EVALU
SHOU
PA

Diagnostics
and its interpretation

Marloes Thoomes-de Graaf

EVALUATING SHOULDER PAIN

Diagnostics
and its interpretation

Marloes Thoomes-de Graaf

Cover design: Erwin Timmerman, Optima Grafische Communicatie
Layout: Optima Grafische Communicatie, Rotterdam, The Netherlands
Printed by: Optima Grafische Communicatie, Rotterdam, The Netherlands
ISBN/EAN: 978-94-6361-101-5

The Medical Ethics Committee of the Erasmus Medical Center in Rotterdam approved the study (MEC-2011-414).
This study was financed by the SIA-RAAK grant serving exclusively for lectureships and knowledge networks at Universities of Applied Sciences. This study is partly funded by a program grant of the Dutch Arthritis Foundation.

The printing of this thesis was financially supported by the Erasmus Medical Center, the Scientific College Physical Therapy (WCF) of the Royal Dutch Society for Physical Therapy (KNGF) and the department of General Practice of the Erasmus Medical Centre, Rotterdam.

Copyright © 2018 Marloes Thoomes-de Graaf, the Netherlands. No part of this thesis may be reproduced or transmitted in any form or by any other means, electronical or mechanical, included photocopy, recording or any information storage or retrieval system, without permission of the copyright holder.

EVALUATING SHOULDER PAIN; DIAGNOSTICS AND ITS INTERPRETATION

De evaluatie van schouderpijn; diagnostiek en de interpretatie

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Erasmus Universiteit Rotterdam

op gezag van de rector magnificus Prof. dr. R.C.M.E. Engels
en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op donderdag 21 juni 2018 om 15.30 uur
door

Marloes Thoomes-de Graaf
geboren te Alphen aan den Rijn

Erasmus University Rotterdam

PROMOTIECOMMISSIE

Promotor:	Prof. dr. B.W. Koes
Overige leden:	Prof. dr. H.J. Stam Prof. dr. J.J. van Busschbach Prof. dr. ir. H.C.W. de Vet
Copromotoren:	Prof. dr. A.P. Verhagen Dr. G.G.M. Scholten-Peeters

TABLE OF CONTENTS

GENERAL INTRODUCTION

CHAPTER 1

Epidemiology

Shoulder pain is a common disorder in western society. In the Netherlands, it is the second most reported musculoskeletal complaint in the general population, with a point prevalence of 20.9% [1]. Only the prevalence of low back pain exceeds shoulder pain [1]. International data reports a point prevalence ranging from 7 to 26%, a 12-month prevalence ranging from 5 to 47% and a lifetime prevalence ranging from 7 to 67%, depending on case definitions and age [2].

The point prevalence in the Netherlands is highest in the age group of 45-64 for both men and women. In all age categories, the point prevalence is higher for women than for men [1]. The point prevalence of chronic pain in the shoulder region in the general population has been estimated at 15.1% [1]. Chronic was defined here as current pain lasting for more than 3 months [1].

Consequences of shoulder pain

There is a general lack of knowledge regarding pathophysiology and etiology of shoulder pain. However, the onset of shoulder pain is assumed to be related to a multiple set of combined factors, including individual factors (e.g. age, gender, BMI), physical work load factors and psychosocial work environment factors (e.g. stress, work organization) [3-5]. In general, patients with shoulder problems, apart from pain, report having functional disabilities [6, 7], especially when the dominant shoulder is affected [8]. The reported functional disabilities in patients with shoulder pain range from difficulties with moving their arm/hand, self-care to impeding sleep [7, 9] and can reach a level of severity where they preclude work-related tasks which can result in sick leave and indirect costs [1, 7, 10]. There are five main functional limitations that are mentioned by the majority of patients (in a sports medicine orthopedic surgeon setting), namely; hand and arm use (performing coordinated actions required to move objects or to manipulate them by using hands and arms, such as when turning door handles), lifting and carrying objects (e.g. lifting a cup), exercise tolerance functions (related to respiratory and cardiovascular capacity as required for enduring physical exertion), recreation and leisure activities and sleep function [6].

Prognosis

The prognosis of patients with a new episode of shoulder pain is not always favorable, as reported recovery rates at 6 months after initial consultation to a general practitioner (GP) vary from 21% [9] to 54% [11] and increase up to 49% [9] to 59% [12] after 12 months. Prognostic factors are widely used and have been the subject of research for a considerable period of time now [13-16]. There is moderate to strong evidence from three systematic reviews, with slightly different results, that a longer duration of complaints [17-19] a high level of disability (SPADI) at baseline, a high level of pain intensity

at baseline and increasing age predict a poorer outcome in patients with shoulder pain [17-19].

Research in other areas such as psychotherapy and psychology has shown that the working/therapeutic alliance between a therapist and a patient could be a predictor for improvement [20-23]. Moreover, physiotherapy students indicate that the therapeutic alliance has become very important within their future profession [24].

Health care consumption

The most recently published annual average incidence in the Netherlands of people consulting their GP with shoulder pain was 29.3 (95% confidence interval (CI): 28.5-30.0) per 1000 person-years, calculated for the period 1998 to 2007 [25]. In concordance with the National Guideline for GPs, a Dutch retrospective cohort study indicated that usual care after a first GP consultation, consists of a prescription for oral NSAIDs (50%), wait-and- see policy (32%), a referral for physiotherapy (15%), a cortisone injection (3%), or a combination of physiotherapy and medication [26, 27]. The majority of patients that have been referred by their GP during their episode of shoulder pain, have been referred to physiotherapy (84%) followed by rehabilitation medicine (6%) and orthopedic surgery (6%) [26]. These numbers are to a large extend comparable with international data [28].

Since 2006, patients in the Netherlands have the possibility to consult a **physiotherapist (PT)** without contacting their GP (direct access). In the year of its introduction about 27% of patients contacted a PT through direct access, versus 73% after referral [29]. This number has increased since then, from 35.4% in 2011 to 53.5% in 2016 [30]. This Dutch report also indicates that shoulder pain is a frequently occurring health care problem within a population consulting physiotherapy [30].

There is no information however, with regards to the total number of patients with shoulder pain consulting a PT. Only one study, using data of a registration network of physiotherapy practices in the Netherlands (gathered between 2006 and 2010), provides an indication of the number of patients visiting a PT due to shoulder pain [31]. Originally this study focused on shoulder syndromes, which according to this study were responsible for 2.6% (1182) of the total number of patients visiting a PT. They stated these 1182 patients were a proportion of 27% of all patients visiting a PT due to shoulder pain [31].

Economic consequences

The economic consequences of shoulder pain in a primary care population are moderate [32]. On average the costs per patients during the 6 months after first consultation, were €689 (standard deviation (SD) ± €1965), a large proportion of which was due to indirect costs of productivity losses [32]. The direct costs consisted mainly (37%) of charges for treatment sessions by a therapist (mostly for physiotherapy) [32]. Remarkably, a small percentage of patients (12%) was responsible for the majority of costs (74%). These pa-

tients reported more sick leave, a higher pain severity score and more shoulder disability at baseline [32].

Physiotherapy assessment

Ultimately, a proportional part of patients visits a PT, either via direct access or via their GP. Patients consulting a PT expect information regarding their condition, advice and explanation about self- management [33, 34]. Physiotherapy assessment usually starts with history taking and in addition might include the use of relevant health related patient reported outcome measures **(PROMs).** As functional disability is one of the main complaints for patients with shoulder pain, outcome measurements should include an instrument to objectify functional disabilities/ perceived "activity limitations" in terms of assessing the physical impairment in patients with shoulder pain [6, 7, 35-37]. Several of these PROMs have been developed and a number of reviews have been performed to assess their psychometric properties and some assessed the quality of the individual studies (using self- constructed checklists). The COSMIN checklist, has been developed to evaluate the methodological quality of studies investigating the measurement properties of PROMs [38].

These reviews all have included studies with mixed populations, such as upper extremity disorders, which impacts their recommendations. Furthermore, these reviews presented their results per PROM not taking into account language variations of the PROM at issue. Due to differences in cultural context however, a translation of the original version does not guarantee similar psychometric properties [39, 40]. Therefore, the psychometric qualities of (translated) PROMs should be evaluated, for patients with shoulder pain, before they can be used in daily practice or research.

A number of reviews have encouraged the use of the Shoulder Pain and Disability Index (SPADI) in clinical and research settings [41-43]. Moreover, the Royal Dutch Society for Physical Therapy (KNGF) has recommended implementation of the Dutch SPADI (SPADI-D) in the evidence statement [44]. Despite its frequent use internationally, the SPADI-D has not yet been validated and tested for reliability in a Dutch setting.

Findings during history taking combined with those from physical examination and possibly the use of PROMs, leads to a **(physiotherapy) diagnosis** [44]. Physical examination, including specific tests, alone is not valid to differentiate between various disorders, because of low sensitivity, specificity and reproducibility [45-47]. Nowadays, diagnostic musculoskeletal ultrasonography (DMUS) is increasingly used by PTs to overcome this problem [48]. Medical specialists (most often radiologists) are able to accurately diagnose several shoulder disorders (full thickness tear, partial thickness tear, subacromial bursitis and calcifying tendonitis) using DMUS [49-51]. Only a small number of studies

evaluated subacromial bursitis and calcifying tendonitis and although promising, the results should be interpreted with caution [50].

However, research regarding the diagnostic accuracy of DMUS for full thickness rotator cuff tears, showed a pooled sensitivity of at least 0.92 and specificity higher than 0.94 for medical specialists [49-51]. Besides, the reliability between radiologists for full thickness tears is good (κ = 0.90-0.95) [52, 53]. The learning curve for a non-musculoskeletal radiologist appears to be relatively short, as the agreement between an experienced musculoskeletal radiologist and a less experienced (half year) radiologist in DMUS is good (κ = 0.90) [53]. Moreover, the agreement between a general radiologist and an experienced musculoskeletal radiologist increased from good (κ = 0.81) during the first 50 consultations to excellent (κ = 0.96-1.00) thereafter [52]. The recommended operator experience for surgeons, based upon the increase rate of sensitivity and specificity of the DMUS compared to Magnetic Resonance Imaging (MRI) or arthroscopy, is 100 diagnostic ultrasounds of the shoulder [54, 55].

With regards to assessing partial thickness rotator cuff tears, specificity remained high, but sensitivity decreased (ranging from 0.67 to 0.84) for medical specialists [49-51]. Also, the reliability between radiologists decreased as the overall kappa ranged between 0.63 and 0.79 [52, 53].

However, little is known about the reliability and validity (and the influence of experience) of DMUS in primary (physiotherapy) care settings. Interestingly, only a small percentage (13.3%) of orthopedic surgeons and radiologist trusts the results of a PT when using DMUS. Therefore, in the majority of patients the DMUS is repeated in secondary care [56]. In case a DMUS is not valid and reliable, it is not in the best interest of the patient, as well as the therapist, to use DMUS for defining diagnostic labels for their symptoms.

Patient satisfaction largely depends on the communication skills of the PT, such as "explaining" and "teaching" abilities [33, 57]. As the prognosis of patients with shoulder pain is not particularly favorable, it is likely the patient will see more than one health care professional. It can be frustrating and confusing if a patient receives different diagnostic labels from different health care professionals (e.g. GP, PTs, etc.) such as 'tendinitis' or 'impingement'. Moreover, diagnostic labels have implications on the perceptions of patients and this should be taken into account when using them [58].

Ideally, a diagnostic tool should assist in differentiating between **clinically important subgroups**, as it immediately impacts the therapeutic process. E.g. diagnostic ultrasound could hypothetically be used to distinguish between patients that need referral to secondary care (potentially specific or serious pathology), the ones that could benefit from physiotherapy management and those that should just be monitored and receive a wait-and-see approach.

In order to complete the diagnostic process, the PT has to make an estimate with regards to the clinical course and the prognosis in order to inform the patient. PROMs can be of help, such as the SPADI [17-19]. The Working Alliance Inventory (WAV-12) is one of the most commonly used and validated questionnaires to measure working alliance [59], although it has not yet been validated in Dutch.

Evaluating treatment effect

Physiotherapy usually consist of 'information and advice', exercise therapy and mobilization and is effective for a number of shoulder conditions [60-68]. As treatment of shoulder pain is usually aimed at pain reduction and improvement of functional disabilities, it is important to measure whether physiotherapy treatment is effective concerning these outcomes [35]. In order to do so, the responsiveness and interpretability of change scores of PROMs targeting limitations in activity should be assessed. The study population can have an impact on the responsiveness of PROMs, both in terms of generalizability and in affecting the results. It is therefore important to assess the responsiveness in a population that is reflective of daily practice; patients with non-specific shoulder pain in primary care with/without conservative treatment. The SPADI-D has not been assessed on responsiveness yet.

PROMs are not being implemented in clinical care yet

Although the use of PROMs has been highly recommended in guidelines, PROMs are not (fully) integrated into clinical practice. A survey in 2008 among nearly 500 American PTs concluded that only half of them regularly used a PROM during their work [69]. PTs indicate that the most common reasons for not using PROMs are that it is too time consuming for patients to complete (43%) and for clinicians to analyze, calculate, and score (30%) [69]. Similar findings were reported in a study assessing the 'barriers and facilitators' for the implementation of standardized measures in physiotherapy in the Netherlands. This study focused not only on the use of PROMs but the use of standardized measures in general. A total of 468 Dutch PTs participated, of which 394 worked in primary care. Even though the majority of PTs had a positive attitude towards the use of standardized outcome measures and was convinced of the advantages of the use of measurement instruments, it was hard to implement standardized measures into their daily clinical care. The main barriers mentioned were a lack of knowledge with regards to appropriate measures and a lack of time. PTs stated goniometry and pain assessment using the Visual Analogue Scale (VAS) were the most often used measures. However, the assessment of activity and participation clearly was not routinely used in primary physiotherapy care [70].

Therefore, it would be useful to create a less time-consuming PROM to measure limitations in activity and to assess its predictive value, as there is consistent evidence

that a high level of disability is one of the predictors of poor recovery for patients with shoulder pain [18].

AIMS

Based on the lack of knowledge and insight in the diagnosis and prognosis of patients with shoulder pain in physiotherapy care, the aims of this thesis are to:

- Critically appraise and compare the measurement properties of both the original versions as well as the translated versions of self-administered PROMs focusing on the shoulder assessing "activity limitations" for patients with nonspecific shoulder pain (**Chapter 2**).
- Evaluate the reliability and construct validity of the SPADI-D for patients with shoulder pain in primary care (**Chapter 3**).
- Assess the interrater-reliability of DMUS between physiotherapists and radiologists in patients with shoulder pain for full thickness tears, partial thickness tear, calcification and subacromial bursitis and to assess if experience or training of the physiotherapist influences the overall reliability (**Chapter 4**).
- Develop new labeling strategies based on the therapeutic consequences according to the literature; to explore a new clinical pathway and the inter-professional agreement of DMUS in patients with shoulder pain between physiotherapists and radiologists, using these new labeling strategies (**Chapter 5**).
- Assess whether the WAV-12 is a valid measurement instrument in terms of the construct and discriminative abilities for a population of patients with shoulder pain in physiotherapy care (**Chapter 6**).
- Evaluate the measurement error, interpretability and responsiveness of the SPADI-D on patients with shoulder pain seeking help by a physiotherapist in primary care setting (**Chapter 7**).
- Develop a single substitute question for the SPADI and evaluate its convergent/divergent validity, responsiveness and predictive power as this might be helpful to integrate a PROM into clinical practice (**Chapter 8**).

REFERENCES

1. Picavet, H.S. and J.S. Schouten, *Musculoskeletal pain in the Netherlands: prevalences, consequences and risk groups, the DMC(3)-study.* Pain, 2003. **102**(1-2): p. 167-78.
2. Luime, J.J., et al., *Prevalence and incidence of shoulder pain in the general population; a systematic review.* Scand J Rheumatol, 2004. **33**(2): p. 73-81.
3. van der Windt, D.A., et al., *Occupational risk factors for shoulder pain: a systematic review.* Occup Environ Med, 2000. **57**(7): p. 433-42.
4. Bodin, J., et al., *Risk factors for shoulder pain in a cohort of French workers: A Structural Equation Model.* Am J Epidemiol, 2017.
5. van Rijn, R.M., et al., *Associations between work-related factors and specific disorders of the shoulder--a systematic review of the literature.* Scand J Work Environ Health, 2010. **36**(3): p. 189-201.
6. Smith-Forbes, E.V., et al., *Descriptive analysis of common functional limitations identified by patients with shoulder pain.* J Sport Rehabil, 2015. **24**(2): p. 179-88.
7. Roe, Y., et al., *A systematic review of measures of shoulder pain and functioning using the International classification of functioning, disability and health (ICF).* BMC Musculoskelet Disord, 2013. **14**: p. 73.
8. Ozaras, N., et al., *Shoulder pain and functional consequences: does it differ when it is at dominant side or not?* J Back Musculoskelet Rehabil, 2009. **22**(4): p. 223-5.
9. Croft, P., D. Pope, and A. Silman, *The clinical course of shoulder pain: prospective cohort study in primary care. Primary Care Rheumatology Society Shoulder Study Group.* BMJ, 1996. **313**(7057): p. 601-2.
10. Feleus, A., et al., *Management in non-traumatic arm, neck and shoulder complaints: differences between diagnostic groups.* Eur Spine J, 2008. **17**(9): p. 1218-29.
11. Reilingh, M.L., et al., *Course and prognosis of shoulder symptoms in general practice.* Rheumatology (Oxford), 2008. **47**(5): p. 724-30.
12. van der Windt, D.A., et al., *Shoulder disorders in general practice: prognostic indicators of outcome.* Br J Gen Pract, 1996. **46**(410): p. 519-23.
13. Beneciuk, J.M., M.D. Bishop, and S.Z. George, *Clinical prediction rules for physical therapy interventions: a systematic review.* Phys Ther, 2009. **89**(2): p. 114-24.
14. Stanton, T.R., et al., *Critical appraisal of clinical prediction rules that aim to optimize treatment selection for musculoskeletal conditions.* Phys Ther, 2010. **90**(6): p. 843-54.
15. van Oort, L., et al., *Preliminary state of development of prediction models for primary care physical therapy: a systematic review.* J Clin Epidemiol, 2012. **65**(12): p. 1257-66.
16. Artus, M., et al., *Generic prognostic factors for musculoskeletal pain in primary care: a systematic review.* BMJ Open, 2017. **7**(1): p. e012901.
17. Struyf, F., et al., *A Multivariable Prediction Model for the Chronification of Non-traumatic Shoulder Pain: A Systematic Review.* Pain Physician, 2016. **19**(2): p. 1-10.
18. Kuijpers, T., et al., *Systematic review of prognostic cohort studies on shoulder disorders.* Pain, 2004. **109**(3): p. 420-31.
19. Chester, R., et al., *Predicting response to physiotherapy treatment for musculoskeletal shoulder pain: a systematic review.* BMC Musculoskelet Disord, 2013. **14**: p. 203.
20. Martin, D.J., J.P. Garske, and M.K. Davis, *Relation of the therapeutic alliance with outcome and other variables: a meta-analytic review.* J Consult Clin Psychol, 2000. **68**(3): p. 438-50.
21. Welmers-van de Poll, M.J., et al., *Alliance and Treatment Outcome in Family-Involved Treatment for Youth Problems: A Three-Level Meta-analysis.* Clin Child Fam Psychol Rev, 2017.

22. Ferreira, P.H., et al., *The therapeutic alliance between clinicians and patients predicts outcome in chronic low back pain.* Phys Ther, 2013. **93**(4): p. 470-8.
23. Lumley, M.A., et al., *The working alliance and Clinician-assisted Emotional Disclosure for rheumatoid arthritis.* J Psychosom Res, 2018. **104**: p. 9-15.
24. Barradell, S., T. Peseta, and S. Barrie, *'There's so much to it': the ways physiotherapy students and recent graduates experience practice.* Adv Health Sci Educ Theory Pract, 2017.
25. Greving, K., et al., *Incidence, prevalence, and consultation rates of shoulder complaints in general practice.* Scand J Rheumatol, 2012. **41**(2): p. 150-5.
26. Dorrestijn, O., et al., *Patients with shoulder complaints in general practice: consumption of medical care.* Rheumatology (Oxford), 2011. **50**(2): p. 389-95.
27. Winters, J.C., et al., *NHG-Standaard Schouderklachten (Tweede herziening).* Huisarts Wet 2008(2008:51(11):555-565).
28. Linsell, L., et al., *Prevalence and incidence of adults consulting for shoulder conditions in UK primary care; patterns of diagnosis and referral.* Rheumatology (Oxford), 2006. **45**(2): p. 215-21.
29. Leemrijse, C.J., I.C. Swinkels, and C. Veenhof, *Direct access to physical therapy in the Netherlands: results from the first year in community-based physical therapy.* Phys Ther, 2008. **88**(8): p. 936-46.
30. Barten, D.J. and L. Koppes, *Zorg door de fysiotherapeut; jaarcijfers 2016 en trendcijfers 2012-2016.* NIVEL zorgregistraties., 2017.
31. Kooijman, M., et al., *Patients with shoulder syndromes in general and physiotherapy practice: an observational study.* BMC Musculoskelet Disord, 2013. **14**: p. 128.
32. Kuijpers, T., et al., *Costs of shoulder pain in primary care consulters: a prospective cohort study in The Netherlands.* BMC Musculoskelet Disord, 2006. **7**: p. 83.
33. Potter, M., S. Gordon, and P. Hamer, *The physiotherapy experience in private practice: the patients' perspective.* Aust J Physiother, 2003. **49**(3): p. 195-202.
34. Hush, J.M., et al., *Patient satisfaction with musculoskeletal physiotherapy care in Australia: an international comparison.* J Man Manip Ther, 2012. **20**(4): p. 201-8.
35. van der Windt, D.A., et al., *The responsiveness of the Shoulder Disability Questionnaire.* Ann Rheum Dis, 1998. **57**(2): p. 82-7.
36. Mintken, P.E., P. Glynn, and J.A. Cleland, *Psychometric properties of the shortened disabilities of the Arm, Shoulder, and Hand Questionnaire (QuickDASH) and Numeric Pain Rating Scale in patients with shoulder pain.* J Shoulder Elbow Surg, 2009. **18**(6): p. 920-6.
37. Page, M.J., et al., *Identifying a core set of outcome domains to measure in clinical trials for shoulder disorders: a modified Delphi study.* RMD Open, 2016. **2**(2): p. e000380.
38. Mokkink, L.B., et al., *The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content.* BMC Med Res Methodol, 2010. **10**: p. 22.
39. Beaton, D.E., et al., *Guidelines for the process of cross-cultural adaptation of self-report measures.* Spine (Phila Pa 1976), 2000. **25**(24): p. 3186-91.
40. Wang, W.L., H.L. Lee, and S.J. Fetzer, *Challenges and strategies of instrument translation.* West J Nurs Res, 2006. **28**(3): p. 310-21.
41. Bot, S.D., et al., *Clinimetric evaluation of shoulder disability questionnaires: a systematic review of the literature.* Ann Rheum Dis, 2004. **63**(4): p. 335-41.
42. Roy, J.S., J.C. MacDermid, and L.J. Woodhouse, *Measuring shoulder function: a systematic review of four questionnaires.* Arthritis Rheum, 2009. **61**(5): p. 623-32.
43. Breckenridge, J.D. and J.H. McAuley, *Shoulder Pain and Disability Index (SPADI).* J Physiother, 2011. **57**(3): p. 197.

