type of treatment and the number of treatment sessions. Patients were recruited to answer baseline questionnaires, and follow-up questionnaires at various intervals after the first consultation. In four different phases of the research project a variety of patient questionnaires were used, with different follow-up moments. With this automated web-based register a large amount of data was collected that was used to evaluate a number of research questions. In this thesis studies are presented in which the characteristics of MSK physicians and of their patient population were evaluated. Other studies evaluated the course of low back pain after treatment by MSK physicians, and the occurrence of adverse events reported by patients. Furthermore, the web-based registry was used to assess the validity of two patient reported outcome measurement instruments (PROMS) that can be used to evaluate the complaints of patients with musculoskeletal complaints.

Chapter 2

In chapter 2 the results were presented of a study in which the characteristics of MSK physicians and of their patient population were assessed. Data about the characteristics of MSK physicians was collected with a survey and with telephone interviews. Questionnaires were sent to 138 physicians of whom 90 responded (65%). Most physicians were trained in MSK medicine after a career in other medical specialities. They reported to combine their SMT treatment with a variety of diagnostic and treatment options part of which were only permissible for physicians, such as prescription medication and injections. Data about patient characteristics was extracted from the web-based register. The majority of patients presented with complaints of long duration (62.1% > 1 year), most frequently low back pain

(48.1%) or neck pain (16.9%), with mean scores of 6.0 and 6.2, respectively, on a 0 to10 numerical rating scale (NRS) for pain intensity. Mean scores on all PROMs showed moderate impairment. Patients most frequently reported previous treatment by physical therapists (68.1%), manual therapists (37.7%) or chiropractors (17.0%). It was concluded that MSK physicians in The Netherlands reported to use an array of SMT techniques. They embedded their SMT techniques in a variety of other diagnostic and treatment options, part of which were limited to medical doctors.

Chapter 3

In chapter 3 the results were presented of a study assessing the course of low back pain (LBP) after treatment by MSK physicians. Data was collected using the web-based register. MSK physicians recorded various baseline and treatment variables. Patient questionnaires included information about previous medical consumption, together with PROMs measuring the level of pain and functional status at baseline, and at 6-weekly intervals during a follow-up period of six months. Latent Class Growth Analysis (LCGA) was used to classify patients into different groups according to their pain trajectories. Baseline variables were evaluated as predictors of a favourable trajectory using logistic regression analyses. In a period of two years 1377 Patients were recruited, of whom 1117 patients (81%) answered at least one follow-up measurement. LCGA identified three groups of patients with distinct pain trajectories. A first group (N=226) with high pain levels showed no improvement, a second group (N=578) with high pain levels showed strong improvement, and a third group (N=313) with mild pain levels showed moderate, but clinically relevant improvement. The two groups of patients presenting with high baseline pain scores were compared, and a prediction model of a favourable course was constructed. Male gender, previous specialist visit, previous pain clinic visit, having work, a shorter duration of the current episode, and a longer time since the complaints first started were predictors of a favourable course. The prediction model showed a moderate area under the curve (0.68) and a low explained variance (0.09). It was concluded that a large proportion of patients with low back pain improved after treatment by MSK physicians , but that the clinical value of the prediction model presented will be limited.

Chapter 4

In chapter 4 the results were presented of a study in which the occurrence of adverse events after treatment by an MSK physician was assessed in patients with low back pain (LBP) or neck pain (NP). MSK physicians recorded various baseline and treatment variables. Patients were asked to answer questionnaires at baseline including PROMs measuring the level

of pain and functional status. Three months after the start of the treatment, patients were invited to answer questionnaires enquiring after the type, the severity, and the duration of adverse events. A total of 1391 LBP patients and 549 NP patients answered the baseline questionnaire, of whom 823 (59%) LBP and 315 (57%) NP patients answered the adverse events questionnaire. Of these patients, 362 (31.8%) reported a total of 683 adverse events. All patients except five were treated with a manipulative or mobilising technique, or both, in, on average, 3-6 sessions (range 1-12). The highest proportion of patients (15.8%) reporting any adverse event reported only one adverse event, and the adverse event most frequently reported was fatigue (10.9% of all patients). Patients with a main complaint of NP reported adverse events more frequently (38.4%) than patients with a main complaint of LBP (29.3%). Most adverse events were not severe and resolved within a week, but some patients reported adverse events to be more severe (6.9%) or lasting longer (7.1%). It was concluded that adverse events after spinal manipulative treatment by musculoskeletal physicians were common but generally short-lived and not severe. Neck pain patients displayed different adverse events than low back pain patients. Patients in whom the neck had been treated with a mobilising technique more frequently reported adverse events, compared to patient in which a manipulative technique was used, which was largely due to the frequent reporting of fatigue. There was no relation between the report of adverse events and the reported improvement after three months follow-up.