44. Jansen, M.J., et al., *KNGF Evidence Statement Subacromiale klachten.* Nederlands Tijdschrift voor Fysiotherapie, 2011. **121**(1).
45. Hughes, P.C., N.F. Taylor, and R.A. Green, *Most clinical tests cannot accurately diagnose rotator cuff pathology: a systematic review.* Aust J Physiother, 2008. **54**(3): p. 159-70.
46. Hegedus, E.J., et al., *Physical examination tests of the shoulder: a systematic review with meta-analysis of individual tests.* Br J Sports Med, 2008. **42**(2): p. 80-92; discussion 92.
47. Beaudreuil, J., et al., *Contribution of clinical tests to the diagnosis of rotator cuff disease: a systematic literature review.* Joint Bone Spine, 2009. **76**(1): p. 15-9.
48. McKiernan, S., P. Chiarelli, and H. Warren-Forward, *Diagnostic ultrasound use in physiotherapy, emergency medicine, and anaesthesiology.* Radiography, 2010. **16**(2): p. 154-159.
49. de Jesus, J.O., et al., *Accuracy of MRI, MR arthrography, and ultrasound in the diagnosis of rotator cuff tears: a meta-analysis.* AJR Am J Roentgenol, 2009. **192**(6): p. 1701-7.
50. Ottenheijm, R.P., et al., *Accuracy of diagnostic ultrasound in patients with suspected subacromial disorders: a systematic review and meta-analysis.* Arch Phys Med Rehabil, 2010. **91**(10): p. 1616-25.
51. Smith, T.O., et al., *Diagnostic accuracy of ultrasound for rotator cuff tears in adults: a systematic review and meta-analysis.* Clin Radiol, 2011. **66**(11): p. 1036-48.
52. Rutten, M.J., G.J. Jager, and L.A. Kiemeney, *Ultrasound detection of rotator cuff tears: observer agreement related to increasing experience.* AJR Am J Roentgenol, 2010. **195**(6): p. W440-6.
53. Le Corroller, T., et al., *Sonography of the painful shoulder: role of the operator's experience.* Skeletal Radiol, 2008. **37**(11): p. 979-86.
54. Alavekios, D.A., et al., *Longitudinal analysis of effects of operator experience on accuracy for ultrasound detection of supraspinatus tears.* J Shoulder Elbow Surg, 2013.
55. Murphy, R.J., et al., *An independent learning method for orthopaedic surgeons performing shoulder ultrasound to identify full-thickness tears of the rotator cuff.* J Bone Joint Surg Am, 2013. **95**(3): p. 266-72.
56. Scholten-Peeters, G.G., et al., *The opinion and experiences of Dutch orthopedic surgeons and radiologists about diagnostic musculoskeletal ultrasound imaging in primary care: A survey.* Man Ther, 2013.
57. Hush, J.M., K. Cameron, and M. Mackey, *Patient satisfaction with musculoskeletal physical therapy care: a systematic review.* Phys Ther, 2011. **91**(1): p. 25-36.
58. Karran, E.L., et al., *The impact of choosing words carefully: an online investigation into imaging reporting strategies and best practice care for low back pain.* PeerJ, 2017. **5**: p. e4151.
59. Hall, A.M., et al., *The influence of the therapist-patient relationship on treatment outcome in physical rehabilitation: a systematic review.* Phys Ther, 2010. **90**(8): p. 1099-110.
60. Page, M.J., et al., *Manual therapy and exercise for rotator cuff disease.* Cochrane Database Syst Rev, 2016(6): p. CD012224.
61. Green, S., R. Buchbinder, and S. Hetrick, *Physiotherapy interventions for shoulder pain.* Cochrane Database Syst Rev, 2003(2): p. CD004258.
62. Gebremariam, L., et al., *Effectiveness of surgical and postsurgical interventions for the subacromial impingement syndrome: a systematic review.* Arch Phys Med Rehabil, 2011. **92**(11): p. 1900-13.
63. Coghlan, J.A., et al., *Surgery for rotator cuff disease.* Cochrane Database Syst Rev, 2008(1): p. CD005619.
64. Kromer, T.O., et al., *Effects of physiotherapy in patients with shoulder impingement syndrome: a systematic review of the literature.* J Rehabil Med, 2009. **41**(11): p. 870-80.
65. Saltychev, M., et al., *Conservative treatment or surgery for shoulder impingement: systematic review and meta-analysis.* Disabil Rehabil, 2015. **37**(1): p. 1-8.

66. Littlewood, C., et al., *Exercise for rotator cuff tendinopathy: a systematic review.* Physiotherapy, 2012. **98**(2): p. 101-9.
67. Toliopoulos, P., et al., *Efficacy of surgery for rotator cuff tendinopathy: a systematic review.* Clinical Rheumatology, 2014.
68. Karel, Y., et al., *Physiotherapy for patients with shoulder pain in primary care: a descriptive study of diagnostic- and therapeutic management.* Physiotherapy, 2017. **103**(4): p. 369-378.
69. Jette, D.U., et al., *Use of standardized outcome measures in physical therapist practice: perceptions and applications.* Phys Ther, 2009. **89**(2): p. 125-35.
70. Swinkels, R.A., et al., *Current use and barriers and facilitators for implementation of standardised measures in physical therapy in the Netherlands.* BMC Musculoskelet Disord, 2011. **12**: p. 106.

Thoomes-de Graaf M., Scholten-Peeters GGM, Schellingerhout JM, Bourne AM, Buchbinder R, Koehorst M, Terwee CB, Verhagen AP.

Qual Life Res. 2016 Sep;25(9):2141-60. doi: 10.1007/s11136-016-1277-7.
[Epub 2016 Apr 2].

EVALUATION OF MEASUREMENT PROPERTIES OF SELF-ADMINISTERED PROMS AIMED AT PATIENTS WITH NON-SPECIFIC SHOULDER PAIN AND "ACTIVITY LIMITATIONS": A SYSTEMATIC REVIEW

CHAPTER 2

ABSTRACT

Objective: To critically appraise and compare the measurement properties of self-administered patient reported outcome measures (PROMs) focussing on the shoulder, assessing "activity limitations".

Study design: Systematic review. The study population had to consist of patients with shoulder pain. We excluded postoperative patients or patients with generic diseases. The methodological quality of the selected studies and the results of the measurement properties were critically appraised and rated using the COSMIN checklist.

Results: Out of a total of 3427 unique hits, 31 articles, evaluating 7 different questionnaires, were included. The SPADI is the most frequently evaluated PROM and its measurement properties seem adequate apart from a lack of information regarding its measurement error and content validity.

Conclusion: For English, Norwegian and Turkish users, we recommend to use the SPADI. Dutch users could use either the SDQ or the SST. In German we recommend the DASH. In Tamil, Slovene, Spanish and the Danish language, the evaluated PROMs were not yet of acceptable validity. None of these PROMs showed strong positive evidence for all measurement properties. We propose to develop a new shoulder PROM focused on activity limitations, taking new knowledge and techniques into account.

Keywords: shoulder pain, disability, questionnaire, patient outcome assessment, psychometrics, systematic review

INTRODUCTION

The International Classification of Functioning, Disability and Health (ICF) have described the widely accepted definition of functional health status in terms of "impairments", "activity limitations", and "participation restrictions" [1-3]. For patients with shoulder pain, one of the most important consequences in terms of their health is "activity limitations" [4]. As such, health related patient reported outcome measures (PROMs) that assess perceived "activity limitations" are useful in terms of assessing the physical impairment in patients with shoulder pain.

Several PROMs focusing on the shoulder have been developed to measure "activity limitations" in patients with shoulder pain. Examples of these include the Shoulder Disability Questionnaire (SDQ) [5] and the Shoulder Pain and Disability Index (SPADI) [6]. Furthermore, the Disabilities of the arm, shoulder and hand questionnaire (DASH) is also often used for patients with shoulder pain [7]. There is a great variety in PROMs focusing on patients with shoulder pain. Some PROMs, such as the American Shoulder and Elbow Surgeon questionnaire (ASES), include a physical examination component, while others are completely self-administered. Other PROMs are specifically designed for a subgroup of patients, such as the wheelchair user's shoulder pain index (WUSPI), which is specifically designed for wheelchair users.

Several systematic reviews have evaluated the measurement properties of shoulder specific PROMS. A systematic review which included studies until 2002, found that none of the included 16 PROMs demonstrated satisfactory results for all measurement properties, but overall, the DASH received the best ratings [8]. Another review that assessed the measurement properties of four commonly used shoulder PROMs concluded that none of the questionnaires was superior or could be recommended over the other [9]. A recent review, specifically focused on patients with rotator cuff disorders (RCD), evaluated 12 PROMs and concluded that the included questionnaires showed acceptable psychometric properties for individuals with RCD [10]. Several other reviews have summarized the characteristics and measurement properties of a limited number of PROMs, but these reviews did not assess the methodological quality of the included studies and consequently their conclusions have several limitations [11-13].

Despite the fact that several reviews have been performed, we feel there is a need for a more specific and focused research question. If a research question is broad, it can be difficult to reach conclusions applicable to any single population. For example, a specific description of the patient population is important as it can influence the possibility to reach conclusions [14].

All of the above reviews included studies with mixed populations as well, such as upper extremity disorders. Their recommendations, about PROMs that can be used for patients with shoulder pain explicitly, are partly based on mixed populations, such as patients with solely hand or elbow pain (without shoulder pain). We feel that results of research on psychometric properties of shoulder PROMs should be based on data from patients with shoulder pain only, or should be presented separately. Study populations often consist of patients with "nonspecific" shoulder pain (including rotator cuff disease, frozen shoulder etc.), but can also include patients with serious pathology (e.g. malignancy, infection and fracture), specific diseases (e.g. rheumatoid arthritis) or post-surgery patients. Especially if responsiveness is assessed, this can have consequences on the results. Therefore, we prefer to include only questionnaires assessing shoulder-related disability in patients with non-specific shoulder pain with or without conservative treatment.

Furthermore, these reviews presented their results per PROM and not per language, however due to differences in cultural context, a translation of the original version does not guarantee similar psychometric properties [15, 16]. Therefore, the psychometric qualities of translated PROMs should also be evaluated before they can be used in daily practice or research.

Recently, a new instrument known as the COSMIN checklist has been developed to evaluate the methodological quality of studies investigating the measurement properties of PROMs [17]. This checklist showed a high level of agreement between raters [17, 18]. Since its development, several systematic reviews examined the measurement properties of various PROMS by means of the COSMIN checklist [19-22].

Therefore, the aim of this study was to critically appraise and compare the measurement properties of both the original versions as well as the translated versions of self-administered PROMs focusing on the shoulder assessing "activity limitations" for patients with nonspecific shoulder pain, using the COSMIN checklist.

METHODS

Selection criteria

We included publications concerning the development or validation/evaluation of measurement properties of an original or translated version of a self-administered PROM focussing on the shoulder and assessing "activity limitations". Included patients should have nonspecific shoulder pain as a main complaint. As the definition of adhesive capsulitis, subacromial impingement syndrome and rotator cuff disorders is still unclear and there are no generally accepted criteria yet [23], we consider these pathologies as

nonspecific shoulder pain and not as a specific subgroup. Studies including patients with serious pathology (e.g. malignancy, infection and fracture), specific diseases (e.g. rheumatoid arthritis) or where surgery was applied were excluded, as well as studies that did not report their results separately for patients with shoulder pain. Questionnaires including physical examination (e.g. ASES) were excluded, as well as questionnaires specifically designed for specific subgroups, such as RCD (e.g. Western Ontorio Rotator Cuff Index (WORC)), instability (e.g. Western Ontorio Shoulder Instability Index (WOSI)), athletes (e.g. Athletic shoulder outcome rating scale), or wheelchair users (e.g. WUSPI). We explicitly did not exclude studies in which patients with rotator cuff disorders, instability etc. were used, but we chose to exclude all PROMs that were explicitly designed for a specific subgroup of shoulder complaints, as proposed by their developers.

No language restrictions were applied. Abstracts for which full reports were not available were excluded.

Literature search

Electronic searches included MEDLINE, EMBASE, CINAHL and Cochrane from inception to August 2014. Eligible studies were identified using MeSH (Medline), Thesaurus (EMBASE, CINAHL) and free text words also including specific names of identified PROMs. We used the highly sensitive and precise published search filter [24] for Pubmed searches and used it to build the subsequent search strategies. We have added the MEDLINE search in the appendix, the specific search strings for EMBASE, CINAHL and Cochrane are available from the authors on request. Manual searches of review bibliographies and reference lists of primary studies were also undertaken to search for possible studies not captured by the electronic searches.

A research librarian, together with a review author (MTG) performed the electronic search. Two review authors (MTG, GSP) independently selected the studies to be included by first screening the title and abstract and later assessing the full text papers for eligibility. Disagreements were solved by discussion or through arbitration by a third review author (AV). We listed the excluded studies and their bibliographic details with the reason for exclusion.

Methodological quality

Quality assessment

Two reviewer authors (MTG and either JS, AB, MK or CT) independently performed the assessment of methodological quality, using the COSMIN checklist [17]. Disagreements were solved by discussion or by a third review author (AV). The checklist contains nine boxes, with standards for good methodological quality of studies on nine different measurement properties [17]. The appropriate boxes were selected per study and each item

within this box scored on a 4-point rating scale: "poor", "fair", "good" or "excellent" [25]. An overall score for the methodological quality of a study was determined by taking the lowest rate of any items of the box per measurement property. An intra class coefficient (ICC) was calculated to assess the immediate agreement between both raters on the overall score per box, an ICC higher than 0.70 was considered good [26, 27].

Measurement properties

The measurement properties are divided into three domains: reliability, validity and responsiveness. Information on interpretability and feasibility were also extracted from the studies [17].

Interpretability

Interpretability is defined as: "the degree to which one can assign qualitative meaning -that is, clinical or commonly understood connotations- to an instrument's quantitative scores or changes in scores" [28]. Information about clinically meaningful differences in scores between subgroups, floor and ceiling effects and the minimal important change (MIC) should be provided [17].

Reliability

Reliability is defined as: "the extent to which scores for patients who have not changed, are the same for repeated measurement under several conditions." [28].

The reliability domain contains three measurement properties: internal consistency, reliability, and measurement error [28]. Internal consistency is "the degree of the interrelatedness among the items" of the questionnaire [28] and is measured by Cronbach's alpha or Kuder-Richardson Formula 20 or by using IRT methods [17, 27]. Reliability is "the proportion of the total variance in the measurements which is because of 'true' differences among patients" [28] and is reflected by the Intraclass Correlation Coefficient (ICC) or Cohen's Kappa [17, 27]. The measurement error is "the systematic and random error of a patient's score that is not attributed to true changes in the construct to be measured" [28]. This can be expressed by the standard error of measurement (SEM), the smallest detectable change (SDC) or the limits of agreement (LoA) [17, 27].

Validity

Validity is defined as: "the degree to which an instrument measures the construct(s) it purports to measure" [28]. The validity domain also contains three measurement properties: content validity, criterion validity and construct validity [28]. Content validity is "the degree to which the content of an instrument is an adequate reflection of the construct to be measured" and includes face validity [28]. The definition of face validity is "the degree to which (the items of) an instrument indeed looks as though they are an ad-

equate reflection of the construct to be measured" [28]. In assessing this, it is important to consider whether all items are relevant to the originally described construct [17]. Criterion validity is "the degree to which the scores of an instrument are an adequate reflection of a 'gold standard' " [28]. As PROMs do not have a "gold standard", criterion validity is not appropriate [17]. Construct validity consists of three items:

1. Structural validity is "the degree to which the scores of an instrument are an adequate reflection of the dimensionality of the construct to be measured" [28]. Factor analysis should be used to determine or confirm existing subscales, which are subsequently used in the hypotheses that are being tested [28].
2. Hypotheses-testing is "the degree to which the scores of an instrument are consistent with hypotheses (for instance with regard to internal relationships, relationships to scores of other instruments or differences between relevant groups. Based on the assumption that the instrument validly measures the construct to be measured)" [28].
3. Cross-cultural validity is "the degree to which the performance of the items on a translated or culturally adapted instrument is an adequate reflection of the performance of the items of the original version of the instrument" [28].

Responsiveness

Responsiveness is defined as: "the ability of an instrument to detect changes over time in the construct to be measured" [28]. Responsiveness is considered to be similar to validity, however, while validity refers to the validity of a single score, responsiveness refers to the validity of a change score [17].

Data extraction

Two review authors independently performed data extraction (MTG and either JS, AB, MK or CB). Disagreements were resolved by discussion or by a third review author (AV). Descriptive data extracted included the characteristics of the study population (e.g. age, gender, type of shoulder pain, language); general characteristics of the instruments (e.g. construct, subscales, number of items); whether the PROM was an original version or a translated version of the questionnaire and feasibility. Although feasibility is not captured within the COSMIN checklist, the practical use of a questionnaire is important to determine usefulness in clinical practice. Feasibility includes the time needed to complete the questionnaire, its comprehensibility and whether or not it is generally accepted in clinical practice.

Besides result of the measurement properties and of the interpretability were extracted. Only studies that were ranked as being of fair to excellent methodology were rated on their measurement properties, as studies of poor methodology are of limited value [19, 20].

To rate the results of measurement properties, generally accepted criteria were used [27].

Analysis

To determine the overall quality of the measurement properties of the different questionnaires we combined the different studies per PROM (for each language) by combining their results (ratings), adjusted for the methodological quality (fair, good or excellent) and the consistency of their results. The overall rating for a measurement property was recorded as "positive", "indeterminate", or "negative". Furthermore, we assessed a level of evidence (strong, moderate, limited, conflicting, unknown) using the COSMIN-checklist in a similar manner to that proposed by the Cochrane Review Group (see Table 1) [29].

TABLE 1. Levels of evidence for the overall quality of the measurement property

Level	Rating[1]	Criteria[2]
Strong	+++ OR - - -	Consistent findings among multiple studies of good/ excellent methodological quality
Moderate	++ OR --	Consistent finding among multiple studies of fair studies or in one study of good methodological quality
Limited	+ OR -	One study of fair methodological quality
Conflicting	+/-	Conflicting findings
Unknown	?	Only studies of poor methodological quality
No evidence	0	No studies available

Legend:

1. Rating is based on table 1 per study, where + refers to a positive result and – for a negative result.
2. The criteria of methodological quality are based on the COSMIN-checklist.

We made recommendations concerning the use of a certain PROM per language, based upon the best evidence synthesis. Ideally a PROM should have strong positive evidence on all measurement properties; however, if there was moderate evidence a recommendation was still made. In case multiple PROMs showed similar ratings in a specific language, both were presented. If there were no studies with at least fair methodology, no recommendations were made and if there was only limited evidence, caution was advised.

RESULTS

The search strategy resulted in a total of 3421 hits. Of these, 161 articles were selected based on their title and abstract. Reference checking resulted in 6 additional studies. Evaluation of the full text articles resulted in exclusion of 136 articles. Finally, 31 articles, evaluating 7 different questionnaires, were included (see Figure 1).

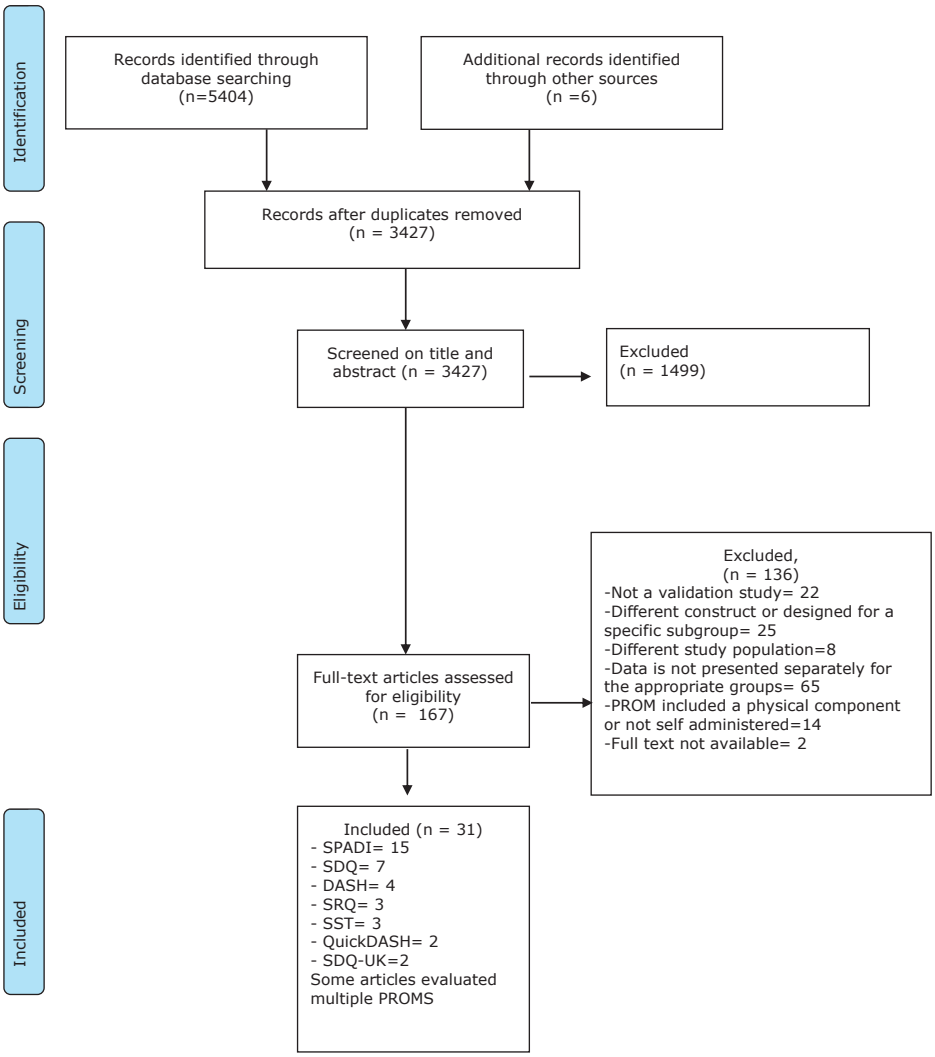


FIGURE 1. Inclusion

The characteristics of the included studies are described in Table 2. For some articles less boxes were scored than described by their original authors, as they did not present these results for our target population separately. The agreement between both raters on the methodological overall quality per box was good (ICC two way random-agreement = 0.88 (95%CI 0.818-0.915). There was no need to discuss disagreement with the third review author. All original versions were developed in English, except the SDQ, which was originally developed in Dutch. The originally described construct and examples of questions of each PROM are described in Table 3. The methodological quality of the studies is presented in Table 4 for each PROM for each measurement property. The main categories with poor

methodology were internal consistency, reliability and cross-cultural validity. The comparator instruments that were used for construct hypothesis testing (except studies of poor methodology) are presented in Table 5. The best evidence synthesis of results per language (per PROM) and their accompanying level of evidence is presented in Table 6.

Below we will describe the results per questionnaire.

TABLE 2. Characteristics of the included studies

Study	Country	PROMs	Setting	Population
English				
Beaton et al. [44]	Canada/ USA	DASH	Hospital	Mixed types of shoulder pain Mean age 53, 43% male *
Cloke et al. [63]	UK	SPADI	Shoulder clinic	Subacromial impingement Mean age 55, 44% male
Croft et al. [54]	UK	SDQ-UK	GP	Shoulder pain Community- mean age 65, 28% male; General practice attendees- mean age 51, 48% male
Fan et al. [64]	USA	QuickDASH	Working population	Shoulder pain Mean age 40, 52% male*
Godfrey et al. [53]	USA	SST	Hospital	Rotator cuff disease Mean age 42, 67% male
Hill et al. [34]	Australia	SPADI	General population	Shoulder pain or stiffness Mean age 56, 41% male
L'Insalata et al. [47]	USA	SRQ	Hospital	Mixed types of shoulder pain Mean age 40, 73% male
MacDermid et al. [39]	Canada	SPADI	General population	Shoulder pain Mean age 44, 49% male
Mintken et al. [52]	USA	QuickDASH	Physiotherapy	Shoulder pain Stable patients- mean age 44, 59% male; Improved patients- mean age 39, 66% male
Paul et al. [31]	UK	SDQ SDQ-UK SPADI SRQ	Shoulder clinic	Shoulder pain Mean age 54, 50% male
Roach et al. [6]	USA	SPADI	GP	Shoulder pain Mean age 58, 100% male
Staples et al. [40]	Australia	SPADI DASH	Physiotherapy	Adhesive capsulitis Mean age 56, 25% male
Tashjian et al [51]	USA	SST	GP	Rotator cuff disease Mean age 51, 48 % male
Dutch				
Heiden, van der et al. [5]	Netherlands	SDQ	Rehabilitation clinic	Shoulder pain and stiffness Mean age 51, 49% male
Kampen van et al. [50]	Netherlands	SST	Hospital	Shoulder pain Mean age 39, 72% male
Vermeulen et al. [48]	Netherlands	SRQ	Hospital	Mixed types of shoulder pain Mean age 52, 23% male

TABLE 2. Characteristics of the included studies *(continued)*

Study	Country	PROMs	Setting	Population
Windt, van der et al. [4]	Netherlands	SDQ	GP	Shoulder pain Mean age 50, 44% male
Winter, de et al. [43]	Netherlands	SDQ	GP	Shoulder pain Mean age 47, 34% male
Norwegian				
Ekeberg et al [37]	Norway	SPADI	GP	Rotator cuff disease Mean age 51, 34% male
Ekeberg et al. [33]	Norway	SPADI	GP	Rotator cuff disease Mean age 51, 37% male
Haldorsen et al. [45]	Norway	DASH	Outpatient clinic	Shoulder impingement Mean age 53, 52% male
Tveita et al. [36]	Norway	SPADI	Hospital	Adhesive capsulitis Not reported
Tveita et al. [35]	Norway	SPADI	Hospital	Adhesive capsulitis Mean age 52, 42% male
Turkish				
Bicer et al. [38]	Turkey	SPADI	Rehabalitation clinic	Shoulder pain Mean age 53, 0% male
Dogu et al. [30]	Turkey	SDQ SPADI	Physiotherapy	Shoulder impingement Mean age 56, 33% male
Ozsahin et al. [42]	Turkey	SDQ	Shoulder clinic	Shoulder pain Mean age 51, 25% male
German				
Offenbacher et al. [65]	Germany	DASH	Hospital	Shoulder pain Mean age 59, 27% male
Danish				
Christiansen et al. [32]	Denmark	SPADI	Hospital	Shoulder pain Mean age 48, 46% male
Spanish				
Alvarez-Nemegyei et al. [66]	Mexico	SDQ	Hospital	Subacromial impingement Mean age 55, 20% male
Slovene				
Jamnik et al. [41]	Slovenia	SPADI	Rehabilitation clinic	Chronic shoulder complaints Mean age 56, 29% male
Tamil				
Jeldi et al. [67]	India	SPADI	Physiotherapy	Shoulder pain or dysfunction Mean age 49, 48% male

*Based on whole cohort, not separately reported for the section of interest

TABLE 3. Overview of PROMs used with their originally described construct and an example of questions used.

PROM	Description of the construct by the original author (and the author of a study assessing content validity)	Example of used questions
SPADI	Pain and disability [6].	1) How severe is your pain when…. When lying on the involved side? 2) How much difficulty did you have…. washing your back?
SDQ	Functional status limitation [5]. Pain related disability [43].	1) My shoulder hurts when I lie on it: Y/N 2) My shoulder is painful when I open or close a door: Y/N
DASH	Symptoms and functional status focused on physical function. The items tap upper extremity-related symptoms and measure functional status at the level of disability. Disability is defined as "difficulty doing activities in any domain of life (the domains typical for one's age-sex group) due to a health or physical problem" [7].	Please circle the number that best describes your physical ability in the past week. Did you have any difficulty: 1) using your usual technique for your work? 2) doing your usual work because of arm, shoulder or hand pain? No difficulty (1)- Unable (5)
SRQ	Symptoms and function [47].	The following questions refer to pain: 1) During the past month, how would you describe the usual pain in your shoulder during activities? Very severe (1) – None (5) The following questions refer to daily activities: 1) During the past month, how much difficulty have you had in each of the following activities due to your shoulder; putting on or removing a pullover sweater or shirt? Unable (1)-No difficulty (5)
SST	Functional limitations of the affected shoulder [49].	1) Can you reach the small of your back to tuck in your shirt with your hand? Y/N 2) Can you place your hand behind your head with the elbow straight out to the side? Y/N
QuickDASH	Physical function and symptoms in persons with any or multiple musculoskeletal disorders of the upper limb [58].	Please rate your ability to do the following activities in the last week by circling the number below the appropriate response. 1) Open a tight or new jar 2) Do heavy household chores (e.g. wash walls, floors) No difficulty (1)-Unable (5)
SDQ-UK	Disability associated with shoulder symptoms [54].	1) Because of my shoulder, I move my arm or hand with some difficulty: Y/N 2) I do not bath myself completely because of my shoulder: Y/N

TABLE 4. Methodological quality of each study per measurement property

Study	Internal consistency	Reliability	Measurement error	Content validity	Structural validity	Hypotheses testing	Cross cultural validity/ *only a translation	Responsive-ness
SPADI developed in English								
Bicer et al. [38]	Poor	Fair				Fair		
Christiansen et al. [32]	Poor	Poor	Poor			Fair	Poor	
Cloke et al. [63]		Poor				Poor		Poor
Dogu et al. [30]								Poor
Ekeberg et al [37]	Poor	Good	Good			Fair		
Ekeberg et al. [33]								Good
Hill et al. [34]	Excellent				Good	Poor		
Jamnik et al. [41]	Poor	Poor			Poor	Fair	Fair*	Poor
Jeldi et al. [67]	Poor	Poor				Poor	Poor	
MacDermid et al. [39]	Fair				Fair	Fair		Poor
Paul et al. [31]						Fair		Fair
Roach et al. [6]	Poor	Poor			Poor	Poor		Poor
Staples et al. [40]						Fair		Fair
Tveita et al. [36]		Fair	Fair				Fair*	Poor
Tveita et al. [35]	Fair				Fair			
SDQ developed in Dutch								
Alvarez-Nemegyei et al. [66]	Poor	Poor					Poor	
Dogu et al. [30]								Poor
Heiden, van der et al. [5]								Fair
Ozsahin et al. [42]	Poor	Fair				Poor	Poor*	
Paul et al. [31]						Fair		Fair
Windt, van der et al. [4]								Good
Winter, de et al. [43]	Poor			Excellent		Fair		
DASH developed in English								
Beaton et al. [44]						Fair		
Haldorsen et al. [45]	Poor	Fair	Fair			Fair		
Offenbacher et al. [65]	Poor	Poor				Fair	Poor*	
Staples et al. [40]						Fair		Fair
SRQ developed in English								
L'Insalata et al. [47]	Poor	Poor				Poor		
Paul et al. [31]						Fair		Fair
Vermeulen et al. [48]	Poor	Fair				Poor	Excellent*	

TABLE 4. Methodological quality of each study per measurement property (*continued*)

Study	Internal consistency	Reliability	Measurement error	Content validity	Structural validity	Hypotheses testing	Cross cultural validity/ *only a translation	Responsive-ness
SST developed in English								
Godfrey et al. [53]						Poor		
Kampen van et al. [50]	Excellent	Fair	Fair		Excellent	Good	Fair*	
Tasjian et al. [51]								Poor
QuickDASH developed in English								
Fan et al. [64]						Poor		
Mintken et al. [52]		Poor	Poor					Fair
SDQ-UK developed in English								
Croft et al. [54]				Poor		Poor		
Paul et al. [31]						Fair		Fair

TABLE 5. Comparator instrument in case of hypothesis testing

Study	Comparator instruments and correlations
SPADI	
Bicer et al. [38]	Convergent; The spearman correlation with the HAQ total score was 0.67 and 0.65 with VAS during AROM.
Christiansen et al. [32]	Known groups; Those currently working, despite their shoulder pain, were found to have significantly lower scores than those not working; the mean difference was –18.3 (95% CI –29.4 to –7.2).
Ekeberg et al. [37]	Convergent; The spearman correlation with the OSS total score was 0.57, -0.67 for the WORC total, -0.75 with WORC physical, -0.46 with WORC Sports, -0.55 with WORC Work and -0.69 with WORC Lifestyle. Divergent; The spearman correlation between the SPADI and the WORC emotions was -0.31.
Jamnik et al. [41]	Known groups; Participants who differed in the severity of the perceived disability self-rating (mild-moderate-severe) differed significantly in the SPADI score in the presumed order.
MacDermid et al. [39]	Known groups; Patients who had diagnosed shoulder problems and those on pain medication reported significantly higher pain and disability scores. Convergent; Convergent scales (Home management 0.59, Work -0.10, Physical dimension 0.51) of the SIP showed a moderate correlation, except the work scale. Divergent; Divergent (emotional) scales of the SIP showed low correlations (0.17-0.33). *
Paul et al. [31]	Convergent; The spearman correlation with other shoulder PROMs was: 0.57 for the SDQ-UK, 0.33 with the SDQ and 0.83 with the SRQ. The correlation with Difficulty VAS 0.62. *
Staples et al. [40]	Convergent; The Pearson correlation with other shoulder PROMs was: 0.55 with the DASH and 0.65 with the Croft index. Correlations with generic PROMs were: 0.17 with PET, 0.60 with Pain and 0.55 with the HAQ.

TABLE 5. Comparator instrument in case of hypothesis testing (*continued*)

Study	Comparator instruments and correlations
SDQ	
Paul et al. [31]	Convergent; The spearman correlation with other shoulder PROMs was: 0.55 for the SDQ-UK, 0.33 with the SPADI and 0.43 with the SRQ. The correlation with Difficulty VAS 0.47. *
Winter, de et al. [43]	Known groups; Significant differences in the SDQ- scores (P < 0.001) were found for subgroups with different pain severity, ability to perform activities in daily life, mobility, muscle force, and levels of disability according to the physical therapists. Convergent; the spearman correlation with severity of disability was 0.58, degree of difficulty for the main functional limitation was 0.32. *
DASH	
Beaton et al. [44]	Known groups; Those currently working with their upper limb condition and able to continue doing so had significantly lower disability than those who were not able to work (26.8 vs. 50.7, t=-7.51, p<0.001). Statistically significant differences were also found between those who were able to do all they want to do as opposed to those who were not able to do so (23.6 vs. 47.1, t = -5.81, P<0.0001). Convergent; the spearman correlation with the overall rating of the problem was 0.68, with the ability to function 0.85, with the ability to work 0.76, with Brigham symptoms 0.71 and 0.90 with Brigham symptoms. The spearman correlation with another shoulder PROM 0.76 with the SPADI pain scale and 0.83 with the SPADI function scale. *
Haldorsen et al. [45]	Convergent: The Pearson correlation with the SPADI was 0.75 and with the NPRS 0.58. The correlations with components of the SF-36 were: physical functioning –0.48, bodily pain –0.62, and physical component summary –0.59. Divergent: The Pearson correlation with the mental component summary score of the SF-36 was –0.17 and -0.35 with the social functioning scale of the SF-36.
Offenbacher et al. [65]	Convergent; the spearman correlation with the HAQ was 0.81, with the SF-36 physical functioning component -0.58, and with global impact 0.76. *
Staples et al. [40]	Convergent; The Pearson correlation with other shoulder PROMs was: 0.55 with the SPADI and 0.65 with the Croft index. Correlations with generic PROMs were: 0.20 with PET and 0.54 with the HAQ. *
SRQ	
Paul et al. [31]	Convergent; The spearman correlation with other shoulder PROMs was: 0.72 for the SDQ-UK, 0.83 with the SPADI and 0.43 with the SDQ. The correlation with Difficulty VAS 0.60. *
SST	
Kampen van et al. [50]	Convergent; the Pearson correlation with other shoulder PROMs was: 0.74 with the OSS, 0.59 with the CM, 0.74 with the DASH. The correlation with the SF-36 subscale physical functioning was 0.56.
SDQ-UK	
Paul et al. [31]	Convergent; The spearman correlation with other shoulder PROMs was: 0.72 for the SRQ, 0.57 with the SPADI and 0.55 with the SDQ. The correlation with Difficulty VAS 0.41. *

** ROM, pain alone and the EQ5D were considered to be inappropriate comparators and were therefore excluded in the rating process.*

TABLE 6. Best evidence synthesis

PROM	Internal consistency	Reliability	Measurement error	Content validity	Structural validity	Hypotheses testing	Cross cultural validity	Responsive-ness
English								
SPADI	+++	?	0	0	++	++	0	++
DASH	0	0	0	0	0	++	0	+
SDQ-UK	0	0	0	?	0	+	0	+
SRQ	?	?	0	0	0	+	0	+
SDQ-English	0	0	0	0	0	-	0	+
SST	0	0	0	0	0	?	0	?****
QuickDASH	0	0	0	0	0	?	0	0
Dutch								
SST-Dutch	+++	+	?*	0	+++	++	0	0
SDQ	?	0	0	?**	0	+	0	++
Quick DASH-Dutch	0	?	?	0	0	0	0	+
SRQ-Dutch	?	+	0	0	0	?	0	0
Norwegian								
SPADI-Norwegian	+	++	?*	0	-	+	0	++
DASH-Norwegian	?	+	?*	0	0	+	0	0
Turkish								
SPADI-Turkish	?	+	0	0	0	+	0	?
SDQ-Turkish	?	+	0	0	0	?	0	?
German								
DASH-German	?	?	0	0	0	+	0	0
Danish								
SPADI-Danish	?	?	?	0	0	?***	?	0
Spanish								
SDQ-Spanish (Mexican)	?	?	0	0	0	0	?	0
Slovene								
SPADI-Slovene	?	?	0	0	?	?***	0	?
Tamil								
SPADI-Tamil	?	?	0	0	0	?	?	0

* Despite fair/good methodology, the level of evidence could not be determined as the appropriate measurement properties were not provided.

** Despite fair/good methodology, the level of evidence could not be determined as the originally described construct differed from the construct described in the current study.

*** Despite fair/good methodology, the level of evidence could not be determined as unclear, as they confirmed their hypothesis with known group validity, but did not assess whether the correlations with related constructs were higher than with unrelated constructs.

**** This study only evaluated the minimal clinical difference.

Shoulder Pain and Disability Index (SPADI)

The SPADI was developed to measure pain and disability associated with shoulder pathology. It consists of 13 items, each scored on a 0-10 numeric rating scale, divided into two subscales: pain (5 items) and disability (8 items). The total score varies between 0 and 100 [6]. It takes approximately 2 to 3 minutes to complete [30, 31]. The SPADI is considered to be easy to understand by patients [31] and no floor or ceiling effects have been detected [32, 33].

Reliability

Internal consistency: There is strong positive evidence for internal consistency within the English SPADI (Cronbach Alpha = 0.85 for pain and 0.90 for disability) [34]. There is also limited positive evidence for the internal consistency of the Norwegian SPADI (Cronbach Alpha = 0.80 for pain and 0.87 for disability) [35]. However, there were inconsistent findings on the factor structure of the SPADI, therefore these results should be interpreted with caution.

Reliability: Both the Norwegian and the Turkish versions showed moderate (ICC= 0.85-0.89) [36, 37] and limited positive evidence (ICC= 0.92)[38] respectively. Studies evaluating other language-versions were rated as having poor methodology.

Measurement error: Two studies (both Norwegian) were rated as having at least "fair" methodology that evaluated measurement error, one study of fair methodology only reported an SDC (17 points), but no MIC was determined [36]. The other study reported an SDC of 19.7 and the Loa was between -20.9 and 18.5 [37], the MIC however, ranged between 15.0 and 31.1 depending on the methods used [33], the authors therefore concluded that a change of approximately 20 points is necessary for patient perceived important change.

Validity

Content validity: There were no studies evaluating content validity.

Construct structural validity: There is moderate evidence that the English SPADI consists of two factors; pain and disability and all factors are loaded accordingly as originally proposed by Roach [34]. In contrast, there is limited evidence that not all items are loaded on the original factor but no explained variance was described [39]. Factor analysis of the Norwegian SPADI resulted in limited evidence that it consists of two factors but the original factor structure could not be confirmed, as not all items loaded as originally intended [35].

Construct hypothesis testing: In terms of construct hypothesis testing, moderate positive evidence was identified for the English SPADI [31, 39, 40]. There was limited positive evidence for the Turkish version [38] and the Norwegian version [37]. The evidence for the Danish SPADI [32] and the Slovenish version [41] was unclear, as they confirmed their

hypothesis with known group validity, but did not assess whether the correlations with related constructs were higher than with unrelated constructs.
Construct cross cultural validity: Only studies that were rated as being of poor methodology have been performed.

Responsiveness
There is moderate positive evidence for responsiveness of the English version (AUC ranging between 0.74 and 0.87) [31, 40] and the Norwegian version (AUC= 0.84 or 0.92 depending on the follow-up period) [33].

Shoulder Disability Questionnaire (SDQ)

The SDQ is 16-item pain-related disability questionnaire that was originally developed in Dutch. Response options are "yes", "no" or "not applicable", resulting in a total score which ranges from 0 to 100, with a higher score indicating more severe disability [4]. It takes about 2 [30, 31] to 4 minutes to complete and patients indicated the SDQ as (very) easy to complete [5, 30, 31]. One study assessed whether there were signs of floor or ceiling effects; however, they did not report the data needed to give a proper indication of it [5].

Reliability
Internal consistency: Only studies that were rated as being of poor methodology have been performed.
Reliability: There were no sound methodological studies evaluating reliability, except for the Turkish version, which showed limited positive evidence, with a Pearson correlation coefficient of 0.88 for the total score [42].
Measurement error: There were no studies evaluating the measurement error.

Validity
Content validity: The evidence regarding content validity of the original SDQ is indeterminate, as the questions are not aimed at the originally described construct (see Table 4).
Construct structural validity: There were no studies evaluating structural validity.
Construct hypothesis testing: There is limited positive evidence for the Dutch version [43] and limited negative evidence for the English version (as three out of the seven expected positive correlations measured were below 0.50) [31].
Construct cross cultural validity: No studies specifically assessed cross cultural validity.

Responsiveness

There is moderate positive evidence for the Dutch version (AUC= 0.84) [4] and limited positive evidence for the English version (AUC= 0.77) [31].

Disability of arm, shoulder and hand (DASH)

The DASH is designed to measure symptoms and physical functioning in patients with pain in the arm, shoulder or hand. It consists of 30 items and the response options for each item are presented as 5-point Likert scales. The total score ranges from 0 to 100 [7]. We did not find studies reporting any item on feasibility. No floor or ceiling effects were detected [44, 45].

Reliability

Internal consistency: Only studies that were rated as being of poor methodology have been performed.

Reliability: There is limited positive evidence for the Norwegian version (ICC= 0.89) [45].

Measurement error: The result of the only study with fair methodology evaluating measurement error is indeterminate, as they did not provide the MIC; the SDC however, was 6.7 points for the Norwegian version [45].

Validity

Content validity: There were no studies evaluating content validity.

Construct structural validity: There were no studies evaluating structural validity.

Construct hypothesis testing: There is moderate positive evidence for construct hypothesis testing of the English version [40, 44] and limited positive evidence for the German [46] and Norwegian version [45].

Construct cross cultural validity: No studies specifically assessed cross cultural validity.

Responsiveness

There is limited positive evidence for the English version for responsiveness (AUC= 0.71-0.86 depending on the anchor used) [40].

Shoulder Rating Questionnaire (SRQ)

The SRQ was developed to measure the severity of symptoms related to and the functional status of the shoulder. It covers seven domains including 21 items -the total score ranges between 17 and 100 [47]- takes about 4 [31] to 7 [48] minutes to complete and is moderately easy to complete according to patients [31].

Reliability

Internal consistency: Only studies that were rated as being of poor methodology have been performed.

Reliability: There was limited positive evidence for the reliability of the Dutch version (ICC=0.85) [48].

Measurement error: There were no studies evaluating the measurement error.

Validity

Content validity: There were no studies evaluating content validity.

Construct structural validity: There were no studies evaluating structural validity.

Construct hypothesis testing: There was limited positive evidence for the English SRQ [31].

Construct cross cultural validity: No studies specifically assessed cross cultural validity.

Responsiveness

There was limited positive evidence for the responsiveness of the English SRQ (AUC= 0.85) [31].

Simple Shoulder Test (SST)

The SST was developed to measure functional limitations in patients with shoulder dysfunction. It consists of 12 items, and the response options are dichotomous. The total score ranges between 0 and 12 [49]. We did not find studies reporting any item on feasibility. No floor or ceiling effects were detected [50].

Reliability

Internal consistency: There was strong positive evidence for the Dutch SST with a Cronbach Alpha of 0.78 [50].

Reliability: There was limited positive evidence for the reliability of the Dutch SST (ICC=0.92) [50].

Measurement error: The result of the only study with fair methodology evaluating measurement error is indeterminate, as they did not provide the MIC; the SDC however was 3.3 [50].

Validity

Content validity: There were no studies evaluating content validity.

Construct structural validity: There was strong evidence for the unidimensionality of the Dutch SST. Confirmatory factor analysis of a 1-factor model showed a moderate fit (CFI 0.94, TLI 0.93, RMSEA 0.07), and three items showed relatively low factor loadings [50].

Construct hypothesis testing: There is moderate positive evidence for construct hypothesis testing of the Dutch SST [50].
Construct cross cultural validity: No studies specifically assessed cross cultural validity.

Responsiveness
There were no studies judged as having a sound methodology evaluating the English version. One study on the English SST only calculated the minimal clinically important difference but did not assess the responsiveness [51].

QuickDASH

The QuickDASH is an 11-item questionnaire that addresses symptoms and physical function in people with disorders of the arm, shoulder or hand. It provides a summative percentage score, with 100 indicating the most disability [52]. We did not find studies reporting on feasibility. No floor or ceiling effects were detected [53].