Chapter 5

In chapter 5 the results are presented of a study in which methods to calculate Smallest Detectable Change (SDC) and Minimal Important Change (MIC) were evaluated for the NDI. In a cohort study 101 patients with chronic neck pain were recruited, who were asked to answer the NDI at baseline, and after a follow-up period of six months. SDC and MIC were calculated using two types of external anchors. For each anchor we applied two different definitions to dichotomise the population in a group of improved and a group of unimproved patients. The influence of patient characteristics was assessed in relevant subgroups: patients with or without radiating pain, patients with or without concomitant headache and patients with high or low baseline scores. It was shown that different anchors and different definitions of improvement hardly influenced estimates of the SDC and the MIC. The SDC and the MIC were similar for subgroups of patients with or without radiation, but differed strongly for subgroups of patients with or without concomitant headache and for patients with high or low baseline scores. It was concluded that the SDC and the MIC are not an invariable characteristic of the NDI but are influenced by patient characteristics.

Chapter 6

In chapter 6 the results were presented of a study in which the validity of the v1.1 Dutch-Flemish PROMIS Pain Behaviour item bank was assessed in a sample of 1602 patients with musculoskeletal complaints. Assumptions of the underlying Item Response Theory were evaluated in a Grade Response Model (GRM): unidimensionality and local dependency with Confirmatory Factor Analyses (CFA), and monotonicity with scalability coefficients. IRT-model fit of all items was evaluated, and item parameters were estimated. Differential Item Functioning (DIF) was studied for age and gender, and DIF for language was studied as a measure of cross-cultural validity. The GRM showed suboptimal fit of a unidimensional model (CFI: 0.816, TLI: 0.806, RSMEA: 0.093), and fifteen item pairs (2%) with local dependence. Five items showed poor scalability (Mokken $H_{(i)}$: 0.14-0.41). A bi-factor model showed low risk of bias when a unidimensional model was assumed (Omega-H 0.92, Explained Common Variance (ECV) 0.70), but the fit of the bi-factor model was still suboptimal (CFI: 0.922, TLI: 0.915, RSMEA: 0.062), with 3 item pairs showing local dependence (0.4%). All items fitted the IRT model; slope parameters ranged from 0.60 to 2.00, and threshold parameters from -2.05 to 6.80. One item showed DIF for age, one item DIF for gender, and five items showed DIF for language, but the impact on total scores was low. It was concluded that the DF-PROMIS-Pain Behaviour item bank can be used in clinical research and in clinical practice, although further research should examine whether problems concerning dimensionality and monotonicity occur in other populations.

Chapter 7

In chapter 7 the results were presented of a study in which the validity of the v1.1 Dutch-Flemish PROMIS Pain Interference item bank was evaluated in a population of 1677 patients with musculoskeletal complaints. Assumptions of the underlying Item Response Theory were evaluated in a Graded Response Model (GRM): unidimensionality and local dependency with Confirmatory Factor Analyses (CFA), and monotonicity with scalability coefficients. IRT-model fit of all items was evaluated, and item parameters were estimated. Differential Item Functioning (DIF) was studied for age and gender, and DIF for language was studied as a measure of cross-cultural validity. Hypotheses concerning construct validity were tested by correlating item bank-scores with scores on several legacy instruments. The GRM showed suboptimal evidence of unidimensionality in confirmatory factor analysis (CFI: 0.903, TLI: 0.897, RSMEA: 0.144), and 99 item pairs with local dependence. A bi-factor model showed good fit (CFI: 0.964, TLI: 0.961, RSMEA: 0.089), with a high Omega-H (0.97), a high Explained Common Variance (ECV: 0.81), and no local dependence. Sufficient monotonicity was shown for all items (Mokken $H_{(i)}$: 0.367-0.686). The unidimensional IRT model showed good

fit (Only two items with S-X^2 < 0.001), with slope parameters ranging from 1.00 to 4.27, and threshold parameters ranging from -1.77 to 3.66. None of the items showed DIF for age or gender. One Item showed DIF for language. Correlations with legacy instruments were high (Pearson's R: 0.53-0.75), supporting construct validity. The item bank showed good item fit, good coverage of the pain interference trait, and good construct validity. It was concluded that the Dutch-Flemish v1.1 PROMIS Pain Interference item bank showed good IRT item fit, good coverage of the pain interference trait, and good construct validity. CFA and analyses of local independence showed evidence of multidimensionality, but omega-H and ECV were high, indicating a low risk of biased parameters when assuming unidimensionality. It was concluded that these results supported the validity of the DF-PROMIS-Pain Interference item bank, and that the item bank can be used as a basis for short forms and computer adaptive testing in clinical research and in clinical practice.

Chapter 8

In chapter 8 of the thesis the findings of the characteristics of MSK physicians and of their patient population are discussed with special attention to the possible role of MSK physicians in Dutch health care. The health care landscape is rapidly changing, and health care for patients with musculoskeletal complaints could benefit from a more patient-centred approach. The acceptance of MSK as an additional competence for medical specialists by the UEMS, together with recent changes in the educational program for MSK physicians, including a new description of their competences, could make MSK physicians suitable for a more central role in musculoskeletal health. Because of their user friendliness the PROMIS questionnaires will be perfectly suited for routine outcome measurement. A web-based measurement system using PROMIS questionnaires is currently under construction and will be used for further study.

Samenvatting

Dit proefschrift is een product van een groot onderzoeksproject dat uitgevoerd kon worden dankzij de steun van de Nederlandse Vereniging voor Musculoskeletale Geneeskunde (NVAMG). De wens om onderzoek te doen kwam voort uit het feit dat er geen eerder onderzoek was waaruit gegevens bekend waren over MSK artsen en hun patiëntenpopulatie. Op het onderzoek naar de psychometrische eigenschappen van de *Neck Disability Index* (NDI) na is voor alle studies gebruik gemaakt van een computerprogramma waarmee geautomatiseerd gegevens konden worden verzameld met web-based vragenlijsten. Deze vragenlijsten zijn meetinstrumenten waarmee veranderingen op het gebied van pijn en functioneren gemeten kunnen worden. Het computerprogramma werd speciaal voor dit onderzoek ontwikkeld, en maakte gebruik van email adressen om patiënten uit te nodigen om vragenlijsten in te vullen. De email bevatte daarvoor een link naar de web-based vragenlijsten. Deelnemende artsen werd gevraagd om bij een eerste consult een aantal algemene gegevens over de patiënt in te voeren in een web-based register, zoals de hoofdklacht, nevenklachten, de duur van de klachten en leeftijd. Na de laatste behandeling konden gegevens ingevoerd worden over de toegepaste behandeling en het aantal behandelsessies. Aan patiënten die daarvoor toestemming hadden gegeven werd gevraagd om vragenlijsten in te vullen aan het begin van de behandelingen, en op verschillende momenten na de behandelingen. Het programma is gebruikt in vier fases, waarbij in iedere fase andere vragenlijsten werden gebruikt en ook verschillende meetmomenten. Op deze manier konden grote aantallen gegevens worden verzameld waarmee een aantal onderzoeksvragen beantwoord konden worden. In de eerste hoofdstukken wordt de achtergrond van MSK artsen en de kenmerken van hun patiëntenpopulatie beschreven, wordt het verloop van pijnklachten bij patiënten met lage rugklachten na behandeling door een MSK arts geëvalueerd, en worden bijwerkingen beschreven zoals die gerapporteerd zijn door patiënten na MSK behandeling. In de andere hoofdstukken wordt de validiteit geëvalueerd van twee recent ontwikkelde meetinstrumenten die gebruikt kunnen worden om onderzoek te doen bij patiënten met klachten van het bewegingsapparaat.