Reliability
Internal consistency: There were no studies evaluating internal consistency.
Reliability: Only studies that were rated as being of poor methodology have been performed.
Measurement error: Only studies that were rated as being of poor methodology have been performed.

Validity
Content validity: There were no studies evaluating content validity.
Construct structural validity: There were no studies evaluating structural validity.
Construct hypothesis testing: Only studies that were rated as being of poor methodology have been performed.
Construct cross cultural validity: No studies specifically assessed cross cultural validity.

Responsiveness
There was limited positive evidence for responsiveness in the Dutch version (AUC= 0.82) [52].

Shoulder Disability Questionnaire (SDQ-UK)

The SDQ-UK is a 22-item questionnaire [54]. The questionnaire contains some statements that people have used to describe themselves when they have trouble with their shoulder. Participants are asked to answer "yes" or "no" depending on whether they recognize the statement as applying to them, with a total score ranging between 0 and 100. It takes about 3 minutes to complete and patients describe it as easy to understand [31].

Reliability

Internal consistency: There were no studies evaluating internal consistency.
Reliability: There were no studies evaluating reliability.
Measurement error: There were no studies evaluating the measurement error.

Validity

Content validity: Only studies that were rated as being of poor methodology have been performed.
Construct structural validity: There were no studies evaluating structural validity.
Construct hypothesis testing: There was limited positive evidence for construct hypothesis testing [31].
Construct cross cultural validity: No studies specifically assessed cross cultural validity.

Responsiveness

There was limited positive evidence for the responsiveness (AUC= 0.77) [31].

Recommended PROMS per language

English

All seven PROMs were available and assessed in English. For English users we recommend using the English SPADI as it was rated best in the best evidence synthesis. It consists of two factors, there is strong positive evidence for the internal consistency and moderate evidence for construct hypothesis testing and the responsiveness.

Dutch

Four questionnaires were available and assessed in Dutch in this specific population. The SDQ was developed in Dutch, the other three were developed in English. Both the SDQ and SST showed acceptable ratings in the best evidence synthesis. There was strong evidence for the reliability as well as for the construct validity for the Dutch SST. Strong positive evidence was found for the internal consistency and limited positive evidence for the reliability of the Dutch SST, and inconclusive evidence for the measurement error. The construct validity of the SST was strong, as there was strong evidence for the unidimensionality and moderate positive evidence for construct hypothesis testing.

There is limited positive evidence for construct hypothesis testing of the Dutch SDQ, and there is moderate positive evidence for responsiveness. We recommend choosing between either the SST or the SDQ depending on the purpose of its use.

Norwegian

Out of the two available instruments, the SPADI showed the best ratings. There is moderate positive evidence for the reliability and inconclusive evidence for the measurement error. There was limited evidence that the Norwegian SPADI did not follow the original factor structure and limited positive evidence for the internal consistency. There was limited positive evidence for construct hypothesis testing and moderate positive evidence for the responsiveness.

Turkish

In Turkish both the SDQ and the SPADI were evaluated, both only showed limited evidence; however, the SPADI also had limited evidence for construct hypothesis testing instead of only limited evidence for reliability. We therefore recommend using the SPADI, however caution is advised.

German

We only found one study using a PROM in German when using our search criteria. There is limited positive evidence for the construct hypothesis of the German DASH. We recommend using the DASH in the German language; however, it is important to be aware of the lack of information available about this PROM in German.

Other languages

In Danish, Tamil and Slovene, the only instrument evaluated was the SPADI, in Spanish the only questionnaire assessed was the SDQ. For all four languages we only found studies with poor methodology or information was missing regarding a measurement property. We could therefore not make a recommendation in these languages.

DISCUSSION

The SPADI has been the most frequently evaluated questionnaire in this review on patients with shoulder pain and its measurement properties seem adequate apart from a lack of information regarding its reliability, measurement error and content validity. For English users, we recommend its use, as this is the PROM with the best measurement properties.

For Norwegian users the SPADI is recommended, as well for Turkish users, although for the latter caution is advised as the evidence is limited and information on some measurement properties is lacking. Dutch users could use either the SDQ or the SST, depending on the intended purpose. Germans could use the DASH, although caution is although there is still a lack of information regarding many measurement properties.

In Danish, Spanish, Tamil and Slovene, the evaluated PROMs were not yet of acceptable validity. We found no studies concerning PROMs in other languages, which met our inclusion-criteria.

Comparison with the literature

One systematic review, assessing the methodological quality of measurement properties of shoulder PROMs, concluded that the DASH received the best ratings [8]. This is in contrast with our findings. A possible reason for this difference is the search period. Most studies reporting on the SPADI in our review, were published after the search period (2002) of the previous review. Moreover, we excluded studies evaluating the DASH that did not report their results for shoulder pain patients separately.

Another recent review, concluded that all of the included PROMs showed acceptable psychometric properties [10]. This study recommended PROMs that we excluded in our review [10]. The methodological quality of the studies included, ranged from 33.3 to 95.9%. No evidence synthesis was performed, the psychometric properties per PROM were presented but without the methodological quality per study [10].

A review that evaluated the DASH, ASES, SPADI and SST only, concluded that their measurement properties were acceptable and that none of the questionnaires was superior or could be recommended over the other. The quality of the individual studies ranged from 25% to 96% [9]. This study presented the psychometric properties of all included studies but did not use the methodological quality of the studies themselves in their conclusions about the psychometric properties of an instrument.

Our search strategy was designed to be highly sensitive rather than specific, resulting in a higher number of hits (3421) compared to other reviews [8-10, 12]. Two reviews did not describe their search strategy [11, 13], and two reviews also included studies that were not designed to validate a PROM [9, 10].

Most importantly these reviews used an unspecified study population (e.g. including post-operative patients), included PROMs focused on a specific pathology (e.g. instability) and PROMs that included a physical component. We specified our study population and excluded studies that did not report their results for patients with shoulder pain separately. As a consequence, we excluded a high amount of studies that were focused on the DASH. Due to our strict selection criteria, we also excluded a number of well-known PROMs, due to our specific research question, such as the WOSI, a PROM that is designed specifically for patients with instability, or the ASES, which includes a physical component.

The major flaws we found with respect to the methodology are comparable with another study on measurement properties of neck pain and disability questionnaires

[55]. For internal consistency most studies did not measure the unidimensionality of the scale. The time interval and the sample size were the main problems within the reliability category, and sample size or performing a confirmatory analysis for cross-cultural validity.

Strengths and limitations

We excluded two studies because we could not retrieve them as full text papers. One was written in Turkish. This could potentially have led to selection bias. However, the leading journals, and consequently the most important papers, are published in English.

We pooled our results by language rather than by country although we recognize that cultural differences may exist between countries. This means that for the English versions of PROMs, we pooled data from the UK, USA, Canada and Australia, hereby neglecting possible cultural differences. If countries are very close in location/ culture/ use of language and the text does not contain wording about education, health systems, brand names or IT, it is acceptable to use the same language version and to pool data from trials [56]. With respect to this, we assumed there are no insurmountable differences between the UK, USA, Canada and Australia. Moreover, our results did not show inconsistencies regarding measurement properties.

We excluded patients with generic and serious conditions (e.g. rheumatoid arthritis, fractures) and post-operative patients; therefore, our results cannot be extrapolated to these kinds of patients. The DASH is designed for patients with upper extremity disorders. Our conclusion on the DASH and its measurement properties are based on patients with shoulder pain only. Our results are therefore incomplete regarding the measurement properties of the DASH itself and cannot be extrapolated to other groups of patients on which the DASH can be used.

Considerations regarding the results

We found that content validity of most PROMs is still unknown (a PROM should have evidence supporting its content validity, including evidence that patients and/or experts consider the content of the PROM relevant and comprehensive for the concept, population, and aim of the measurement application [57]), although content validity is often considered to be the most important measurement property [57]. We could only rate the SDQ and the SDQ-UK on content validity, as some development studies did not involve patients or did not present their results separately for patients with shoulder pain [6, 7, 47, 49, 58, 59]. Originally the construct of the SDQ was described as "functional status" [5], but the items used were focussed on pain e.g. "my shoulder hurts when I lie on it", resulting in a lack of face validity. However, the study which assessed the content validity of the SDQ, used "pain related disability" [43] as the construct to be measured, which would be a more appropriate term. It is therefore important to clearly

describe the construct to be measured. All other PROMs did not show much discrepancy between the described construct and its items. However, in case of the SRQ, SDQ-UK and SST, the construct was not described in generally accepted terms (ICF terminology) or an extensive description, which makes it difficult to assess whether the items are an adequate reflection of the construct to be measured.

Most studies focused on validity. However internal consistency, reliability and responsiveness were also well represented. For hypothesis testing various comparator instruments were used, shoulder PROMs focused on activity limitation/pain related disability (e.g. SDQ, SDQ-UK, SRQ, DASH, SPADI), known groups (e.g. medication, specific diagnosis, currently working), general PROMs (e.g. pain intensity, HAQ) and range of motion. An important aspect of the methodological quality assessment is whether the comparator instruments measures the same construct and shows adequate measurement properties. We considered that range of motion measures a different construct and we therefore rated studies that solely used range of motion as a comparator instrument as being of poor methodology. We also excluded the comparisons with pain alone and the EQ5D as these also measure a different construct, although in most cases this did not influenced the final ratings.

Recommendations for future research

Further research is recommended to fill the gaps in knowledge regarding the measurement properties of shoulder-specific PROMs, especially with respect to their content validity, starting with a clear description of the construct, but also whether all items seem to be relevant to patients.

Although all of the evaluated instruments were developed in the 90s, none of these PROMs showed strong positive evidence for all measurement properties after twenty years of research. Meanwhile, knowledge regarding the development of a PROM has increased and instrument-developers must articulate how a particular conceptual framework guided their construct selection, item development (including e.g. in-depth interviews and focus groups with patients and experts in the field), and psychometric testing [60]. Also, important issues concerning the limitation of functional activities have changed over time, e.g. computer use is nowadays completely integrated into everyday life, but this is not included in most PROMs. Not only relevant items have been changed, but also the available methodology and technology has reached a new level of sophistication, including "modern" psychometric techniques of item banking, item response theory (IRT) and computer-adaptive testing (CAT) [60]. Recently, the Patient-Reported Outcomes Measurement Information System (PROMIS) was developed using sample qualitative input from patients and IRT methods, to construct and evaluate a preliminary item bank for measuring physical function [61]. At this moment, there are

upper-extremity and mobility subdomain scores from the PROMIS physical functioning adult item bank [62].

Computer-adaptive testing has tremendous potential for yielding precise PROM assessment quickly and with significantly reduced respondent burden [60]. The methods of the PROMIS project are likely to substantially improve measures of physical function and to increase the efficiency of their administration using CAT [61].

We therefore propose to develop a new shoulder PROM focused on activity limitations, or evaluate the usefulness of an instrument such as the upper extremity PROMIS scale on patients with shoulder pain, taking new knowledge and techniques into account.

Our study showed that there is a lack of high quality studies measuring cross-cultural validation. Most often PROMs are being translated and some measurement properties are assessed. We feel it is of great importance to perform cross-cultural validation for PROMs [57].

Funding

This study was not funded and we declare to have no competing interests.

REFERENCES

1. Stucki, G., et al., *ICF-based classification and measurement of functioning.* Eur J Phys Rehabil Med, 2008. **44**(3): p. 315-28.
2. Cieza, A. and G. Stucki, *The International Classification of Functioning Disability and Health: its development process and content validity.* Eur J Phys Rehabil Med, 2008. **44**(3): p. 303-13.
3. Jelsma, J., *Use of the International Classification of Functioning, Disability and Health: a literature survey.* J Rehabil Med, 2009. **41**(1): p. 1-12.
4. Van Der Windt, D.A.W.M., et al., *The responsiveness of the Shoulder Disability Questionnaire.* Ann Rheum Dis, 1998. **57**(2): p. 82-7.
5. Van Der Heijden, G.J.M.G., P. Leffers, and L.M. Bouter, *Shoulder disability questionnaire design and responsiveness of a functional status measure.* J Clin Epidemiol, 2000. **53**(1): p. 29-38.
6. Roach, K.E., et al., *Development of a shoulder pain and disability index.* Arthritis Care Res, 1991. **4**(4): p. 143-9.
7. Hudak, P.L., P.C. Amadio, and C. Bombardier, *Development of an upper extremity outcome measure: The DASH (disabilities of the arm, shoulder, and head).* AM J IND MED, 1996. **29**(6): p. 602-8.
8. Bot, S.D., et al., *Clinimetric evaluation of shoulder disability questionnaires: a systematic review of the literature.* Ann Rheum Dis, 2004. **63**(4): p. 335-41.
9. Roy, J.S., J.C. MacDermid, and L.J. Woodhouse, *Measuring shoulder function: a systematic review of four questionnaires.* Arthritis Rheum, 2009. **61**(5): p. 623-32.
10. St-Pierre, C., et al., *Psychometric properties of self-reported questionnaires for the evaluation of symptoms and functional limitations in individuals with rotator cuff disorders: a systematic review.* Disabil Rehabil, 2015: p. 1-20.
11. Angst, F., et al., *Measures of adult shoulder function: Disabilities of the Arm, Shoulder, and Hand Questionnaire (DASH) and Its Short Version (QuickDASH), Shoulder Pain and Disability Index (SPADI), American Shoulder and Elbow Surgeons (ASES) Society Standardized Shoulder Assessment Form, Constant (Murley) Score (CS), Simple Shoulder Test (SST), Oxford Shoulder Score (OSS), Shoulder Disability Questionnaire.* Arthritis Care Res, 2011. **63**(SUPPL. 11): p. S174-S88.
12. Desai, A.S., A. Dramis, and A.J. Hearnden, *Critical appraisal of subjective outcome measures used in the assessment of shoulder disability.* Ann R Coll Surg Engl, 2010. **92**(1): p. 9-13.
13. Fayad, F., Y. Mace, and M.M. Lefevre-Colau, *[Shoulder disability questionnaires: a systematic review].* Ann Readapt Med Phys, 2005. **48**(6): p. 298-306.
14. Wright, R.W., et al., *How to write a systematic review.* Clin Orthop Relat Res, 2007. **455**: p. 23-9.
15. Beaton, D.E., et al., *Guidelines for the process of cross-cultural adaptation of self-report measures.* Spine (Phila Pa 1976), 2000. **25**(24): p. 3186-91.
16. Wang, W.L., H.L. Lee, and S.J. Fetzer, *Challenges and strategies of instrument translation.* West J Nurs Res, 2006. **28**(3): p. 310-21.
17. Mokkink, L.B., et al., *The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content.* BMC Med Res Methodol, 2010. **10**: p. 22.
18. Mokkink, L.B., et al., *Inter-rater agreement and reliability of the COSMIN (COnsensus-based Standards for the selection of health status Measurement Instruments) checklist.* BMC Med Res Methodol, 2010. **10**: p. 82.
19. Schellingerhout, J.M., et al., *Measurement properties of translated versions of neck-specific questionnaires: a systematic review.* BMC Med Res Methodol, 2011. **11**: p. 87.
20. Schellingerhout, J.M., et al., *Measurement properties of disease-specific questionnaires in patients with neck pain: a systematic review.* Qual Life Res, 2012. **21**(4): p. 659-70.

21. Mutsaers, J.H., et al., *Psychometric properties of the Pain Attitudes and Beliefs Scale for Physiotherapists: a systematic review.* Man Ther, 2012. **17**(3): p. 213-8.
22. van Bloemendaal, M., A.T. van de Water, and I.G. van de Port, *Walking tests for stroke survivors: a systematic review of their measurement properties.* Disabil Rehabil, 2012.
23. Schellingerhout, J.M., et al., *Lack of uniformity in diagnostic labeling of shoulder pain: time for a different approach.* Man Ther, 2008. **13**(6): p. 478-83.
24. Terwee, C.B., et al., *Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments.* Qual Life Res, 2009. **18**(8): p. 1115-23.
25. Terwee, C.B., et al., *Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist.* Qual Life Res, 2012. **21**(4): p. 651-7.
26. Nunally JC, B.I., *Psychometric theory. 3rd edition. New York: McGraw-Hill.* 1994.
27. Terwee, C.B., et al., *Quality criteria were proposed for measurement properties of health status questionnaires.* J Clin Epidemiol, 2007. **60**(1): p. 34-42.
28. Mokkink, L.B., et al., *The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes.* J Clin Epidemiol, 2010. **63**(7): p. 737-45.
29. van Tulder, M., et al., *Updated method guidelines for systematic reviews in the cochrane collaboration back review group.* Spine (Phila Pa 1976), 2003. **28**(12): p. 1290-9.
30. Dogu, B., et al., *Which questionnaire is more effective for follow-up diagnosed subacromial impingement syndrome? A comparison of the responsiveness of SDQ, SPADI and WORC index.* J Back Musculoskelet Rehabil, 2013. **26**(1): p. 1-7.
31. Paul, A., et al., *A comparison of four shoulder-specific questionnaires in primary care.* Ann Rheum Dis, 2004. **63**(10): p. 1293-9.
32. Christiansen, D.H., J.H. Andersen, and J.P. Haahr, *Cross-cultural adaption and measurement properties of the Danish version of the Shoulder Pain and Disability Index.* Clin Rehabil, 2013. **27**(4): p. 355-60.
33. Ekeberg, O.M., et al., *A questionnaire found disease-specific WORC index is not more responsive than SPADI and OSS in rotator cuff disease.* J Clin Epidemiol, 2010. **63**(5): p. 575-84.
34. Hill, C.L., et al., *Factor structure and validity of the shoulder pain and disability index in a population-based study of people with shoulder symptoms.* BMC Musculoskelet Disord, 2011. **12**: p. 8.
35. Tveita, E.K., et al., *Factor structure of the Shoulder Pain and Disability Index in patients with adhesive capsulitis.* BMC Musculoskelet Disord, 2008. **9**.
36. Tveita, E.K., et al., *Responsiveness of the shoulder pain and disability index in patients with adhesive capsulitis.* BMC Musculoskelet Disord, 2008. **9**: p. 161.
37. Ekeberg, O.M., et al., *Agreement, reliability and validity in 3 shoulder questionnaires in patients with rotator cuff disease.* BMC Musculoskelet Disord, 2008. **9**: p. 68.
38. Bicer, A. and H. Ankarali, *Shoulder Pain and Disability Index: A validation study in Turkish women.* Singapore Med J, 2010. **51**(11): p. 865-70.
39. MacDermid, J.C., P. Solomon, and K. Prkachin, *The Shoulder Pain and Disability Index demonstrates factor, construct and longitudinal validity.* BMC Musculoskelet Disord, 2006. **7**: p. 12.
40. Staples, M.P., et al., *Shoulder-specific disability measures showed acceptable construct validity and responsiveness.* J Clin Epidemiol, 2010. **63**(2): p. 163-70.
41. Jamnik, H. and M.K. Spevak, *Shoulder pain and disability Index: Validation of slovene version.* Int J Rehabil Res, 2008. **31**(4): p. 337-41.
42. Ozsahin, M., et al., *Adaptation of the shoulder disability questionnaire to the Turkish population, its reliability and validity.* Int J Rehabil Res, 2008. **31**(3): p. 241-5.

43. de Winter, A.F., et al., *The Shoulder Disability Questionnaire differentiated well between high and low disability levels in patients in primary care, in a cross-sectional study.* J Clin Epidemiol, 2007. **60**(11): p. 1156-63.
44. Beaton, D.E., et al., *Measuring the whole or the parts? Validity, reliability, and responsiveness of the disabilities of the arm, shoulder and hand outcome measure in different regions of the upper extremity.* J Hand Ther, 2001. **14**(2): p. 128-46.
45. Haldorsen, B., et al., *Reliability and validity of the Norwegian version of the Disabilities of the Arm, Shoulder and Hand questionnaire in patients with shoulder impingement syndrome.* BMC Musculoskelet Disord, 2014. **15**(1).
46. Offenbaecher, M., et al., *Validation of a German version of the disabilities of arm, shoulder, and hand questionnaire (DASH-G).* J RHEUMATOL, 2002. **29**(2): p. 401-2.
47. L'Insalata, J.C., et al., *A self-administered questionnaire for assessment of symptoms and function of the shoulder.* J BONE JT SURG SER A, 1997. **79**(5): p. 738-48.
48. Vermeulen, H.M., et al., *Translation, adaptation and validation of the Shoulder Rating Questionnaire (SRQ) into the Dutch language.* Clin Rehabil, 2005. **19**(3): p. 300-11.
49. Lippitt, S.B., D.T. Harryman, and F.A. Matsen, *A practical tool for evaluation of function: the Simple Shoulder Test. In The Shoulder: a Balance of Mobility and Stability.* The American Academy of Orthopaedic Surgeons, 1993: p. pp. 501-518. .
50. van Kampen, D.A., et al., *Validation of the Dutch version of the Simple Shoulder Test.* J Shoulder Elbow Surg, 2012. **21**(6): p. 808-14.
51. Tashjian, R.Z., et al., *Minimal clinically important differences in ASES and simple shoulder test scores after nonoperative treatment of rotator cuff disease.* J Bone Joint Surg Am, 2010. **92**(2): p. 296-303.
52. Mintken, P.E., P. Glynn, and J.A. Cleland, *Psychometric properties of the shortened disabilities of the Arm, Shoulder, and Hand Questionnaire (QuickDASH) and Numeric Pain Rating Scale in patients with shoulder pain.* J Shoulder Elbow Surg, 2009. **18**(6): p. 920-6.
53. Godfrey, J., et al., *Reliability, validity, and responsiveness of the simple shoulder test: psychometric properties by age and injury type.* J Shoulder Elbow Surg, 2007. **16**(3): p. 260-7.
54. Croft, P., et al., *Measurement of shoulder related disability: Results of a validation study.* ANN RHEUM DIS, 1994. **53**(8): p. 525-8.
55. Terwee, C.B., et al., *Methodological quality of studies on the measurement properties of neck pain and disability questionnaires: a systematic review.* J Manipulative Physiol Ther, 2011. **34**(4): p. 261-72.
56. Wild, D., et al., *Multinational trials-recommendations on the translations required, approaches to using the same language in different countries, and the approaches to support pooling the data: the ISPOR Patient-Reported Outcomes Translation and Linguistic Validation Good Research Practices Task Force report.* Value Health, 2009. **12**(4): p. 430-40.
57. Reeve, B.B., et al., *ISOQOL recommends minimum standards for patient-reported outcome measures used in patient-centered outcomes and comparative effectiveness research.* Qual Life Res, 2013. **22**(8): p. 1889-905.
58. Beaton, D.E., J.G. Wright, and J.N. Katz, *Development of the QuickDASH: comparison of three item-reduction approaches.* J Bone Joint Surg Am, 2005. **87**(5): p. 1038-46.
59. Beaton, D.E., J.G. Wright, and J.N. Katz, *Development of the QuickDASH: comparison of three item-reduction approaches.* J Bone Joint Surg (Am), 2005. **87A**(5): p. 1038-1046.
60. Turner, R.R., et al., *Patient-reported outcomes: instrument development and selection issues.* Value Health, 2007. **10 Suppl 2**: p. S86-93.