Hoofdstuk 2

In hoofdstuk 2 worden de resultaten beschreven van een studie waarin de achtergrond van MSK artsen en de kenmerken van hun patiëntenpopulatie werden geëvalueerd. Gegevens over MSK artsen werden verzameld met behulp van een enquête en telefonische interviews. De enquête werd verstuurd naar 138 MSK artsen, waarvan 90 de vragenlijst beantwoordden (65%). De meeste artsen specialiseerden in MSK geneeskunde na een carrière in

andere medische specialismen. MSK artsen maakten gebruik van diverse diagnostische mogelijkheden en behandeltechnieken, waarvan een gedeelte voorbehouden is aan artsen, zoals injecties of voorgeschreven medicatie. Veelal maakte mobilisatie of manipulatie deel uit van de behandeling. Patiëntengegevens werden verkregen uit het web-based onderzoek. Functionele beperkingen werden gemeten met specifieke vragenlijsten voor lage rug en nekklachten (*Patient Reported Outcome Measures*, PROMs). De meerderheid van de patiënten had lang bestaande (62% > 1 jaar) lage rug (48%) en nekklachten (17%), met gemiddelde scores van respectievelijk 6,0 en 6,2 op een numerieke schaal voor pijn intensiteit, en met matige functionele beperkingen. De meeste patiënten rapporteerden eerdere behandelingen door fysiotherapeuten (68%), manueel therapeuten (38%) of chiropractors (17%). De conclusie van dit onderzoek met betrekking tot de MSK artsen was dat deze gebruik maken van verschillende manipulatieve technieken. Deze technieken zijn een onderdeel van een bredere aanpak, waarbij gebruik wordt gemaakt van verschillende diagnostische mogelijkheden en behandeltechnieken, gedeeltelijk voorbehouden aan artsen.

Hoofdstuk 3

In hoofdstuk 3 worden de resultaten gepresenteerd van een onderzoek waarin het verloop van de pijnklachten werd geëvalueerd bij patiënten die behandeld waren in verband met lage rugklachten. Gegevens werden verzameld met behulp van het web-based register. MSK artsen registreerden patiëntengegevens en enkele details over de toegepaste behandelingen. Patiënten kregen vragenlijsten waarin onder meer geïnformeerd werd naar eerdere medische consumptie, samen met PROMs waarmee het pijnniveau en functionele beperkingen werden gemeten. Dit gebeurde aan het begin van de behandelingen, en iedere 6 weken voor een periode van 6 maanden. Met behulp van latente klasse analyse (*Latent Class Growth Anlayses*, LCGA) werden patiënten in verschillende groepen ingedeeld op grond van het verloop van de pijnscore. Verschillende anamnestische gegevens werden geëvalueerd als mogelijke predictoren van een gunstig verloop. In twee jaar werden 1377 patiënten gerekruteerd, waarvan 1117 patiënten minstens één vervolgvragenlijst beantwoorden (81%). LCGA identificeerde drie groepen met een verschillend verloop. Een eerste groep met een hoog initieel pijnniveau toonde geen verbetering (N=226), een tweede groep met een hoog initieel pijnniveau toonde een sterke verbetering (N=578), en een derde groep met een lager initieel pijnniveau toonde een matige, maar wel klinisch relevante verbetering (N=313). De twee groepen met een hoog initieel pijnniveau werden met elkaar vergeleken, en een predictiemodel werd gemaakt voor een gunstig verloop van de pijn. Manlijk geslacht, eerder specialist bezoek, eerdere behandeling bij een pijnkliniek, werkend, een kortere duur van de huidige pijnepisode, en een lagere tijd sinds de klachten voor het eerst begonnen waren predictoren voor een gunstig verloop. Het predictiemodel had een lage *'Area Under the*

Curve (AUC)' (0,68), en een lage verklaarde variantie (0,09). De conclusie van dit onderzoek was dat een grote groep patiënten verbeterde in het eerste half jaar na de aanvang van de MSK behandeling, maar dat het niet goed mogelijk was om te voorspellen welke patiënten een gunstige verloop zouden hebben.