61. Rose, M., et al., *Evaluation of a preliminary physical function item bank supported the expected advantages of the Patient-Reported Outcomes Measurement Information System (PROMIS).* J Clin Epidemiol, 2008. **61**(1): p. 17-33.
62. Hays, R.D., et al., *Upper-extremity and mobility subdomains from the Patient-Reported Outcomes Measurement Information System (PROMIS) adult physical functioning item bank.* Arch Phys Med Rehabil, 2013. **94**(11): p. 2291-6.
63. Cloke, D.J., et al., *A comparison of functional, patient-based scores in subacromial impingement.* J Shoulder Elbow Surg, 2005. **14**(4): p. 380-4.
64. Fan, Z.J., C.K. Smith, and B.A. Silverstein, *Assessing validity of the QuickDASH and SF-12 as surveillance tools among workers with neck or upper extremity musculoskeletal disorders.* J Hand Ther, 2008. **21**(4): p. 354-365.
65. Offenbacher, M., et al., *Validation of a German version of the 'Disabilities of Arm, Shoulder and Hand' questionnaire (DASH-G).* Z Rheumatol, 2003. **62**(2): p. 168-77.
66. Alvarez-Nemegyei, J., et al., *Development of a Spanish-language version of the Shoulder Disability Questionnaire.* J Clin Rheumatol, 2005. **11**(4): p. 185-7.
67. Jeldi, A.J., et al., *Cross-cultural adaption, reliability and validity of an Indian (Tamil) version for the Shoulder Pain and Disability Index.* Hong Kong Physiother J, 2012. **30**(2): p. 99-104.

APPENDIX. SEARCH

("Shoulder Pain"/ OR ((pain* OR complaint* OR disorder* OR lesion* OR injur* OR stiff* OR tight* OR patholog* OR impingem* OR disease*) ADJ3 shoulder*).ab,ti.) OR ((shoulder/ OR shoulder joint/ OR (shoulder* OR (joint* ADJ3 (glenohumeral OR humeroscapular OR scapulohumeral OR "scapulo humeral"))).ab,ti.) AND (pain/ OR "Wounds and Injuries"/ OR "Arm Injuries"/ OR ((functional ADJ3 (disorder* OR illness* OR impairment* OR limitation* OR disabilit* OR status* OR complaint*)) OR ((activit* OR participat*) ADJ6 (limit* OR complicat* OR interfer*)) OR (Disabilit* ADJ3 Evaluat*)).ab,ti.)) AND (exp questionnaires/ OR (questionnaire* OR ((self OR patient*) ADJ3 report*) OR PRO OR PROM).ab,ti.) AND (instrumentation.xs. OR methods.xs. OR validation studies.pt. OR comparative study.pt. OR exp psychometrics/ OR exp "outcome assessment (health care)"/ OR observer variation/ OR exp Health Status Indicators/ OR Reproducibility of Results/ OR Discriminant Analysis/ OR (psychometr* OR clinimetr* OR clinometr* OR (outcome ADJ3 (measure* OR assess*)) OR (observer* ADJ3 variation*) OR reproducib* OR reliab* OR unreliab* OR valid* OR coefficient OR homogen* OR "internal consistency" OR (cronbach* ADJ3 (alpha OR alphas)) OR (item* ADJ3 (correlation* OR selection* OR reduction*)) OR agreement OR precision OR imprecision OR "precise values" OR (test ADJ3 retest) OR (reliab* ADJ3 (test OR retest)) OR stability OR interrater OR intrarater OR ((intra OR inter) ADJ (rater OR tester OR observer OR technician OR examiner OR assay OR individual OR participant)) OR intertester OR intratester OR interobserver OR intraobserver OR intertechnician OR intratechnician OR interexaminer OR intraexaminer OR interassay OR intraassay OR interindividual OR intraindividual OR interparticipant OR intraparticipant OR kappa OR "kappa s" OR kappas OR repeatab* OR ((replicab* OR repeated) ADJ6 (measure OR measures OR findings OR result OR results OR test OR tests)) OR generali* OR concordance OR (intraclass ADJ3 correlation*) OR discriminative OR "known group" OR (factor ADJ (structure* OR analy*)) OR dimension* OR subscale* OR (multitrait AND (scaling ADJ3 analy*)) OR "item discriminant" OR (interscale ADJ correlation*) OR error OR errors OR ((individual OR interval OR rate) ADJ variability) OR (variability ADJ3 (analy* OR values)) OR (uncertainty ADJ3 (measurement OR measuring)) OR "standard error of measurement" OR sensitiv* OR responsive* OR (limit ADJ3 detection) OR "minimal detectable concentration" OR interpretab* OR ((minimal OR minimally OR clinical OR clinically) ADJ3 (important OR significant OR detectable) AND (change OR difference)) OR (small* ADJ3 (real OR detectable) AND (change OR difference)) OR "meaningful change" OR "ceiling effect" OR "floor effect" OR "Item response model" OR IRT OR Rasch OR "Differential item functioning" OR DIF OR "computer adaptive testing" OR "item bank" OR "cross-cultural equivalence").ab,ti.)

This chapter is an extended version of:
Thoomes-de Graaf M, Scholten-Peeters GG, Duijn E, Karel Y, Koes BW, Verhagen AP.

Qual Life Res. 2015 Jun;24(6):1515-9. doi: 10.1007/s11136-014-0879-1. [Epub 2014 Dec 4].

THE DUTCH SHOULDER PAIN AND DISABILITY INDEX (SPADI): A RELIABILITY AND VALIDATION STUDY

CHAPTER 3

ABSTRACT

Purpose: To evaluate the reliability and validity of the Dutch Shoulder Pain and Disability Index (SPADI-D).

Background: The SPADI is recommended and frequently used. However, the validity and reliability of the SPADI-D are unknown.

Methods: The study population consisted of patients consulting a physiotherapist for shoulder pain. We assessed construct validity, using known groups, convergent validity (SDQ) and divergent validity (EQ5D) for which the mean difference or Spearman correlations coefficients were calculated. The factor structure was assessed using principal component factor analysis and we calculated Cronbach's alpha and the ICC to assess the reliability.

Results: A total of 356 patients and a randomly selected group of 74 subjects for the reliability analysis were included. There was a significant difference between extreme groups (a high/low level of pain and work absence/presence) in SPADI score. The correlation between the SPADI and the SDQ was 0.69, with the EQ5D mobility-item 0.25 and with the depression-item 0.14. The SPADI consisted of one factor according to principal component factor analysis, showed high internal consistency (Cronbach alpha=0.94 for the total score), and the test-retest reliability was good (ICC=0.89).

Conclusion: The Dutch SPADI is a valid and reliable questionnaire for patients in primary care in assessing functional disability.

Keywords: SPADI, validation, reliability, shoulder

INTRODUCTION

Shoulder pain is a common disorder in western society [1]. The point prevalence ranges from 7 to 27% [2], which makes it the second most reported musculoskeletal complaint in general practice [3]. The 12-month prevalence ranges from 5 to 47% and the lifetime prevalence from 7 to 67%, depending on case definitions and age [2]. One of the main complaints in patients with shoulder pain is functional disability [1]. The reported functional disabilities in these patients range from difficulties with opening a jar and getting dressed, to impeding sleep. Functional disabilities can reach a level of severity where they preclude work-related tasks [1]. Shoulder pain presents an economic burden on society due to costs of sick leave and health care and also impacts patient's quality of life [4]. Treatment of shoulder pain is usually aimed at pain reduction and improvement of functional disabilities [5]. Consequently, outcome measurements should include an instrument (e.g. questionnaire) for the evaluation of functional disabilities [6].

There are several self-administered shoulder pain and disability questionnaires, e.g.; Shoulder Disability Questionnaire (SDQ) [7] and the Shoulder Pain and Disability Index (SPADI) [8]. Most of these questionnaires were developed and tested in secondary care settings [9]. Patients ranked the SDQ and the SPADI as the most relevant questionnaire for their shoulder problem [9]. The SPADI was rated as the least time-consuming shoulder questionnaire, both the SDQ and the SPADI appear to be convenient and easy to complete [9].

The SDQ was originally designed and validated in Dutch, it showed acceptable convergent and divergent validity and is a responsive instrument [5, 7, 10]. The reliability of the SDQ however is still unknown. The SPADI is the most frequently used questionnaire and was originally developed in English [8]. It has been translated and validated into Danish, Norwegian, Tamil, German, Turkish and Slovene [11-16]. The SPADI showed excellent reliability and responsiveness and has been validated in a range of clinical settings [17-19].

Several reviews have encouraged the use of the SPADI in clinical and research settings [17-19]. The Royal Dutch Society for Physical Therapy has recommended implementation of the Dutch SPADI (SPADI-D) in a clinical guideline for patients with shoulder pain [20]. Despite, it's frequent use and recommendations in reviews and in a clinical guideline, the SPADI has not been validated and tested for reliability in Dutch.

Therefore, the aim of this study is to evaluate the reliability and construct validity of the SPADI-D for patients with shoulder pain in primary care.

METHODS

This is a validation and reliability study, which is part of a larger cohort study (ShoCoDiP-study), including patients with shoulder pain in physiotherapy setting. Details of the design are presented elsewhere [21]. Construct validity consists of discriminative validation (using extreme groups), convergent validity (a high correlation with SDQ) and divergent validity (a low correlation with items of the EuroQol five-item quality of life questionnaire (EQ-5D-3L)). Reliability consists of internal consistency and test-retest measurement. The Medical Ethics Committee of the Erasmus Medical Center in Rotterdam approved the study (MEC-2011-414). Informed consent was obtained from all patients.

Study population

Patients were recruited from primary care physical therapy clinics between November 2011 and December 2012. Patients with shoulder pain were eligible for inclusion if they were at least 18 years old and adequately understood the Dutch language. Patients with serious pathology (infection, cancer or fracture), previous surgery or diagnostic imaging techniques of the shoulder, such as Magnetic Resonance Imaging or Ultrasound in the previous 3 months, were excluded.

Baseline measurement

Patients received an online questionnaire that included the SPADI-D, SDQ and the Euro-Qol five-item quality of life questionnaire (EQ-5D-3L).

The SPADI is designed to measure pain and disability associated with shoulder pain. It consists of 13 items and response options range from 0 to 10, where 0 represents "no pain/no difficulty" and 10 "worst pain imaginable/very difficult". The total score varies between 0 and 100; a higher score indicates a higher level of pain related disability [8].

The SDQ is a pain-related disability questionnaire consisting of 16 items. Response options are "yes", "no" or "not applicable". The SDQ-score can range from 0 to 100 with a higher score indicating more severe disability [5]. The SDQ was originally designed and validated in Dutch, internal consistency and responsiveness are good [5, 10].

The Dutch EQ-5D-3L is a quality of life questionnaire covering 5 dimensions of health: mobility, self-care, usual activities, pain/discomfort and anxiety/depression and an official language version [22]. Response options are "no problems", "some problems", "extreme problems". The EQ-5D-3L has been used frequently, most often as part of a cost-effectiveness study [23-25]. The Dutch EQ-5D-3L is an official language version and has been validated [22].

Test-retest measurement

A randomly selected group of patients received a second SPADI-D after one week. On both occasions, test conditions were considered equal (online). The time interval was chosen to minimize recall bias as well as progression bias and is often considered appropriate [26]. A sample size of approximately 80 is considered acceptable [27].

Analysis

Analyses were performed with SPPS22. Handling of missing items was performed as described by the original authors of the SPADI and SDQ [8, 10].

All data was checked on normality, using a Stem-and-leaf Plot, Q-Plot and whisker box. Nonparametric tests were used if data was not normally distributed.

Known groups validity. We assumed that patients with high initial pain (>7 on the Numeric Rating Scale in the preceding 24 hours) and work absence would have a higher level of perceived disability. Both groups have been chosen a priory [10, 15]. The independent T-test was used to test the difference between known groups.

Convergent validity relates to the extent to which a particular instrument corresponds to the construct (theoretical concept) of shoulder pain and function [27]. Therefore, the correlation between the total score of the SPADI-D and the total score of the SDQ was evaluated as both questionnaires aim to measure the same construct [5, 7, 9, 10]. Convergent validity was quantified by the Pearson correlation coefficient in case of a normal distribution of the data, otherwise a Spearman correlation coefficient was used. Correlations were rated as follows: $r < 0.30$ as low/insignificant; $0.30 \leq r < 0.45$ as moderate; $0.45 \leq r < 0.60$ as substantial and $r \geq 0.60$ as high [28]. High correlations ($r \geq 0.60$) were expected [27] between the scores on SPADI-D and the SDQ, as both aim to measure the same construct.

Divergent validity. Based on the concept of divergent validity, where the correlation is expected to be low ($r < 0.30$), between instruments based on different constructs, the correlation between some items of the EQ-5D-3L and the SPADI-D were analysed [27]. For this, we used items of the EQ-5D-3L that were considered as items which would not likely to be affected as a consequence of shoulder pain, we assumed the items 'mobility' (Mobility: I have no problems in walking about- I am confined to bed) and 'anxiety/depression' (Anxiety/Depression: I am not anxious or depressed-I am extremely anxious or depressed) were not directly associated with shoulder pain. In the literature, no significant differences were found between patients with shoulder pain and healthy subjects in the amount of time spent walking [29]. Previous research showed that depression and anxiety only showed low correlations, or just above (0.31, where 0.30 is considered as a low correlation) [30-33] with activity limitations (using several questionnaires focused on activity limitations of the shoulder and depression scales).

Factor structure. The SPADI-D was developed as a two-factor scale, measuring pain and disability [8], which was confirmed with factor analysis [34]. Contrary, other studies found only one factor, or a different factor loading on both factors of the SPADI than originally described [13, 35-37]. Patients were included in the analysis if they answered all items. To evaluate the factor structure in our data we conducted a principal component factor analysis with and without varimax rotation. We checked whether the data was suitable for factor analysis using Kaiser-Meyer-Olkin Measure (should be 0.60 or greater) and Barlett's Test of Sphericity (should be significant). We used the eigenvalue (>1) criteria, checked the elbow of the scree plot and used parallel analysis [38-40] to extract the number of factors. In case, another number of items then two was extracted, we also used a two-fixed factor analysis, to assess whether the items load as originally described. Items loading higher than 0.40 on one factor and lower than 0.30 on any other factor were acceptable [41]. Ultimately, the stability of our model was assessed using two random splitting halves (subsamples) [42], we performed this five times to assess if our findings were consistent.

Internal consistency. Internal consistency was calculated using Cronbach's alpha and only for the scale(s) that were extracted from our factor analysis. A Cronbach's alpha between 0.70 and 0.95 is considered "good" [43].

Test-retest. The intraclass correlation coefficient (ICC) using a two-way mixed model was used to calculate the test–retest reliability. The ICC can range from 0.00 (no stability/ agreement) to 1.00 (perfect stability/ agreement). An ICC of 0.70 is considered to be acceptable [43]. We checked the test-retest data for extreme values and assessed whether this influenced the ICC.

RESULTS

Patient characteristics

Due to missing variables out of 389 patients, 356 patients were included in this analysis and 74 in the test-retest reliability analysis. The mean age was 49.5 (SD 13) years and 47% was male. Demographic characteristics are reported in Table 1.

TABLE 1. Demographic characteristics of the included patients.

Demographic characteristics of patients		
	Total (356)	Test-retest (74)
Gender (male) (%)	166 (47%)	29 (39%)
Age (SD)	49.5 (13.1)	51.4 (12.7)
Duration of shoulder pain in weeks Median (IQR)	12 (6-26)	16 (8-40)
SPADI baseline score Mean (SD)	46.7 (21.3)	50.8 (22.6)
Use of medication (%)	171 (49%)	37 (51%)
Initial pain (NRS) Median (IQR)	6 (4-7)	6 (4-8)
Work absence (%)	38 (out of the 255 working patients) (15%)*	5 (out of the 48 working patients) (11%)**

Abbreviations: SD; Standard deviation, NRS; Numeric Rating Scale, IQR; Inter quartile range.
* based on data of 252 working patients ** based on 47 working patients

A total of 298 patients (83%) answered all questions of the SPADI-D. The average percentage of missing items per question was 1,7%. Question 1 (20; 5.6%) and question 7 (14; 3.9%) showed the highest amount of missing scores and question; question 10 and 13 both had none missing scores. Therefore, there is no indication of inappropriate or hard to answer questions.

The data of the SPADI-D at baseline and at re-test were considered as normally distributed, in contrast to the data of the SDQ and EQ-5D-3L.

Validity

Differences between "known groups" were statistically significant and considered clinically relevant (Table 2). This means that the SPADI-D is able to differentiate between different groups.

The Spearman correlation between the SPADI-D and SDQ was high (r = 0.69), meaning the convergent validity of SPADI-D with SDQ is good. The Spearman correlation between the SPADI-D and EQ-5D-3L mobility-item (r = 0.25) and the EQ-5D-3L_depression-item (r = 0.14) was low. This means the SPADI-D and EQ-5D-3L measure a different construct.

TABLE 2. Extreme groups correlation coefficients.

	Group	SPADI-D (mean, SD)	Mean difference
Pain- SPADI-D	High initial pain >7	59.4 (18.0)	-21.4* p=0.00 (95%CI: -25.4 to -17.5)
	Low initial pain <7	37.9 (18.9)	
Work absence- SPADI-D	Work absence +	50.5 (20.6)	-7.6* p=0.04 (95%CI: -14.8- to 0.3)
	Work absence -	43.0 (21.1)	

Abbreviations: SD; standard deviation, CI; confidence Interval, *; significant

Factor structure

The results are based on 298 patients. Parallel analysis revealed that the eigenvalue of the first factor should be above 1.44 and of the second factor above 1.33 to be extracted. Only one factor was extracted (see Figure 1), the eigenvalue of the second factor was 0.97. A one-factor solution explained 57.9%. Factor loadings on this one factor ranged between 0.62 and 0.84; the items "How much pain at its worst?" and "How much pain when lying on the involved side?" showed the lowest loading (0.62 and 0.65). When looking at our two-factor solution (using fixed factors), the explained variance was 65.4% and the items 4 (P1), 11 (D6) and 12 (D7) did not load on our two factors as the original factor loading did (see Figure 2 and Table 3).

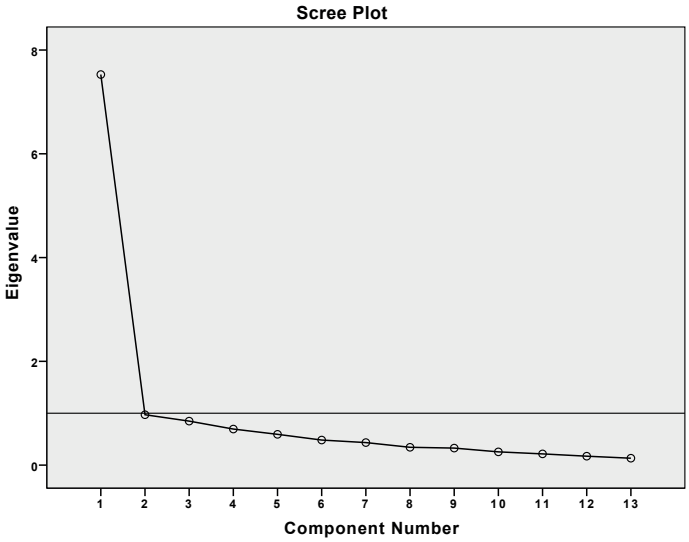


FIGURE 1. Scree plot
A scree plot of eigenvalues, the demarcation point indicates one factor. The results are based on 298 patients.

Findings were consistent with all five analyses based on two random subsamples. This means we consider the SPADI-D to have one factor.

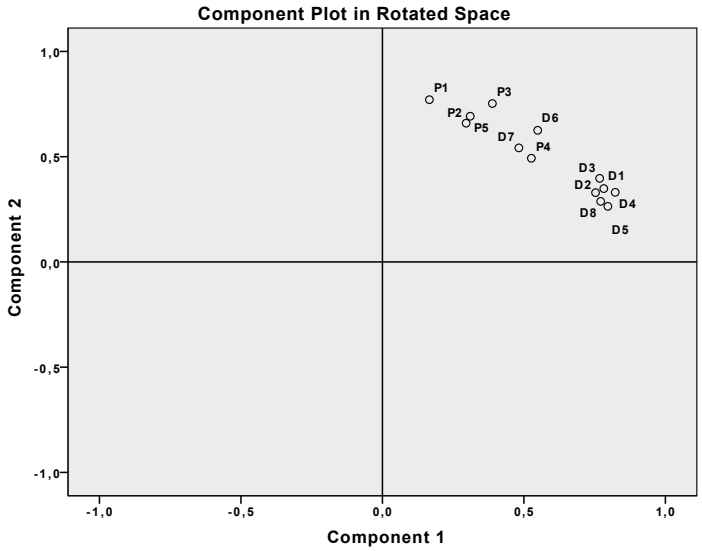


FIGURE 2. Plot of varimax rotated factor loadings on each SPADI-D item using two fixed factors
This figure shows the original items of pain and disability do not load solely on those factors and there is actually only one factor.

TABLE 3. Factor loading using two fixed factors and varimax rotation

Items	Component 1	Component 2
Original Pain scale		
1 (P1): At it's worst	0.166	**0.771**
2 (P2): When lying on the involved side	0.296	**0.659**
3 (P3): Reaching for something on a high shelf	0.389	**0.753**
4 (P4): Touching the back of your neck	**0.526**	0.492
5 (P5): Pushing with the involved arm	0.310	**0.692**
Original Disability Scale		
6 (D1): Washing your hair	**0.782**	0.348
7 (D2): Washing your back	**0.753**	0.329
8 (D3): Putting on an undershirt or pullover sweater	**0.768**	0.397
9 (D4): Putting on a shirt that buttons down the front	**0.823**	0.331
10 (D5): Putting on your pants	**0.797**	0.264
11 (D6): Placing an object on a high shelf	0.549	**0.625**
12 (D7): Carrying a heavy object of 10 pounds	0.482	**0.542**
13 (D8): Removing something from your back pocket	**0.771**	0.288

Bold represents the highest amount of loading on that specific factor of the SPADI-D item.

Reliability

The internal consistency and test-retest reliability were good (Cronbach's alpha = 0.94; ICC = 0.89 (95%CI 0.83-0.93)). After exclusion of two patients with extreme values, the ICC was 0.90 (95%CI 0.85-0.94). Both indicate a high level of agreement.

DISCUSSION

This study shows that the SPADI-D consists of one factor only and can be considered as a valid and reliable questionnaire. It discriminates well between known groups, correlates well with the SDQ and internal consistency and test-retest reliability are high.

One SPADI validation study used similar "known groups", showing a higher mean difference for work absence compared to ours [15]. Differences with this study were that their population was smaller, had a higher baseline SPADI-score and they did not present the percentage of people that could not work due to their shoulder pain.

Correlation coefficients found in other studies for convergent validity, varied between moderate and high (0.33 to 0.85) depending on the comparator [9, 34, 44, 45]. Few studies evaluated divergent validity of the SPADI and none used the EQ-5D-3L [34, 45].