Hoofdstuk 4

In hoofdstuk 4 worden de resultaten gepresenteerd van een studie waarin bijwerkingen werden geëvalueerd na MSK behandeling bij patiënten met lage rug en nekklachten. MSK artsen registreerden patiëntengegevens en enkele details over de toegepaste behandeling. Patiënten werd gevraagd om vragenlijsten in te vullen aan het begin van de behandeling waarmee het pijnniveau en functionele beperkingen werden gemeten. Drie maanden na het begin van de behandelingen werd aan patiënten gevraagd om een vragenlijst in te vullen waarin gevraagd werd naar het soort bijwerkingen die ze hadden ervaren na MSK behandeling. Daarbij werd voor iedere bijwerking apart gevraagd naar de ernst en de duur van deze bijwerking. Een totaal van 1391 lage rugpijn patiënten en 549 nekpijn werden gerekruteerd, waarvan 823 lage rugpijn patiënten (59%) en 315 nekpijn patiënten (57%) de bijwerkingenvragenlijst beantwoorden. Van deze patiënten rapporteerden 362 patiënten (31,8%) een totaal van 683 bijwerkingen. Op vijf na werden alle patiënten behandeld met een manipulatieve of mobiliserende techniek, in gemiddeld 3-6 behandelsessies (bereik 1-12). De meeste patiënten rapporteerden een enkele bijwerking, en de meest voorkomende bijwerking was vermoeidheid. Patiënten met nekpijn rapporteerden vaker bijwerkingen (38,4%) dan patiënten met lage rugklachten (29,3%). De meeste bijwerkingen waren niet ernstig en verbeterden binnen een week, maar sommige patiënten rapporteerden ernstigere bijwerkingen (6,9%) die langer duurden (7,1%). De conclusie van dit onderzoek was dat bijwerkingen na behandeling door MSK artsen vaak voorkwamen, maar meestal niet ernstig waren en van korte duur. Nekpijn patiënten rapporteerden andere bijwerkingen dan patiënten met lage rugklachten. Met name vermoeidheid werd vaker gerapporteerd door nekpijn patiënten. Patiënten die behandeld waren met een mobiliserende techniek rapporteerden iets meer bijwerkingen dan patiënten die behandeld waren met een manipulatieve techniek, wat voor een groot gedeelte verklaard werd door het frequenter voorkomen van vermoeidheid. Er was geen significante relatie tussen het voorkomen van bijwerkingen en het ervaren effect van de behandeling na drie maanden.

Hoofdstuk 5

In hoofdstuk 5 worden de resultaten gepresenteerd van een studie waarin verschillende methoden om veranderscores te evalueren voor een vragenlijst over beperkingen vanwege

nekpijn (de *Neck Disability Index*, NDI). Veranderscores voor de NDI werden geëvalueerd d.m.v. de *Smallest Detectable Change* (SDC) en de *Minimal Important Change* (MIC). In een cohortstudie werden 101 patiënten gerekruteerd die de NDI beantwoorden aan het begin van de behandeling en opnieuw zes maanden na het begin van de behandeling. De SDC en de MIC werden berekend met twee verschillende vergelijkingswaarden (*Anchors*). Daarnaast werd met deze vergelijkingswaarden de populatie op twee verschillende manieren verdeelt in groep patiënte die zichzelf verbeterd achtten en een groep patiënten die vonden dat de klachten onveranderd waren. Verder werden de SDC en de MIC apart berekend voor klinische relevante subgroepen: patiënten met of zonder uitstralende pijn, patiënten met of zonder bijkomende hoofdpijn, en patiënten met hoge en lage beginscores. De resultaten lieten zien dat de SDC en de MIC vrijwel niet werden beïnvloed door verschillende definities van verbetering te hanteren. De waarden waren vergelijkbaar voor patiënten met of zonder uitstralende pijn, maar verschilden sterk voor patiënten met of zonder bijkomende hoofdpijn, en voor patiënten met hoge en lage beginscores. De conclusie van dit onderzoek was dat de SDC en de MIC geen onveranderbare eigenschappen van de NDI zijn, maar sterk beïnvloed kunnen worden door patiënten kenmerken.

Hoofdstuk 6

In hoofdstuk 6 worden de resultaten gepresenteerd van een studie waarbij de validiteit werd onderzocht van de Nederlands-Vlaamse versie van de v1.1. PROMIS item bank 'Pijngedrag' in een populatie van 1602 patiënten met klachten van het bewegingsapparaat. Aannames die ten grondslag liggen aan het gebruikte *Item Response Theory* (IRT) model werden geëvalueerd in een *Graded Response Model* (GRM): unidimensionaliteit en lokale afhankelijkheid door middel van confirmatieve factoranalyse, en monotoniciteit door middel van schaalbaarheidsparameters. De fit van het IRT model werd geëvalueerd voor alle items, en de item parameters werden geschat. Meetinvariantie, ofwel *Differential Item Functioning* (DIF) werd geëvalueerd voor leeftijd en geslacht, en DIF voor taal werd geëvalueerd als maat voor cross-culturele validiteit. Het GRM toonde een suboptimale fit van het unidimensionele model (CFI: 0,816; TLI: 0,806; RSMEA: 0,093), en vijftien item-paren toonden lokale afhankelijkheid. Vijf items toonden slechte schaalbaarheid (Mokken $H_{(i)}$: 0,14-0,41). Een bi-factor model toonde een laag risico voor vertekening als een unidimensioneel model werd aangenomen (Omega-H 0,92, *Explained Common Variance* (ECV) 0,70), maar de fit van dit unidimensionele model was nog steeds suboptimaal (CFI: 0,922; TLI: 0,915; RSMEA: 0,062), waarbij drie item paren nog steeds lokale afhankelijkheid toonden (0.4%). Alle items hadden een voldoende fit in het IRT model; slope parameters varieerden van 0,60 tot 2,00, en threshold parameters van -2,05 tot 6,80. Eén item toonde DIF voor geslacht, en vijf items toonden DIF voor taal, waarbij de invloed van DIF op de totale score laag bleek. De