Only one study reported a factor structure as originally described by Roach [34], the majority of studies could not confirm this loading pattern or reported a one-factor structure [13, 35-37]. One study concluded the SPADI consisted of two factors, but they did not report the explained variance or the eigenvalue for the second factor. They found the exact same items (4, 11 and 12) that did not follow the originally proposed loading pattern, as we found when using our two fixed factors [36].

Another study concluded that people do not distinguish between pain and disability and a possible explanation for this finding could be the wording of the SPADI items. The disability items ask respondents to indicate the amount of difficulty they have with specified functions. It is possible that, when people report their difficulty in performing an activity, they consider pain to be part of what makes the activity difficult [37].

The Cronbach's alpha found in other studies ranged between 0.90 and 0.95 [8, 12, 34-36] and ICC values ranged between 0.88 and 0.94 [12-15, 46], both consistent with ours.

Our study has some limitations. First, the translation process of the SPADI-D was not published and it is unknown if it is performed as recommended [47]. Nevertheless, the SPADI-D is commonly used in clinical practice and research and is also integrated in multiple patient-management software programs in The Netherlands. Secondly, we did not use the General Perceived Effect scale to check if patients were indeed stable between the test and the re-test. However, it is unlikely that patients would have been recovered within one week, due to the duration of complaints and the mean number

of weeks patients usually need to recover [48]. The extreme value analysis showed that differences after exclusion were minimal.

On the other hand, we used an adequate sample size to perform factor analysis [42]. There is increasing consensus among statisticians that parallel analysis is superior to other procedures and typically yields optimal solutions to the number of components problem [38].

REFERENCES

1. Feleus, A., et al., *Management in non-traumatic arm, neck and shoulder complaints: differences between diagnostic groups.* Eur Spine J, 2008. **17**(9): p. 1218-29.
2. Luime, J.J., et al., *Prevalence and incidence of shoulder pain in the general population; a systematic review.* Scand J Rheumatol, 2004. **33**(2): p. 73-81.
3. Picavet, H.S. and J.S. Schouten, *Musculoskeletal pain in the Netherlands: prevalences, consequences and risk groups, the DMC(3)-study.* Pain, 2003. **102**(1-2): p. 167-78.
4. Huisstede, B.M., et al., *Incidence and prevalence of upper-extremity musculoskeletal disorders. A systematic appraisal of the literature.* BMC Musculoskelet Disord, 2006. **7**: p. 7.
5. van der Windt, D.A., et al., *The responsiveness of the Shoulder Disability Questionnaire.* Ann Rheum Dis, 1998. **57**(2): p. 82-7.
6. Mintken, P.E., P. Glynn, and J.A. Cleland, *Psychometric properties of the shortened disabilities of the Arm, Shoulder, and Hand Questionnaire (QuickDASH) and Numeric Pain Rating Scale in patients with shoulder pain.* J Shoulder Elbow Surg, 2009. **18**(6): p. 920-6.
7. van der Heijden, G.J., P. Leffers, and L.M. Bouter, *Shoulder disability questionnaire design and responsiveness of a functional status measure.* J Clin Epidemiol, 2000. **53**(1): p. 29-38.
8. Roach, K.E., et al., *Development of a shoulder pain and disability index.* Arthritis Care Res, 1991. **4**(4): p. 143-9.
9. Paul, A., et al., *A comparison of four shoulder-specific questionnaires in primary care.* Ann Rheum Dis, 2004. **63**(10): p. 1293-9.
10. de Winter, A.F., et al., *The Shoulder Disability Questionnaire differentiated well between high and low disability levels in patients in primary care, in a cross-sectional study.* J Clin Epidemiol, 2007. **60**(11): p. 1156-63.
11. Angst, F., et al., *Cross-cultural adaptation, reliability and validity of the German Shoulder Pain and Disability Index (SPADI).* Rheumatology (Oxford), 2007. **46**(1): p. 87-92.
12. Bicer, A. and H. Ankarali, *Shoulder Pain and Disability Index: A validation study in Turkish women.* Singapore Med J, 2010. **51**(11): p. 865-70.
13. Jamnik, H. and M.K. Spevak, *Shoulder pain and disability Index: Validation of slovene version.* Int J Rehabil Res, 2008. **31**(4): p. 337-41.
14. Jeldi, A.J., et al., *Cross-cultural adaptation, reliability and validity of the Indian (Tamil) version of the Shoulder Pain and Disability Index.* Hong Kong Physiotherapy Journal, 2012.
15. Christiansen, D.H., J.H. Andersen, and J.P. Haahr, *Cross-cultural adaption and measurement properties of the Danish version of the Shoulder Pain and Disability Index.* Clin Rehabil, 2013. **27**(4): p. 355-60.
16. Ekeberg, O.M., et al., *Agreement, reliability and validity in 3 shoulder questionnaires in patients with rotator cuff disease.* BMC Musculoskelet Disord, 2008. **9**: p. 68.
17. Bot, S.D., et al., *Clinimetric evaluation of shoulder disability questionnaires: a systematic review of the literature.* Ann Rheum Dis, 2004. **63**(4): p. 335-41.
18. Roy, J.S., J.C. MacDermid, and L.J. Woodhouse, *Measuring shoulder function: a systematic review of four questionnaires.* Arthritis Rheum, 2009. **61**(5): p. 623-32.
19. Breckenridge, J.D. and J.H. McAuley, *Shoulder Pain and Disability Index (SPADI).* J Physiother, 2011. **57**(3): p. 197.
20. Jansen, M.J., et al., *KNGF Evidence Statement Subacromiale klachten.* Nederlands Tijdschrift voor Fysiotherapie, 2011. **121**(1).

21. Karel, Y.H., et al., *Current management and prognostic factors in physiotherapy practice for patients with shoulder pain: design of a prospective cohort study.* BMC Musculoskelet Disord, 2013. **14**(1): p. 62.
22. Lamers, L.M., et al., *The Dutch tariff: results and arguments for an effective design for national EQ-5D valuation studies.* Health Econ, 2006. **15**(10): p. 1121-32.
23. Rabin R, O.M., Oppe M, Janssen B, Herdman M. , *EQ-5D User Guide; Basic information on how to use the EQ-5D-5L instrument.* The EuroQol Group, 2011.
24. Szende, A., M. Oppe, and N. Devlin, *EQ5D value sets- inventory, comparative review and user guide.* 2007, Dordrecht: Springer.
25. Johnson, J.A. and A.S. Pickard, *Comparison of the EQ-5D and SF-12 health surveys in a general population survey in Alberta, Canada.* Med Care, 2000. **38**(1): p. 115-21.
26. Streiner, D.L. and G.R. Norman, *Health measurement scales a practical guide to the development and use.* . 2008: Oxford university press.
27. Mokkink, L.B., et al., *The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content.* BMC Med Res Methodol, 2010. **10**: p. 22.
28. Burnand, B., W.N. Kernan, and A.R. Feinstein, *Indexes and boundaries for "quantitative significance" in statistical decisions.* J Clin Epidemiol, 1990. **43**(12): p. 1273-84.
29. Hallman, D.M., A.H. Ekman, and E. Lyskov, *Changes in physical activity and heart rate variability in chronic neck-shoulder pain: monitoring during work and leisure time.* Int Arch Occup Environ Health, 2013.
30. Roh, Y.H., et al., *To what degree do shoulder outcome instruments reflect patients' psychologic distress?* Clin Orthop Relat Res, 2012. **470**(12): p. 3470-7.
31. Menendez, M.E., et al., *Computerized adaptive testing of psychological factors: relation to upper-extremity disability.* J Bone Joint Surg Am, 2013. **95**(20): p. e149.
32. Cho, C.H., et al., *Is shoulder pain for three months or longer correlated with depression, anxiety, and sleep disturbance?* J Shoulder Elbow Surg, 2013. **22**(2): p. 222-8.
33. Niekel, M.C., et al., *Correlation of DASH and QuickDASH with measures of psychological distress.* J Hand Surg Am, 2009. **34**(8): p. 1499-505.
34. Hill, C.L., et al., *Factor structure and validity of the shoulder pain and disability index in a population-based study of people with shoulder symptoms.* BMC Musculoskelet Disord, 2011. **12**: p. 8.
35. Tveita, E.K., et al., *Factor structure of the Shoulder Pain and Disability Index in patients with adhesive capsulitis.* BMC Musculoskelet Disord, 2008. **9**.
36. MacDermid, J.C., P. Solomon, and K. Prkachin, *The Shoulder Pain and Disability Index demonstrates factor, construct and longitudinal validity.* BMC Musculoskelet Disord, 2006. **7**: p. 12.
37. Roddey, T.S., et al., *Comparison of the University of California - Los Angeles Shoulder Scale and the Simple Shoulder Test with the Shoulder Pain and Disability Index: Single-administration reliability and validity.* Phys Ther, 2000. **80**(8): p. 759-68.
38. O'Connor, B.P., *SPSS and SAS programs for determining the number of components using parallel analysis and velicer's MAP test.* Behav Res Methods Instrum Comput, 2000. **32**(3): p. 396-402.
39. Horn, J.L., *A Rationale and Test for the Number of Factors in Factor Analysis.* Psychometrika, 1965. **30**: p. 179-85.
40. Patil, V.H., et al., *"Parallel Analysis Engine to Aid Determining Number of Factors to Retain [Computer software]. Available from http://smishra.faculty.ku.edu/parallelengine.htm; Utility developed as part of Patil, Vivek H., Surendra N. Singh, Sanjay Mishra, and Todd Donovan (2008), "Efficient Theory Development and Factor Retention Criteria: A Case for Abandoning the 'Eigenvalue Greater Than One' Criterion," Journal of Business Research, 61 (2), 162-170.,.* 2008.

41. Floyd, F.J. and K.F. Widaman, *Factor analysis in the develop- ment and refinement of clinical assessment instruments. Psychol Assess.* Psychol Assess, 1995. **7: 286–299**.
42. de Vet, H.C., et al., *Are factor analytical techniques used appropriately in the validation of health status questionnaires? A systematic review on the quality of factor analysis of the SF-36.* Qual Life Res, 2005. **14**(5): p. 1203-18; dicussion 1219-21, 1223-4.
43. Terwee, C.B., et al., *Quality criteria were proposed for measurement properties of health status questionnaires.* J Clin Epidemiol, 2007. **60**(1): p. 34-42.
44. Staples, M.P., et al., *Shoulder-specific disability measures showed acceptable construct validity and responsiveness.* J Clin Epidemiol, 2010. **63**(2): p. 163-70.
45. Cloke, D.J., et al., *A comparison of functional, patient-based scores in subacromial impingement.* J Shoulder Elbow Surg, 2005. **14**(4): p. 380-4.
46. Tveita, E.K., et al., *Responsiveness of the shoulder pain and disability index in patients with adhesive capsulitis.* BMC Musculoskelet Disord, 2008. **9**: p. 161.
47. Wild, D., et al., *Principles of Good Practice for the Translation and Cultural Adaptation Process for Patient-Reported Outcomes (PRO) Measures: report of the ISPOR Task Force for Translation and Cultural Adaptation.* Value Health, 2005. **8**(2): p. 94-104.
48. Kooijman, M., et al., *Jaarcijfers 2010 en trendcijfers 2006-2010 fysiotherapie.* Landelijke Informatievoorziening Paramedische Zorg. Utrecht: NIVEL, http://www.nivel.nl/lipz.

Marloes Thoomes-de Graaf, Wendy GM Scholten-Peeters, Edwin Duijn, Yasmaine HJM Karel, Maaike PJ van den Borne, Annechien Beumer, Ramon PG Ottenheijm, Geert Jan Dinant, Eric Tetteroo, Cees Lucas, Bart W Koes, Arianne P Verhagen.

Man Ther. 2014 Oct;19(5):478-83. doi: 10.1016/j.math.2014.04.018. [Epub 2014 May 14].

INTER-PROFESSIONAL AGREEMENT OF ULTRASOUND-BASED DIAGNOSES IN PATIENTS WITH SHOULDER PAIN BETWEEN PHYSICAL THERAPISTS AND RADIOLOGISTS IN THE NETHERLANDS

CHAPTER 4

ABSTRACT

Study design: Reliability study.

Objectives: The aim of this study was to evaluate the interrater-reliability of the interpretation of diagnostic ultrasound in patients with shoulder pain between physiotherapists and radiologists.

Background: Although physiotherapists in The Netherlands increasingly use diagnostic ultrasound in clinical practice, there is no evidence available on its reliability.

Methods: A cohort study included patients with shoulder pain from primary care physiotherapy. Patients followed the usual diagnostic pathway of which diagnostic ultrasound could be a part. Patients that received diagnostic ultrasound also visited a radiologist within one week for a second one. Patients and radiologists were blinded for the diagnostic ultrasound diagnosis of the physiotherapists. Agreement was assessed using Cohen's kappa statistics. Subgroup analysis was performed on education and experience.

Results: A total of 65 patients were enrolled and 13 physiotherapists and 9 radiologists performed diagnostic ultrasound. We found substantial agreement (0.63K) between physiotherapists and radiologists on the assessment of full thickness tears. The overall kappa of all four diagnostic categories was 0.36, indicating fair agreement. The more experienced and highly trained physiotherapists showed moderate agreement (0.43K) compared to only slight agreement (0.17 and 0.09K) from the less experienced and trained physiotherapists with radiologists.

Conclusion: The reliability between physiotherapists and radiologist on diagnostic ultrasound of shoulder patients in primary care is borderline substantial (Kappa = 0.63) for full thickness tears only. This level of reliability is relatively low when compared with the high reliability between radiologists. More experience and training of physiotherapists may increase the reliability of diagnostic ultrasound.

Key words: agreement, kappa, physiotherapy, radiology, ultrasonography

INTRODUCTION

Shoulder pain is a common disorder; the second most reported musculoskeletal complaint in general practice and presents an economic and social burden on society [1-3]. When a patient consults a general practitioner, about 15-17% of cases leads to a referral, of which 64% are referred to physiotherapy and 27% to secondary care [4, 5]. Since 2006 the patient has the possibility for direct access in the Netherlands, and about 28% of the patients consult the physiotherapist (PT) via direct access [6].

Physiotherapy assessment, which includes history taking and physical examination, provides a diagnosis. Physical examination, including specific and widely used tests, alone is not valid to differentiate between various disorders, because of low sensitivity, specificity and reproducibility [7-9]. An accurate diagnosis however, is regarded essential to ensure that patients receive appropriate treatment and correct information regarding their prognosis [10]. Adequate diagnosis of a full thickness tear, for example, is important, as this might warrant medical intervention [11, 12].

As indicated in previous studies, medical specialists (most often radiologists) were able to accurately diagnose several shoulder disorders using diagnostic musculoskeletal ultrasound [10, 13, 14]. The diagnostic accuracy in detecting full thickness rotator cuff tears, showed a pooled sensitivity of at least 0.92 and specificity higher than 0.94 for medical specialists [10, 13, 14]. The overall kappa, concerning the reliability between radiologists ranged from 0.90 to 0.95 [15, 16]. The learning curve for a non-musculoskeletal radiologist appears to be relatively short: after having performed 100 diagnostic ultrasounds of the shoulder, general radiologists showed excellent agreement [15, 17, 18]. Magnetic Resonance Imaging (MRI) and diagnostic ultrasound appeared to be equally accurate in detecting full thickness tears [19, 20]. Diagnostic ultrasound however, was the more cost- effective test procedure [13, 19].

In the detection of partial thickness rotator cuff tears, specificity remained high, but sensitivity decreased (ranging from 0.67 to 0.84) [10, 13, 14]. Also, the reliability between radiologists decreased as the overall kappa ranged between 0.63 and 0.79 [15, 16].

Only a small number of studies evaluated subacromial bursitis, calcifying tendonitis and tendinopathy; these results should therefore be interpreted with caution. The diagnostic accuracy for calcifying tendonitis and subacromial bursitis appeared to be high (sensitivity varying from 0.79 – 1 and specificity ranging from 0.83-1), while the sensitivity for a tendinopathy ranged from 0.67 – 0.93 [10].

Originally, PTs predominantly used ultrasound as a rehabilitative tool, to guide rehabilitation decisions [21], but recently diagnostic imaging is increasingly used by PTs [22, 23]. These two uses of ultrasound by PTs are different, requiring different training and skills. Rehabilitative ultrasound imaging (RUSI) is not the topic of the present paper. In clinical practice, GPs and PTs combine clinical history and physical examination with

diagnostic ultrasound aiming at diagnostic certainty to guide treatment decisions [20]. However, little is known about the reliability and validity of diagnostic ultrasound in primary care settings. The reliability of diagnostic ultrasound performed by PTs has only been studied on healthy subjects [24, 25].

It is estimated that a large amount (50-85%) of patients referred to secondary care (orthopaedic surgeons and radiologists) have already received a diagnostic ultrasound in primary care [26]. Despite its increased use, orthopaedic surgeons and radiologist feel there are more disadvantages than advantages in the use of diagnostic ultrasound performed by PTs. They assumed there was a high degree of false positives and negatives [26]. Only 2.9% of orthopaedic surgeons thought that PTs were the appropriate persons to perform a diagnostic ultrasound. Only 13.3% of orthopaedic surgeons trusted the results of a specific PT and in the majority of cases diagnostic ultrasound was repeated in secondary care [26].

Therefore, the aim of this study is to assess the interrater-reliability of diagnostic ultrasound between PTs and radiologists in patients with shoulder pain for full thickness tears. In case of substantial or high agreement, diagnostic ultrasound by PTs might be appropriate. The secondary aim of this study is to assess the interrater-reliability of a partial thickness tear, calcification and subacromial bursitis as well, as the diagnostic accuracy in these categories of diagnostic ultrasound by radiologists seems to be high. We are also interested if experience or training of the PT influences the overall reliability.

METHODS

Design

This is an interrater-reliability study which is part of a prognostic cohort study: 'Shoulder Complaints and Diagnostic Ultrasound in Physiotherapy' (ShoCoDiP) [27]. The Medical Ethical Committee of the Erasmus University approved this study, nr mec-2011-414. All patients provided informed consent prior to participating in the study.

Raters

Dutch PTs trained in the use of diagnostic ultrasound, were asked to participate. Several learning institutes in the Netherlands provide private post-graduate courses in diagnostic ultrasound for PTs. The use of diagnostic ultrasound is not limited by legal boundaries or to specific health care providers. All participating PTs should at least have one year of experience in diagnostic ultrasound and have made more than 100 diagnostic ultrasounds of the shoulder after graduation from their diagnostic ultrasound-education [15]. Experienced radiologists specialized in musculoskeletal complaints and who per-

formed diagnostic ultrasound of the shoulder on a regular basis where recruited from local hospitals.

PTs and radiologists were trained in a consensus meeting to use an international scanning protocol, which is focused on anatomy, technique and scanning pitfalls [28] to standardize the procedure. Data were collected concerning years of experience, education and type of ultrasound equipment used by radiologists and PTs.

Ultrasound equipment

To be eligible to participate, PTs should work with equipment with a minimum transducer frequency range of 5-10MHz, as a transducer frequency of 7.5 MHz seems to be recommended for detecting rotator cuff tears [14] and a minimal feature of digital beamformer technology. No specific demands concerning the equipment of the radiologists were made.

Patients

Patients were recruited from primary care physiotherapy clinics between November 2011 and December 2012. Patients with shoulder pain were eligible if they were over 18 years of age and adequately understood the Dutch language. Patients were excluded in the presence of serious pathology (such as infection, cancer or fracture), previous surgery and/or diagnostic imaging techniques of the shoulder such as MRI and in the previous three months.

Baseline measurement

Patients received usual physiotherapy assessment, including normal history taking and physical examination. Additionally, a diagnostic ultrasound made by the PT could be part of the diagnostic pathway. These patients were included in the reliability study and received a second diagnostic ultrasound within one week, performed by a radiologist, at the nearest participating hospital. When this was not possible, patients were excluded from the study because of possible progression bias and reasons were recorded. The standard procedure for a radiologist is to perform a hypothesis-generated diagnostic ultrasound; the hypothesis being supplied by either a GP or orthopaedic specialist. Therefore, the radiologist received the PT's hypothesis based on history taking alone; as the diagnostic ultrasound made by the PT was performed after history taking.

Outcome measurement

For the purpose of this manuscript, we were only interested in full thickness tears, partial thickness tears, bursitis and calcifications. However, based on a consensus meeting between PTs and radiologists, other diagnoses were also integrated into a standardized form, to reflect the diagnostic labels most commonly used in clinical practice.

Both PTs and radiologists used the same standardized form to record their diagnostic ultrasound diagnoses [27]. All forms were available online, using Limesurvey software.

Diagnostic ultrasound diagnoses were standardized in terms of a total of 24 possible outcomes, leading to 10 primary diagnostic outcome categories: 1) tendinopathy, 2) calcification, 3) full thickness tear or 4) partial thickness tear, 5) biceps tendon tear, 6) subacromial-subdeltoid bursitis, 7) subacromial impingement, 8) osteoarthritis of the acromio-clavicular joint, 9) no pathology, or 10) other (e.g. labral tear, capsular thickening).

In case a diagnosis in category 1-4 was made, it was specified by adding the affected tendon; supraspinatus, subscapularis and infraspinatus / teres minor.

The PT and radiologist were not limited in the number of positively scored outcomes.

Blinding

The radiologist was blinded for the PT's diagnostic ultrasound diagnosis. The patient did not receive the diagnostic ultrasound diagnosis of the PT or radiologist.

Analysis

Descriptive statistics, including frequencies for categorical variables and means and standard deviations (in case the data did not show a normal distribution, medians and interquartile ranges (IQR)) for continuous variables were calculated to summarize the characteristics of the patients, PTs and radiologists. The prevalence of positive findings and the frequencies of particular diagnosis were calculated.

Only a Cohen's kappa and 95% confidence interval (95%CI) [29, 30] was calculated for the "full thickness tear", "partial thickness tear", "calcification" and "bursitis" categories, to evaluate the interrater-reliability of PTs and radiologists. The information of the other diagnostic categories will only be expressed in raw data. Next an overall Cohen's kappa was calculated, based on all four categories. Finally, observed agreement (OA), specific positive agreement (SPA) and specific negative agreement (SNA) was calculated, as these are regarded relevant for clinicians. The specific positive agreement, is calculated by the following formula: $SPA=2a/(2a+b+c)$, while specific negative agreement, is calculated using the formula: $SNA=2d/(2d+b+c)$ [31]. For the interpretation of the kappa values, the following criteria were used: almost perfect (0.81–1.00), substantial (0.61–0.80), moderate (0.41–0.60), fair (0.21–0.40), slight (0.01–0.20) or poor (<0.00) agreement [32, 33].

Two subgroup analyses were performed based on a) experience (more or less than 200 diagnostic ultrasounds of the shoulder) [15, 17] and b) education level (basic or advanced). All PTs were trained to use diagnostic ultrasound of the shoulder and some PTs followed specific additional courses and were therefore labelled as advanced.

All statistical analyses were performed using SPPS 20 software.

RESULTS

Raters

In this study a total of 13 PTs and 9 radiologists met the selection criteria and participated. Table 1 presents the characteristics of the participating physiotherapists and radiologists.

TABLE 1. Characteristics of the raters.