conclusie van dit onderzoek was dat de PROMIS Pijngedrag item bank kan worden gebruikt in de klinische praktijk en in onderzoek, maar dat de problemen met de dimensionaliteit verder onderzocht moeten worden in andere populaties.

Hoofdstuk 7

In hoofdstuk 7 worden de resultaten gepresenteerd van een studie waarbij de validiteit werd onderzocht van de Nederlands-Vlaamse versie van de v1.1. PROMIS item bank 'Belemmeringen door pijn' in een populatie van 1677 patiënten met klachten van het bewegingsapparaat. Aannames die ten gronde liggen aan het gebruikte *Item Response Theory* (IRT) model werden geëvalueerd in een *Graded Response Model* (GRM): unidimensionaliteit en lokale afhankelijkheid door middel van confirmatieve factoranalyse, en monotoniciteit door middel van schaalbaarheidsparameters. De fit van het IRT model werd geëvalueerd voor alle items, en de item parameters werden geschat. '*Differential Item Functioning* (DIF)' werd geëvalueerd voor leeftijd en geslacht, en DIF voor taal werd geëvalueerd als maat voor cross-culturele validiteit. Hypothesen werden geformuleerd waarmee de construct validiteit geëvalueerd kon worden door de scores te vergelijken met de scores op andere, vergelijkbare vragenlijsten. Het GRM toonde een suboptimale fit van het unidimensionele model (CFI: 0,903; TLI: 0,897; RSMEA: 0,144), en 99 item paren toonden lokale afhankelijkheid. Een bi-factor model toonde een goede fit(CFI: 0,964; TLI: 0,961; RSMEA: 0,089), met een Omega-H (0,97), een hoge *Explained Common Variance* (ECV: 0,81), en geen lokale afhankelijkheid. Alle items waren voldoende schaalbaar (Mokken $H_{(i)}$: 0,367-0,686). Het unidimensionele IRT model toonde geode fit (Slechts twee item $S\text{-}X^2 < 0{,}001$), met slope parameters van 1,00 tot 4,27, en threshold parameters van -1,77 tot 3,66. Geen van de items toonden DIF voor leeftijd of geslacht, en een item toonde DIF voor taal. De vragenlijst correleerde goed met vergelijkbare vragenlijsten (Pearson's R: 0,53-0,75), wat de construct validiteit ondersteunde. De item bank toonde een goede IRT item fit, een goede dekking van belemmeringen door pijn, en een goede construct validiteit. Confirmatieve factoranalyses gaven aanwijzingen voor multidimensionaliteit, maar de hoge Omega-H en ECV gaven aan dat het risico op vertekende scores door een unidimensioneel model aan te houden beperkt is. De resultaten ondersteunen de validiteit van de PROMIS item bank 'Belemmeringen door pijn', en de conclusie van dit onderzoek was dat deze item bank gebruikt kan worden al basis voor korte versies en computer adaptief testen in de klinische praktijk en in onderzoek.

Hoofdstuk 8

Hoofdstuk 8 betreft een discussie van de resultaten, met speciale aandacht voor de mogelijke rol van MSK artsen in de Nederlandse gezondheidszorg. Het zorglandschap

is momenteel erg aan het veranderen, en de zorg voor patiënten met klachten van het bewegingsapparaat kan mogelijk verbeterd worden door een meer patiënt gerichte aanpak. Gesteund door de erkenning van MSK als een toegevoegde competentie voor medisch specialisten door de UEMS, samen met recente veranderingen in de opleiding tot MSK arts en het recent aangepaste competentieprofiel kan de MSK arts mogelijk een meer centrale rol vervullen in de zorg voor patiënten met klachten van het bewegingsapparaat. Door hun gebruiksvriendelijkheid zijn meetinstrumenten gebaseerd op de PROMIS item banken zeer geschikt om routinematig uitkomsten te meten. Een web-based systeem waarbij gebruik gemaakt wordt van PROMIS item banken wordt momenteel gebouwd en zal gebruikt worden voor verder onderzoek.

Dankwoord

Deze dissertatie was niet mogelijk geweest zonder de inzet van velen. Als eerste wil ik de NVAMG, en met name **Anja**, **Bram** en **Henk** bedanken voor het mogelijk maken van dit onderzoeksproject. Het gehele bestuur heeft het onderzoek altijd enthousiast gesteund, en ook de ledenvergadering was altijd blij met het onderzoeksproject. Daarnaast heeft een aantal collega's jarenlang meegedaan met het onderzoek. Ze woonden voorlichtingsbijeenkomsten bij en overlegden met de onderzoeksassistente. Sommige collega's moesten hun praktijkprocessen aanpassen om patiënten voor het onderzoek te rekruteren. Ondanks de extra belasting heb ik nooit een onvertogen woord gehoord. Ik kan hier niet alle collega's persoonlijk noemen, maar jullie vormden de basis waarop ik het hele project heb gebouwd. Mijn directe collega's nemen een speciale plaats in, en dan met name **Marianne**. Als jij niet het geregel van heel veel praktijkzaken op zich had genomen, dan was het mij nooit gelukt om zo'n groot onderzoeksproject te volbrengen.

Ik ben heel erg verwend geweest met het onderzoeksteam aan de VU. Allereerst had ik het geluk dat **Daphne** zich aandiende als onderzoeksassistent. Jij vulde mij aan op gebieden waar ik zelf onzeker over was, zoals het werken met grote databestanden, waarmee jij al ervaring had opgedaan bij TNO. Samen hebben wij een enorm cohortonderzoek opgezet waar meestal een heel team aan zou werken - heel bijzonder eigenlijk. De begeleiding door **Riekie** en **Raymond** was buitengewoon goed. Jullie hadden er soms wel wat werk aan om mij gefocust te houden en deden dit steeds met veel interesse en enthousiasme. Het EMGO heb ik altijd als een van de vooraanstaande onderzoeksinstituten gezien, niet alleen in Nederland, maar in de hele onderzoekswereld. Op internationale congressen bleek steeds weer dat heel veel bepalende publicaties uit onze koker kwamen. Ik ben er dan ook trots op dat ik hier het woord 'onze' kan gebruiken. Binnen het EMGO heb ik het bovendien getroffen met mijn begeleidingsteam, **Riekie**, **Raymond**, **Caroline** en **Martijn**. Anderen waren wel eens jaloers op de begeleiding die ik kreeg.

In het begin voelde ik mij wat verlaten, achter een computer zittend op de VU, maar in de loop van de jaren heb ik er toch goede vrienden gemaakt. Vooral met **Tsjitske** en **Steve** heb ik lange discussies gevoerd over onderzoeksmethodes, maar we spraken ook af voor gezellige etentjes en gingen samen naar het Low Back Pain congres in Odense. Met **Adri** heb ik veel gesproken over het nut van orthomanuele geneeskunde en fysiotherapie bij de behandeling van rugklachten. We zitten allebei misschien wel in hetzelfde schuitje, doordat we aan de ene kant in de praktijk werken waar we dingen doen die heel nuttig lijken, maar die matig ondersteund worden door wetenschappelijk onderzoek, en aan de

andere kant actief zijn in het wetenschappelijk onderzoek waarmee we zoeken naar een betere onderbouwing. Dit plaatst ons in een moeilijke spagaat, waarbij je aan collega's in het onderzoek moet uitleggen waarom je patiënten zo behandelt als je doet, en waarbij je aan collega behandelaars moet uitleggen wat het nut is van wetenschappelijk onderzoek. Dit was ook vaak het onderwerp van gesprek bij onze etentjes bij Nam Kee, met **Maurits**, **Raymond** en **Maurice**. Ik heb mij verder ook erg gesteund, en zelfs gewaardeerd gevoeld door de andere collega's van de afdeling klinimetrie, **Wieneke** en **Sanna**. En **Luc**, ik vond het fantastisch om paranimf te zijn bij jouw promotie.