Variable	PTs (N=13)	Radiologist (N=9)
Gender: N (% male)	Male: 13 (100%)	Male: 9 (100%)
Age: Mean (SD)	40 (8)	49 (3)
Years of experience: Median (IQR)	5 (1.5-6)	10 (5-20)
Number of DMUS made: N (%)	<200: 6 (46%) >200: 7 (54%)	>200: 9 (100%)
Education or setting: N (%)	Basic course: 5 (38%) Master course: 8 (62%)	MD and radiology training
Transducer frequencies MHz: N (%)	5-10: 8 (62%) 5-12: 3 (23%) 5-13: 2 (15%)	3-10: 3 (33%) 3-12: 4 (44%) 3-15: 2 (22%)

Abbreviations: PT, Physiotherapist; N, Number; SD, Standard deviation; IQR, Interquartile range; MHz, Megahertz; DMUS, Diagnostic Musculoskeletal Ultrasound.

The 13 PTs were able to include patients who received a second diagnostic ultrasound, with a median of three patients per PT (IQR 1-6). All participating PTs and radiologists were male.

Patients

A total of 65 patients participated and received both diagnostic ultrasounds. Another 41 patients received a diagnostic ultrasound from their PT only. Of these, 12 patients were unwilling to participate in the reliability study and 29 patients were excluded as a result of late scheduling of the second diagnostic ultrasound.

Demographic characteristics of the 65 patients are presented in table 2. The mean age of the patients was 56 years and 54% were male. The median duration of shoulder pain was 12 weeks.

TABLE 2. Characteristics of the patients.

Variable	Frequencies
Gender: N (%male)	35 (54%)
Age: Mean (SD)	56 (12)
Duration of complaints in weeks: Median (IQR)	12 (6-29)
Medication use: N (%yes)	31 (52%)
Pain Score[1]: Median (IQR)	6 (5-7)
SPADI[2]: Median (IQR)	51 (35-67)
SDQ[3]: Median (IQR)	71 (50-87)
EQ5D health status[4]: Median (IQR)	7 (6-8)
Data of the questionnaires of three patients missing.	

Abbreviations: N, Number; SD, Standard deviation; IQR, Interquartile range.

Legend:

1. The pain score has been measured using the Numeric Rating Scale (NRS) ranging from 0 to 10, with 0 no pain and 10 worst pain ever.
2. The Shoulder Pain and Disability Index (SPADI) ranges from 0 to 100, a higher score indicates a higher level of disability.
3. The Shoulder Disability Questionnaire (SDQ) ranges from 0 to 100, a higher score indicating more severe disability.
4. The Euroqol (EQ5D) health status ranges from 0 to 10, 0 represents the worst possible health status and 10 the best possible health status.

Outcomes

The prevalence of positive findings on the main categories (based on the high diagnostic accuracy of diagnostic ultrasound performed by a radiologist as reported in the literature) and kappa values are reported in Table 3.

TABLE 3. Results of agreement.

Diagnostic category	Frequency	Cohen's kappa	Agreement
Overall		Kappa: 0.36 (95%CI 0.29- 0.43)	OA: 80% SPA: 51% SNA: 86%
Full thickness tear	PT: 6 Radiologist: 9 Both: 5	Kappa: 0.63 (95%CI: 0.31-0.94)	OA: 92% SPA: 67% SNA: 96%
Partial thickness tear	PT: 13 Radiologist: 6 Both: 2	Kappa: 0.10 (95%CI: 0.00-0.49)	OA: 77% SPA: 21% SNA: 86%
Calcification	PT: 20 Radiologist: 31 Both: 14	Kappa: 0.28 (95%CI: 0.04-0.52)	OA: 65% SPA: 55% SNA: 71%
Subacromial bursitis	PT: 7 Radiologist: 16 Both: 7	Kappa: 0.54 (95%CI: 0.26-0.82)	OA: 86% SPA: 61% SNA: 92%

Abbreviations: PT, Physiotherapist; OA, Observed agreement; SPA, Specific positive agreement; SNA, Specific negative agreement.

Prevalence

The prevalence of a calcification was highest, 22% [14/65], based on how often the PT and the radiologist both placed the same subject in this diagnostic category.

The prevalence of a bursitis, 25% [16/65] versus 11% [7/65], and a calcification, 48% [31/65] versus 31% [20/65], was substantially higher according to the radiologist as opposed to the PT.

Reliability

The kappa for the full thickness tears was 0.63, indicating substantial agreement. We found moderate agreement (0.54K) for bursitis, fair agreement (0.28K) for calcification, and slight agreement (0.10K) for partial thickness tears. The overall kappa of all four main diagnostic categories was 0.36 (95%CI 0.29- 0.43), indicating fair agreement. The overall observed agreement, based on these four categories, was 80%, the SPA was 51% and the SNA was 86%.

Subgroups

Subgroup analysis showed an overall kappa in the more experienced group of 0.43 (95%CI 0.25-0.63) (moderate) compared to a kappa of 0.17 (95%CI -0.15-0.50) (slight) in the less experienced group. Furthermore, we found a kappa of 0.43 (95%CI 0.27-0.60) (moderate) in the advanced course group compared to of 0.09 (95%CI -0.30-0.48) (slight) in the basic course group.

The prevalence of positive findings on the other categories is presented in Table 4.

TABLE 4. Prevalence's of ultrasound diagnoses.

Diagnostic category	Frequency
Tendinopathy	PT: 28 Radiologist: 23 Both: 13
Biceps tendon tear	PT: 2 Radiologist: 2 Both: 2
Subacromial impingement	PT: 14 Radiologist: 6 Both: 2
Arthritis of arthrosis of the acromio- clavicular joint	PT: 7 Radiologist: 8 Both: 3
No pathology	PT: 3 Radiologist: 7 Both: 2
Other	PT: 13 Radiologist: 9

Abbreviations: PT, Physiotherapist; OA, Observed agreement; SPA, Specific positive agreement; SNA, Specific negative agreement.

The prevalence of the other shoulder disorders ranged from 3 to 20%. In three percent of patients both the PT and radiologist found no pathology. The PT most frequently labelled patients with a tendinopathy (43%).

DISCUSSION

The results from this study were disappointing and indicative of low trustworthiness of current shoulder pathology-related ultrasound diagnoses by Dutch PTs. Our study showed borderline substantial agreement (0.63K) between PTs and radiologists in diagnosing a full thickness tear only. The overall kappa of all four categories was 0.36, indicating fair agreement. Subgroup analysis on both experience and education level showed that the more experienced and higher trained PTs showed moderate agreement with the radiologist compared to only slight agreement for the less experienced and basic trained PTs.

We found higher specific negative agreement compared to specific positive agreement in all diagnostic categories. This indicated that disagreement was more often found in diagnosing a patient with a certain pathology as compared to ruling the presence of a pathology out.

Comparison with the literature

To our knowledge, this is the first reliability study between different professions on diagnostic ultrasound in primary care in symptomatic patients. Secondary care reliability studies between radiologists showed substantial to almost perfect agreement for the full thickness tear and partial thickness tear categories [15, 16]. Compared to these results, the agreement between PTs and radiologist in our study, showed only limited promising results for diagnosing full thickness tears. Unfortunately, because of the relatively low numbers, a firm conclusion is not possible. Compared to other studies, our study population differed, as it was a primary care population and the prevalence of specific pathology was therefore low. Most likely the severity of disorders was also lower in our study population compared to a hospital care population. The pre-test probability therefore is much higher in a secondary care setting, which influences the sensitivity and specificity and possibly also the reliability [34]. The reliability on calcification and bursitis between radiologists has not been mentioned in the literature, but sensitivity and specificity were high [10].

Strengths and limitations

Multiple physiotherapy and radiology departments were included to account for a large as possible sample size and generalizability. We only included patients from primary care selected by the PT to receive a diagnostic ultrasound. Therefore, prevalence of separate diagnostic categories was relatively low and so conclusions should be interpreted with caution.

Scanning methods were standardized using the international scanning protocol [28] which is focused on anatomy, technique and scanning pitfalls. The protocol was used to standardize the procedure, but with respect to daily clinical practice. The protocol did not include "pre-sets" (e.g. gain, depth, transducer frequency) and is less strict than others. To our knowledge no published scanning protocol has proven to be superior towards other protocols. However, we did not measure the adherence to this protocol, which might have influenced the reliability negatively.

Disagreement often consisted of one labelling the patient as having a partial thickness tear and the other clinician labelling the patient as having a tendinopathy or calcification of the same structure. This might be due to a bias of PTs towards a certain diagnosis. The difficulty in differentiating between both pathologies has been mentioned before in the literature [35, 36]. It is unknown if this discrepancy can be influenced by stricter diagnostic criteria. Possibly this could improve reliability in the future.

We defined a minimum transducer frequency for ultrasound equipment of the PTs, in order to increase the possibility of high agreement. The demands on transducer frequencies were based on the literature, where frequencies in the range of 5 to 13 MHz are most commonly used [19]. A frequency of 7.5 MHz seems to be recommended for

detecting rotator cuff tears [14]. Most useful for diagnostic ultrasound are linear array transducers with in-line piezoelectric elements and a flat probe surface [37]. Although the equipment used by PTs and radiologists does not differ on transducer frequency and beam former technology, there is a difference in costs of the machines. However, a study measuring the muscle thickness, medio- lateral length and cross-sectional area of the hallucis muscle showed that regardless the use of two different machines (a higher end Philips HD11 Ultrasound Machine and clinically orientated Chison 8300 Deluxe Digital Portable Ultrasound System) the intra-reliability was very high [38]. Excellent reliability was also reported using two different machines measuring the patellar tendon [39]. It is unknown, however, if the differences in equipment in this study affected the level of agreement but these differences reflect usual care.

Besides the difference in equipment, radiologists were more experienced in performing ultrasounds than the PTs in our study, due to the nature of their profession (performing ultrasounds on a daily basis) and the number of years they had been working with ultrasonography.

This broad participation (the number of raters), the use of their own equipment and a less strict protocol increases the external validity of our results.

In order to minimize progression bias, we chose a maximum period of one week between both tests. We assume that the conditions of interest (full thickness tear, partial thickness tear, calcification and bursitis) did not change within this time frame.

Due to the absence of a true gold standard we chose to perform a reliability study instead of a validation study. Only in case of the full thickness tear, a validation study could have been a possibility, due to the very high diagnostic accuracy of the radiologists. However, it is unknown if this high level of diagnostic accuracy can be extrapolated to a different study population. The diagnostic accuracy of a partial thickness tear, bursitis and calcification appears to be lower and is therefore not appropriate to be used as a gold standard. However, radiologists are the experts in the field of diagnostic ultrasound and perform diagnostic ultrasound on a regular basis and although the diagnostic ultrasound of the radiologist cannot be used as gold standard, the diagnostic accuracy of all these diagnostic labels is still high. Therefore, a high level of agreement between PTs and radiologist could have been an indication that the use of diagnostic ultrasound in primary care by PTs is appropriate.

Implications for clinical practice

This study showed substantial agreement between PTs and radiologists in diagnostic ultrasounds for full thickness tears. However, the Kappa value of 0.63 is at the lower end of the range for being classified as substantial, so the clinical significance of that level of reliability is relatively low when compared with the high reliability between radiologists. At this moment it seems PTs use diagnostic ultrasound to diagnose a patient with a

specific pathology. However, based on the overall agreement, which was fair, diagnostic ultrasound performed in primary care by PTs should not be integrated in their diagnostic clinical practice yet.

There was a high level of specific negative agreement across all diagnostic categories. This implies that in ruling pathologies out the agreement is higher compared to ruling them in. This could perhaps be useful in clinical practice. PTs often use exercises in their rehabilitation program. However, it is hypothesized that training a muscle affected by a full or partial thickness tear could potentially be harmful to the structure. Because there is a high level of agreement between the radiologist and the PT in ruling these pathologies out, the PT could perform a diagnostic ultrasound to examine if there is no contra-indication to loading the tendon and there is no limitation in the level of exercising this tendon based on the physiology.

Subgroup analysis showed that both education and experience lead to higher level of agreement with the radiologist, although the agreement is still limited. Diagnostic ultrasound is a relatively young diagnostic technique for PTs and developments within education, equipment and learning curves of PTs could eventually lead to a higher level of diagnostic ultrasound of PTs.

Therefore, it is important to acknowledge there might be a possible added value of diagnostic ultrasound in the future of the physiotherapy profession. At this moment, however, conclusions based on the results of the diagnostic ultrasound of the PT only, should be interpreted with significant caution.

Future research on the reliability of diagnostic ultrasound performed by PTs in patients with shoulder pain should take into account the variability in interpretation and pathology definitions. We therefore recommend stricter diagnostic criteria, definition of outcomes and guidelines in future research. Besides research concerning the inter-rater reliability between PTs and radiologists, the inter- and intra-rater reliability between PTs should be assessed. A larger study is warranted to determine the effects of experience and education on the level of agreement. Additionally, it would be of interest whether the use of diagnostic ultrasound guides differences in treatment processes and if this results in different treatment outcomes.

REFERENCES

1. Luime, J.J., et al., *Prevalence and incidence of shoulder pain in the general population; a systematic review.* Scand J Rheumatol, 2004. **33**(2): p. 73-81.
2. Picavet, H.S. and J.S. Schouten, *Musculoskeletal pain in the Netherlands: prevalences, consequences and risk groups, the DMC(3)-study.* Pain, 2003. **102**(1-2): p. 167-78.
3. Huisstede, B.M., et al., *Incidence and prevalence of upper-extremity musculoskeletal disorders. A systematic appraisal of the literature.* BMC Musculoskelet Disord, 2006. **7**: p. 7.
4. Linsell, L., et al., *Prevalence and incidence of adults consulting for shoulder conditions in UK primary care; patterns of diagnosis and referral.* Rheumatology (Oxford), 2006. **45**(2): p. 215-21.
5. Dorrestijn, O., et al., *Patients with shoulder complaints in general practice: consumption of medical care.* Rheumatology (Oxford), 2011. **50**(2): p. 389-95.
6. Leemrijse, C.J., I.C. Swinkels, and C. Veenhof, *Direct access to physical therapy in the Netherlands: results from the first year in community-based physical therapy.* Phys Ther, 2008. **88**(8): p. 936-46.
7. Beaudreuil, J., et al., *Contribution of clinical tests to the diagnosis of rotator cuff disease: a systematic literature review.* Joint Bone Spine, 2009. **76**(1): p. 15-9.
8. Hughes, P.C., N.F. Taylor, and R.A. Green, *Most clinical tests cannot accurately diagnose rotator cuff pathology: a systematic review.* Aust J Physiother, 2008. **54**(3): p. 159-70.
9. Park, H.B., et al., *Diagnostic accuracy of clinical tests for the different degrees of subacromial impingement syndrome.* J Bone Joint Surg Am, 2005. **87**(7): p. 1446-55.
10. Ottenheijm, R.P., et al., *Accuracy of diagnostic ultrasound in patients with suspected subacromial disorders: a systematic review and meta-analysis.* Arch Phys Med Rehabil, 2010. **91**(10): p. 1616-25.
11. Lahteenmaki, H.E., et al., *Repair of full-thickness rotator cuff tears is recommended regardless of tear size and age: a retrospective study of 218 patients.* J Shoulder Elbow Surg, 2007. **16**(5): p. 586-90.
12. Millett, P.J., et al., *Rehabilitation of the rotator cuff: an evaluation-based approach.* J Am Acad Orthop Surg, 2006. **14**(11): p. 599-609.
13. de Jesus, J.O., et al., *Accuracy of MRI, MR arthrography, and ultrasound in the diagnosis of rotator cuff tears: a meta-analysis.* AJR Am J Roentgenol, 2009. **192**(6): p. 1701-7.
14. Smith, T.O., et al., *Diagnostic accuracy of ultrasound for rotator cuff tears in adults: a systematic review and meta-analysis.* Clin Radiol, 2011. **66**(11): p. 1036-48.
15. Rutten, M.J., G.J. Jager, and L.A. Kiemeney, *Ultrasound detection of rotator cuff tears: observer agreement related to increasing experience.* AJR Am J Roentgenol, 2010. **195**(6): p. W440-6.
16. Le Corroller, T., et al., *Sonography of the painful shoulder: role of the operator's experience.* Skeletal Radiol, 2008. **37**(11): p. 979-86.
17. Alavekios, D.A., et al., *Longitudinal analysis of effects of operator experience on accuracy for ultrasound detection of supraspinatus tears.* J Shoulder Elbow Surg, 2013.
18. Murphy, R.J., et al., *An independent learning method for orthopaedic surgeons performing shoulder ultrasound to identify full-thickness tears of the rotator cuff.* J Bone Joint Surg Am, 2013. **95**(3): p. 266-72.
19. Dinnes, J., et al., *The effectiveness of diagnostic tests for the assessment of shoulder pain due to soft tissue disorders: a systematic review.* Health Technol Assess, 2003. **7**(29): p. iii, 1-166.
20. Rutten, M.J., et al., *Detection of rotator cuff tears: the value of MRI following ultrasound.* Eur Radiol, 2010. **20**(2): p. 450-7.
21. Teyhen, D.S., *Rehabilitative ultrasound imaging for assessment and treatment of musculoskeletal conditions.* Man Ther, 2011. **16**(1): p. 44-5.

22. Potter, C.L., M.C. Cairns, and M. Stokes, *Use of ultrasound imaging by physiotherapists: a pilot study to survey use, skills and training.* Man Ther, 2012. **17**(1): p. 39-46.
23. McKiernan, S., P. Chiarelli, and H. Warren-Forward, *Diagnostic ultrasound use in physiotherapy, emergency medicine, and anaesthesiology.* Radiography, 2010. **16**(2): p. 154-159.
24. Kumar, P., et al., *Interrater and intrarater reliability of ultrasonographic measurements of acromion-greater tuberosity distance in healthy people.* Physiother Theory Pract, 2011. **27**(2): p. 172-5.
25. Kumar, P., M. Bradley, and A. Swinkels, *Within-day and day-to-day intrarater reliability of ultrasonographic measurements of acromion-greater tuberosity distance in healthy people.* Physiother Theory Pract, 2010. **26**(5): p. 347-51.
26. Scholten-Peeters, G.G., et al., *The opinion and experiences of Dutch orthopedic surgeons and radiologists about diagnostic musculoskeletal ultrasound imaging in primary care: A survey.* Man Ther, 2013.
27. Karel, Y.H., et al., *Current management and prognostic factors in physiotherapy practice for patients with shoulder pain: design of a prospective cohort study.* BMC Musculoskelet Disord, 2013. **14**(1): p. 62.
28. Jacobson, J.A., *Shoulder US: anatomy, technique, and scanning pitfalls.* Radiology, 2011. **260**(1): p. 6-16.
29. Sim, J. and C.C. Wright, *The kappa statistic in reliability studies: use, interpretation, and sample size requirements.* Phys Ther, 2005. **85**(3): p. 257-68.
30. Kottner, J., et al., *Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed.* Int J Nurs Stud, 2011. **48**(6): p. 661-71.
31. de Vet, H.C., et al., *Clinicians are right not to like Cohen's kappa.* BMJ, 2013. **346**: p. f2125.
32. Landis, J.R. and G.G. Koch, *The measurement of observer agreement for categorical data.* Biometrics, 1977. **33**(1): p. 159-74.
33. Viera, A.J. and J.M. Garrett, *Understanding interobserver agreement: the kappa statistic.* Fam Med, 2005. **37**(5): p. 360-3.
34. Leeflang, M.M., P.M. Bossuyt, and L. Irwig, *Diagnostic test accuracy may vary with prevalence: implications for evidence-based diagnosis.* J Clin Epidemiol, 2009. **62**(1): p. 5-12.
35. Bianchi, S., C. Martinoli, and I.F. Abdelwahab, *Ultrasound of tendon tears. Part 1: general considerations and upper extremity.* Skeletal Radiol, 2005. **34**(9): p. 500-12.
36. Jamadar, D.A., et al., *Musculoskeletal sonography: important imaging pitfalls.* AJR Am J Roentgenol, 2010. **194**(1): p. 216-25.
37. Force, A.C.o.R.M.U.T., *Ultrasound in American rheumatology practice: report of the American College of Rheumatology musculoskeletal ultrasound task force.* Arthritis Care Res (Hoboken), 2010. **62**(9): p. 1206-19.
38. Hing, W.A., K. Rome, and A.F. Cameron, *Reliability of measuring abductor hallucis muscle parameters using two different diagnostic ultrasound machines.* J Foot Ankle Res, 2009. **2**: p. 33.
39. Gellhorn, A.C. and M.J. Carlson, *Inter-rater, intra-rater, and inter-machine reliability of quantitative ultrasound measurements of the patellar tendon.* Ultrasound Med Biol, 2013. **39**(5): p. 791-6.

Karel Y, Thoomes-De Graaf M, Scholten-Peeters G, Ferreira P, Rizopoulos D, Koes BW, Verhagen AP.

Physiother Theory Pract. 2017 Nov 9:1-9. doi: 10.1080/09593985.2017.1400141. [Epub ahead of print]

VALIDITY OF THE FLEMISH WORKING ALLIANCE INVENTORY IN A DUTCH PHYSIOTHERAPY SETTING IN PATIENTS WITH SHOULDER PAIN

CHAPTER 6

ABSTRACT

Background: Working alliance is the interaction between the patient and therapist. It is a crucial part of the physiotherapeutic process. One instrument to measure working alliance is available in Dutch/Flemish language and validated in psychotherapy setting.

Objective: This study aims to validate the Working Alliance Inventory Short-Form in a Dutch physiotherapy setting.

Design: A prospective cohort study in primary-care physiotherapy.

Method: To validate the Dutch/Flemish version of the working alliance inventory short-form (WAV-12) a RASCH analysis was used.

Results: Sixty-six physiotherapists enrolled in total 389 patients with an average age of 50 years and a mean duration of shoulder pain of 33 weeks. A total of 274 patients filled in one or more items of the WAV-12. The WAV-12 showed good discriminative abilities and all items contributed to a one-dimensional construct. Due to the selective nature of the missing items we believed rewording was necessary to make it more suitable to the physiotherapy setting. We performed a Delphi study and revised the WAV-12 into the PAS (Physio Alliance Scale). The validity of the revised version is unknown and is therefore not sufficiently strong to be implemented as a measurement tool.

Limitations: The response rate for three items especially was low and we found ceiling effects in ten items.

Conclusion: Although the measurement instrument shows good internal consistency and reliability, we made adjustments to the WAV-12 for Dutch physiotherapy setting.

Keywords: Working alliance, Therapeutic alliance, Rasch analysis, WAV-12, physiotherapy, physician-patient relations.

INTRODUCTION

In physiotherapy practice patients usually follow a treatment regimen provided in coherence with the physiotherapist. This interaction between patient and therapist is referred to as a working alliance (WA). WA is first described in psychotherapy as the extent to which a client and therapist work collaboratively, purposefully and connect emotionally. WA is defined as a combination of three factors; agreement about the goals of treatment, the tasks of treatment and the bond between client and therapist [1].

For a treatment to be effective one important factor is that the patient complies with the regimen, after which health outcomes are more likely to improve [2]. Therefore, it is essential for the therapist to provide a proper transfer of information about the goals and tasks of treatment for the patient in order to carry out the treatment regimen [3, 4]. Besides agreement about treatment goals and tasks, co-operation and compliance are achieved by means of bonding and trust between the therapist and the patient. Patients consult a physiotherapist because they seek help and they are in that case vulnerable. Help must therefore be offered and accepted based on trust. How this relationship will develop during the treatment period can have a significant impact on treatment outcome.