Verder moet ik de collega's bedanken die mij hebben geholpen, vooral bij de moeilijke analyses die nodig waren voor de PROMIS artikelen. Samen met **Martine** had ik het nadeel dat er geen statisticus beschikbaar was die deze analyses in zijn geheel kon uitvoeren, iets wat een aantal jaren voor enige stress heeft gezorgd, zacht gezegd. Omdat intussen de manier waarop er naar deze analyses werd gekeken zich verder ontwikkelde nam de complexiteit alleen maar toe, terwijl wij niet vooruit konden. Uiteindelijk is dit opgelost door in een klasje de analyses zelf te leren uitvoeren met behulp van en R-scripts die onder begeleiding van **Thomas** werden gemaakt. Een regenachtige herfstvakantie waarin ik dagelijks tot diep in de avond in een appartementje in Egmond aan Zee bezig was om met behulp van dit script ook daadwerkelijk analyses uit mijn computer te krijgen was wel zo'n beetje het dieptepunt van mijn hele promotietraject. Uiteindelijk heeft dat wel geleid tot een paar mooie artikelen, met de verdere hulp van **Caroline**, **Leo** en **Berend**.

Kimi, jou wil ik apart bedanken. Toen jij je meldde in kamer A505 hadden wij op een of andere manier gelijk een klik die verder ging dan alleen maar het contact tussen collega's. Ik moest erg lachen om het prachtige Amsterdams waarmee jij, na een bezoek aan het Baantjer museum, 'ja shaker' kon zeggen en Sander daarmee volledig op de kast kreeg. Bij een aantal artikelen heb je mij geholpen om er een duidelijk coherent verhaal van te maken, en het Engels mooier te maken. Deze taak is voor latere artikelen een beetje overgenomen door **Sophie**, die nu eenmaal dichterbij was, en als editor ook veel ervaring met dit werk had. Mijn lieve **Sophie**, jij hebt ook eraan bijgedragen dat het proefschrift uiteindelijk toch ook af gemaakt werd, onder meer bij het inhoud geven aan de introductie en de discussie, maar ook door mij te stimuleren praktische keuzes te maken.

Ellen, **Sanne** en **Nils**, zonder jullie was dit onderzoek nooit gelukt. Vanaf 2008 ben ik bezig geweest met dit onderzoek, en ik zat tot vervelens toe daaraan te werken achter de laptop aan de keukentafel. Waar andere gezinnen misschien wel in de weekenden gaan wandelen in het bos, besteedde ik veel weekenden aan het schrijven van dit proefschrift. En dikwijls ging de computer ook mee met vakantie, omdat ik dan de tijd had om er een paar dagen

achtereen aan door te werken. Ik hoop niet dat ik jullie daarmee tekort heb gedaan. Er werd weleens wat gemopperd op mijn 'projectjes', maar ik mocht altijd daarvoor de tijd en ruimte nemen die ik daarvoor nodig achtte, al ging dat misschien ten kosten van jullie tijd en ruimte.

About the author

After training as cryptanalysist with the Royal Dutch Navy, Wouter Schuller studied medicine from 1982-1988 at the VU University. He worked for a while as SHO orthopedics in a third referral hospital in Bagdad and as Casualty officer in Greenwich District Hospital, London. Working with a specialised spinal surgeon sparked his keen interest in the spine. This interest eventually made him decide to specialise in orthomanual medicine, a type of spinal manipulative treatment that had recently been developed in The Netherlands. He subsequently started Spineclinic, a practice in Zaandam, which over the years developed into a center with 4 physicians and 5 physiotherapists. Despite his enthusiasm for his work, he considered the lack of scientific grounding of this profession, and the position within the Dutch Healthcare system a serious drawback. Scientific research would be necessary to improve this, and he followed the post-initial epidemiology master training at the EMGO institute of the VU Medical Center from 1997-2008. This training enabled him to set up the research project that eventually resulted in this dissertation.

Wouter was a board member of the Dutch Association for Musculoskeletal Medicine for 9 years, and as chairman of the scientific committee he was responsible for the scientific policy. Over the years he has sung in various baroque choirs, such as the choir of the Dutch Baroque Society, and he contributed to recordings under Gustav Leonhardt, Sigiswald Kuijken and Jos van Veldhoven. Currently, he sings with the collegium of the Cappella Nicolai. Wouter has been a practitioner of Wado-Ryu karate for 40 years.