Several reviews have found that WA is a strong predictor of improvement in psychotherapy and psychology practices [5, 6]. Later research has established the importance of a good alliance also in other medical settings, such as in patients with ulcer disease, hypertension and diabetes [7, 8]. One review included 14 studies examining the patient-therapist relationship in physical rehabilitation setting [9]. In nine studies a registered physiotherapist delivered the interventions. Results of the individual studies indicated that WA has a consistent positive correlation to treatment outcomes of pain, disability, physical/mental health and patient satisfaction [9]. A recent observational study of therapeutic alliance in patients with chronic low back pain confirmed these findings and found WA to be a consistent predictor of function, pain and disability measures [10]. WA might be more important in some therapies especially in those where treatment adherence represents an important component for treatment effect [11].

The Working Alliance Inventory (WAI) is one of the most commonly used and validated questionnaires to measure the working alliance [9]. It has been originally developed as a 36-item questionnaire based on Bordin's model measuring three domains; goal, task and bond [12, 13]. The WAI exists of one questionnaire for the client (WAI (C)) and one for the therapist (WAI (T)). Evidence suggests that the clients WA rating at the beginning of treatment is superior over the therapist rated version in predicting outcome [5].

The WAI was translated to Flemish, which is closely related to Dutch, named the "werk alliantie vragenlijst" (WAV). The 12 most indicative items were selected using confirma-

tory factor analysis to form the WAV-12 short form [14]. The WAV-12 has been used and validated in patients receiving psychotherapy in Belgium [15]. This study found a good internal consistency for the three-factor model according to Bordin: (1) task scale-correlation coefficient α=0.85; (2) bond scale-correlation coefficient α=0.82; and (3) goal scale- correlation coefficient α=0.83. Correlations between the task and goal scales were good (correlation coefficient r=0.80) but correlations between the other scales were both lower (Cronbach's α=0.49). The WAV-12 used a 5-point Likert scale instead of a 7-point Likert scale in the original WAV-36. Therefore, it is difficult to compare results from this validation study with other data. Literature does describe slightly higher correlation coefficients for the English and French short versions [14, 16]. A review has shown that translated versions of a measurement instrument for the neck do not guarantee similar measurement properties compared with the original instrument [17]. Cross-cultural validation in the Dutch population and physiotherapy setting is an important step to evaluate whether the underlying construct still holds for the WAV-12.

Therefore, this study aims to investigate whether the WAV-12 is a valid measurement instrument in terms of the construct and discriminative abilities for a population of patients with shoulder pain in physiotherapy care.

METHODS

Study design

The study population consisted of patients with shoulder pain that participated in a prospective cohort study in patients consulting a physiotherapist for shoulder pain [18]. Recruitment period was from November 2011 through December 2012. The Research Committee of the Erasmus Medical Centre in Rotterdam approved the project (MEC-2011-414). After signing an informed consent, patients were included and followed up for 6 months.

Participants

A total of 125 physiotherapists were invited to enrol patients. Patients consulting a physiotherapist were included if they suffered from shoulder pain, were aged ≥ 18 years and had adequate understanding of the Dutch language. Patients were excluded if they had serious pathologies (infection, cancer or fracture), surgery of the shoulder in the previous 12 months, or had received diagnostic imaging techniques such as musculoskeletal ultrasound, magnetic resonance imaging or X-ray of the shoulder in the 3 months prior to start of the study. Patients included in the cohort study were followed for 6 months and received usual physiotherapy care. Questionnaires were sent by email at 6, 12 and

26 weeks and 2 reminders were sent after 2 and 4 days whenever the patient had not responded to the questionnaire.

Working Alliance (WA)

WA was measured 6 weeks after baseline for both the patient and physiotherapist, because earlier assessment would not clearly reflect the WA. We used the Flemish version of the WAI (WAV-12). It contains 12 items scored on a five-point scale ranging from 1 ("never") to 5 ("always") and scoring is done for the total score and each subscale (goal, task and bond). The total score ranges from 12 (low WA) to 60 (high WA), and subscales range from 4 to 20. Where the patient had to fill in the name of the therapist we replaced the empty space with the words: "my therapist".

Statistical analysis

Descriptive data for demographic and symptom severity are presented as percentages for nominal variables (i.e. gender, level of education, cause of injury, first episode, reasons for stopping treatment) and as means for continuous variables (i.e. age, symptom duration). T-tests were used to test for differences in demographics between participants scoring all WAV-12 items and those who did not. Cronbach's alpha was used to assess the internal consistency of the WAV-12 and we assessed the correlation between patient and therapist scores using Pearsons' correlation coefficient. Coefficients equal or more than 0.7 were regarded as acceptable. R and SPSS v20.0 were used to conduct the analysis.

Validation

Performance of the items in the WAV-12 questionnaire was assessed with a partial credit Rasch model [19]. The response patterns from the set of available items in the questionnaire were tested against what is expected by the model that works according to a probabilistic form of Guttman scaling [20]. This scale assumes a deterministic pattern with a hierarchical ordering of items (low and high level of item scale). When a higher level of the item is affirmed, there must be a high probability that lower items will also be affirmed. The analysis gives the probability that a person will affirm an item of the difference between the person's level of working alliance and the level of working alliance expressed by the item.

The Rasch model was used to test; (1) internal validity of the construct; (2) whether specific items exhibit different properties in different subgroups in the population (differential item functioning); and (3) whether item redundancy can be considered [21]. Analysis was done using the ltm package in the statistical programing language R [22].

First a one partial credit model with the discrimination parameter fixed at one was tested to check whether it fits the data. If this model did not fit the data an extended

partial credit model with a common discrimination parameter not constrained at one or separate discrimination parameters for each parameter was considered. Uni-dimensionality could further be examined to investigate if the test variance is attributable to the principal factor or construct, estimated with Cronbach's alpha. Due to the fact that some patient responses were missing, multiple imputations were utilized to calculate Cronbach's alpha.

Differential item functioning was examined based on a likelihood ratio X^2 test implemented in the Lordif package in R. Expected scores for each item should remain the same whether, an older or younger person (<50, which was the mean age) and a man or women scores the same item.

Rasch analysis can be useful and psychometrically sound in modifying measurement instruments [23]. Different criteria could be considered for item redundancy: High Item Characteristic Curve (ICC), low ICC or items having similar calibrations.

RESULTS

Study population

Sixty-six physiotherapists enrolled in total 389 patients. Physiotherapists were 72% male and had a mean working experience of 15 years.

Of the 389 patients 43% were male, average age was 50 years with a mean duration of shoulder pain of 33 weeks (see table 1). At baseline, only 4% of the patients did not fill out the baseline questionnaire. At 6 weeks 30% of the responses were lost to follow up.

Working alliance

Seventy-eight patients (22%) filled in all the WAV-12 questions, enabling us to calculate a total score. The mean WAV score was 45 on a total range of 24 to 60, which is slightly above 50% of the maximum score. Most patients did not answer one or more questions of the WAV-12. The population that had responded to all WAV-12 questions did not significantly differ at baseline with the patients that did not (see Table 1). Even though not statistically significant, the difference for duration of complaints appeared to be large. Selective responses can therefore not be excluded. The questions with the most missing values are questions 1, 3, 7 and 9 (see Figure 1). Question 3, 7 and 9 are part of the "bond" subscale and question 1 is part of the "goal" subscale. The working alliance score of therapists was 52 and for patients 45. WAV-12 scores between patient and therapist had a poor correlation (r=0.30).

TABLE 1. Baseline characteristics

Characteristics of cohort	Total n=389	Participants filling in all items of WAV-12; n=87	Participants, missing 1 or more items of WAV-12; n=302
Male (%)	**170 (43)**	**41 (49)**	**129 (44)**
Age (SD)	**50 (13)**	**50 (14)**	**50 (13)**
Duration of complaint in weeks (SD)	**33 (82)**	**27 (58)**	**34 (88)**
Comorbidity (%)			
No	**128 (35)**	**25 (29)**	**103 (34)**
Yes	**236 (65)**	**62 (71)**	**199 (66)**
Medication use (%)	**183 (47)**	**40 (49)**	**144 (50)**
Highest education (%)			
Primary school	**40 (10)**	**12 (15)**	**28 (10)**
High school	**199 (51)**	**44 (54)**	**155 (54)**
University or applied sciences	**127 (33)**	**25 (31)**	**102 (36)**
Paid job (%)	**261 (67)**	**53 (65)**	**208 (72)**
Profession (%)			
Physically intensive job	**65 (17)**	**13 (25)**	**52 (25)**
Static repetitive job	**88 (23)**	**14 (27)**	**74 (35)**
Job with awkward positions/ postures	**11 (3)**	**3 (6)**	**8 (4)**
Other	**99 (25)**	**22 (42)**	**77 (36)**
NRS median (IQR)	**6.0 (3.0)**	**6.0 (2.0)**	**6.0 (3.0)**
SDQ (SD)	**62 (23)**	**63 (24)**	**62 (23)**
EQ-5D (SD)	**0.83 (0.08)**	**0.82 (0.07)**	**0.83 (0.09)**

Abbreviations: NRS Numeric Rating Scale, SDQ Shoulder Disability Index, EQ-5D EuroQol 5 Dimensions, SD standard deviation

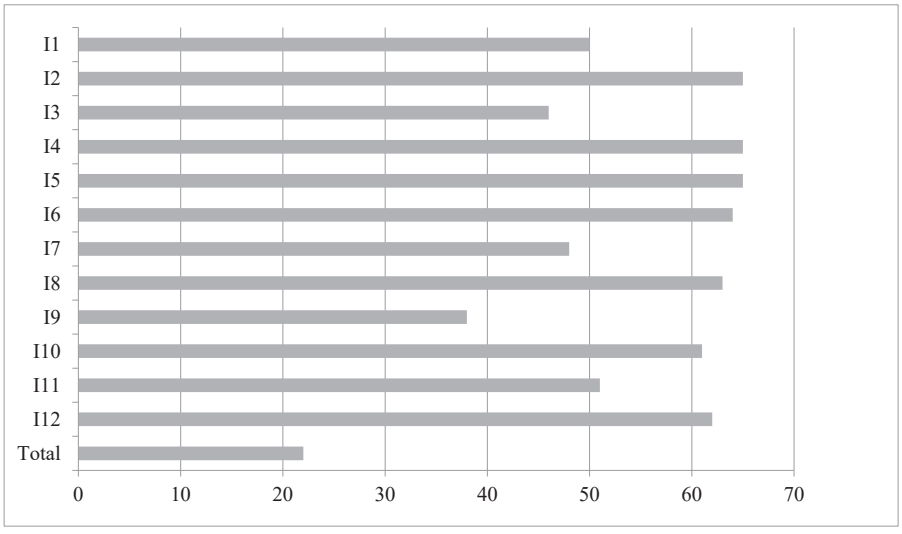


FIGURE 1. Relative response rate per item of WAV-12
I Item 1-12 of the WAV-12

Validity of WAV-12

Of all patients, 274 had at least filled in one or more items of the WAV-12. Three models were fitted to the data. The first model (RASCH) assumes the discrimination parameter is equal for all items and fixed at one. The second model (1PL) assumes the discrimination parameter is equal for all items but is estimated from the data and the third model (gpcm) assumes the discrimination parameter is free to vary across items. Likelihood ratio tests between these models showed that the third model provided the best fit to the data (p=<0.001).

Item properties

All but two items (item 1 and 2), showed ceiling effects, meaning that most of the patients scored a good working alliance. Appendix 1 displays the item characteristic curves for the 12 items from the WAV-12. Items 5, 6 and 8 have a high slope and are endorsed at higher levels of working alliance. Items 1, 2 and 4 have a low slope (discrimination) and are endorsed at lower levels of WA. Considerable variation exists between item discrimination indicating the WAV-12 questionnaire includes items measuring the whole construct and items discriminating at lower and higher levels of working alliance (Table 2). The item information curve showed the amount of information given by the questionnaire is highest between an ability of -2 and 0, implying that the item set is most useful in discriminating among individuals at the lower end of the working alliance trait.

TABLE 2. Discrimination values of WAV-12 items

Item	Discrimination	Standard error	Z value
1	0.496	0.103	4.793
2	0.443	0.088	5.066
3	1.286	0.225	5.716
4	0.761	0.118	6.424
5	2.212	0.457	4.842
6	2.067	0.338	6.114
7	1.377	0.234	5.895
8	2.266	0.369	6.139
9	1.151	0.208	5.537
10	1.068	0.158	6.742
11	1.414	0.224	6.319
12	1.107	0.167	6.613

Unidimensionality

Five imputed datasets were created. Cronbach's alpha's were calculated for the 12 items in each dataset and led to a pooled Cronbach's alpha coefficient of 0.89. Indicating that items correlate highly and measure the same explanatory concept.

Differential Item Functioning (DIF)

The X^2 tested three models. Model 1 is a standard model where the ability for each person remains the same. Model 2 tests whether levels of ability differ among groups and model 3 adds an interaction term for the level of ability and the group in order to test whether discrimination parameters differ among groups.

Age was dichotomized in younger patients (under the mean age of 50) and older patients (50 and over). The X^2 tested flagged item one for differential item functioning where all models were statistically significant. No differential item functioning was found between men and women. Slightly higher factor scores (mean difference = 0.0385) for the WA in patients being treated by a physiotherapist with less than 13 years of experience but was not statistically significant (p=0.73).

Rasch analysis for the WAV-12 questionnaire indicates that items have good discriminative abilities for the lower end of the construct. High correlations coefficients indicate items measure one construct and other factors like age and experience of the physiotherapist did not influence item scoring. Validity for the items in the questionnaire appears to be sound but due to the difference in the percentage of missing data among the items and observed ceiling effects we advise linguistic (Dutch) and contextual (physiotherapeutic setting) adjustments.

Modification of the WAV-12

We believed rewording was necessary due to the selective number of missing responses in some items of the questionnaire and because the researchers had received comments from several patients and physiotherapists about items 3, 7 and 9 of the WAV-12. Therefore, we decided to make adjustments in the questionnaire and did a Delphi study. A two-round survey was employed to ask the panels opinion on the adjustments in the WAV-12. The panel consisted of 11 members (six clinical/research experts and five patients). Panel members were sent a questionnaire via email and these were sent separately to ensure panel members were unaware each of other's identity. For each item the panel member had to give his/her opinion about the adjustments with a five-point Likert scale. If the score was below three (neutral, disagree, totally disagree) the panel member was asked to give their reasoning and/or a suggestion for adjustment. If consensus for one item was < 80% after the first round it was included in the second round containing the suggestions of all panel members (anonymous). Full consensus (100%

response rate) was reached after the second round and the adjusted questionnaire can be found in Appendix 2.

DISCUSSION

Main findings

Just a small proportion of patients filled in the complete WAV-12 compared to other questionnaires at 6-weeks follow-up. A large number of participants only completed a limited number of items. This might indicate that the measurement instrument is not appropriate either in terms of language, setting, or participants had other specific reasons not to complete the questionnaire. The principal investigator also received comments from several patients and therapists, involved in the study, about items 3, 7 and 9 in the WAV-12 questionnaire. The construct theory of the WAV appeared to be sound but ceiling-effects were found in 10 items. Rewording was necessary for the WAV-12.

Comparison with the literature

Items correlated highly and measured the same explanatory concept which is found by several other translated versions of the WAI [13, 14, 16]. A French validation study found a very high correlation between the three subscales indicating that we cannot significantly distinct these subscales [16].

The poor correlation between patient and therapist WA score is consistent with other studies indicating that the two perspectives are not associated, which is confirmed by other studies as well [24, 25]. To ensure unbiased results the patient and the physiotherapist completed the rating forms independently of each other. Nevertheless, contact between the therapist and patient could not have been avoided, resulting in the possibility of deliberation between them.

WA was measured at 6 weeks when alliance might already have evolved into a stable situation; whereas, the first clinical experience between patient and therapist could determine more valid WA scores [26]. The literature is still inconsistent about what the optimal timing would be for measuring WA and some studies report that early WA predicted recovery after controlling for symptom change [27-30], while others have found a reduction of the predictive value of WA [31-33]. In this study WA was measured at six weeks as the first questionnaire was filled in before the first treatment. Nevertheless, we believe multiple measurements during the treatment period might yield more insight into the concept of WA.

Although WA is a valid construct within psychological interventions and research, whether it predicts recovery in a patient population in physiotherapy setting remains unknown. Psychological interventions are usually based on behavioural therapy that

physiotherapists mostly use in chronic patients. The patient population in this study has a new episode of shoulder pain where WA might be less relevant for the therapeutic process.

Strengths and limitations

This is the first study to perform a validation analysis on the Flemish version of the working alliance inventory in a physiotherapy setting. The measurement tool was able to discriminate between patients that experience a good or poor alliance. In ten items we observed ceiling effects, which might have been due to the fact that patients give socially desirable answers or that the items do not properly assess the total construct. There appeared to be a pattern in missing responses, where four items showed more missings than others, indicating that these might need adjustment. The questionnaire was developed in Belgium and applied in a Dutch setting which might not be appropriate given some linguistic characteristic differences of the Belgian Dutch (Flemish) and the Dutch language in the Netherlands. Due to the high number of missings in specific items (item 1, 3 and 9) and low discriminative values (item 1 and 2), we made changes in terms of adjustments in language and specific to the context of physiotherapy.

Implications for future research

The new questionnaire from our Delphi study has not been tested and therefore future research should test the psychometric properties of this questionnaire and evaluate the possible predictive value of the WA throughout the whole process of treatment in patients with musculoskeletal complaints. Whether measuring WA at the beginning or later in therapy is more predictive remains unknown. Studying a relationship between WA and recovery is complex because other factors, like self-adherence, compliance, might influence the relationship and therefore a mediation analysis might find more valid results.

Conclusions

The WAV-12 measurement tool is not suitable for implementation in clinical or research practice yet. However, WA is a concept that needs attention within the field of physiotherapy and therefore we made adjustments to the questionnaire. Previous research has shown a positive correlation between working alliance and recovery in physiotherapy setting. Since shoulder pain can become a chronic condition in more than 50% of patients, interventions from physiotherapy need to be effective and a good WA can possibly contribute to optimal treatment effects.

Funding

The authors report no declarations of interest. This study is financed by the SIA-RAAK grant. The ministry of education has made this funding available for the innovation and promotion of research. This study is partly funded by a program grant of the Dutch Arthritis Foundation.

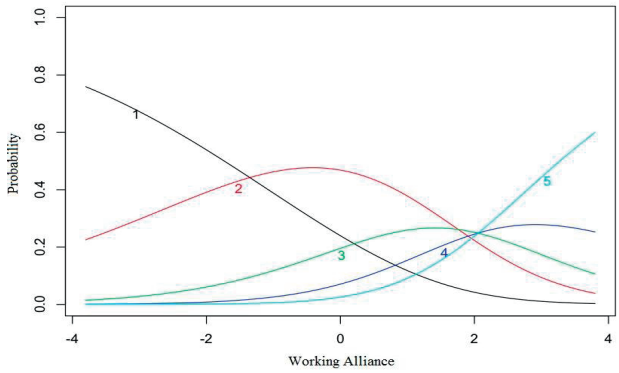
REFERENCES

1. Bordin, E., *The generalizability of the psychoanlaytic concepts of the working alliance.* Psychotherapy: Theory, Research and Practice 1979. **16**: p. 252-60.
2. Bennett, J.K., et al., *The role of patient attachment and working alliance on patient adherence, satisfaction, and health-related quality of life in lupus treatment.* Patient Educ Couns, 2011. **85**(1): p. 53-9.
3. Crow, R., et al., *The role of expectancies in the placebo effect and their use in the delivery of health care: a systematic review.* Health Technol Assess, 1999. **3**(3): p. 1-96.
4. Sluijs, E.M., G.J. Kok, and J. van der Zee, *Correlates of exercise compliance in physical therapy.* Phys Ther, 1993. **73**(11): p. 771-82; discussion 783-6.
5. Horvath, A. and D. Symonds, *Relation between working alliance and outcome in psychotherapy: A meta-analysis.* Journal of Counseling Psychology 1991. **38**: p. 39-49.
6. Martin, D.J., J. Garske, and K. Davis, *Relation of the therapeutic alliance with outcome and other variables: A meta-analytic review* Journal of Consulting and Clinical Psychology, 2000 **68**: p. 438-50.
7. Kaplan, S.H., S. Greenfield, and J.E. Ware, Jr., *Assessing the effects of physician-patient interactions on the outcomes of chronic disease.* Med Care, 1989. **27**(3 Suppl): p. S110-27.
8. Lee, Y.Y. and J.L. Lin, *The effects of trust in physician on self-efficacy, adherence and diabetes outcomes.* Soc Sci Med, 2009. **68**(6): p. 1060-8.
9. Hall, A.M., et al., *The influence of the therapist-patient relationship on treatment outcome in physical rehabilitation: a systematic review.* Phys Ther, 2010. **90**(8): p. 1099-110.
10. Ferreira, P.H., et al., *The therapeutic alliance between clinicians and patients predicts outcome in chronic low back pain.* Phys Ther, 2013. **93**(4): p. 470-8.
11. Siev, J., H. J., and D. Chambless, *The dodo bird, treatment technique, and disseminating empirically supported treatments.* The Behavior Therapist, 2009. **32**: p. 69-75.
12. Horvath, A.O. and L. Luborsky, *The role of the therapeutic alliance in psychotherapy.* J Consult Clin Psychol, 1993. **61**(4): p. 561-73.
13. Horvath, A. and L. Greenberg, *Development and validation of the Working Alliance Inventory.* Journal of Counseling Psychology 1989 **36**: p. 223-33.
14. Tracey, T. and A. Kokotovic, *Factor strructure of the working alliance inventory.* Psychological Assessment: A Journal of Consulting and Clinical Psychology 1989. **1**: p. 207-10.
15. Stinckens, N., A. Ulburghs, and L. Claes, *De werkalliantie als sleutelelement van de WAV-12, de Nederlandstalige verkorte versie van de Working Alliance Inventory.* Tijdschrift voor Klinische Psychologie 2009. **39**: p. 44-60.
16. Corbiere, M., et al., *Factorial validation of a French short-form of the Working Alliance Inventory.* Int J Methods Psychiatr Res, 2006. **15**(1): p. 36-45.
17. Schellingerhout, J.M., et al., *Measurement properties of translated versions of neck-specific questionnaires: a systematic review.* BMC Med Res Methodol, 2011. **11**: p. 87.
18. Karel, Y.H., et al., *Current management and prognostic factors in physiotherapy practice for patients with shoulder pain: design of a prospective cohort study.* BMC Musculoskelet Disord, 2013. **14**(1): p. 62.
19. Masters, G., *A Rasch model for partial credit scoring.* Psychometrika, 1982. **7**: p. 149-74.
20. Guttman, L., *The basis for Scalogram analysis*. 4th edn. (I. S. Prediction, Ed.) ed, ed. N.Y.W.T.A. Soldier. 1950.

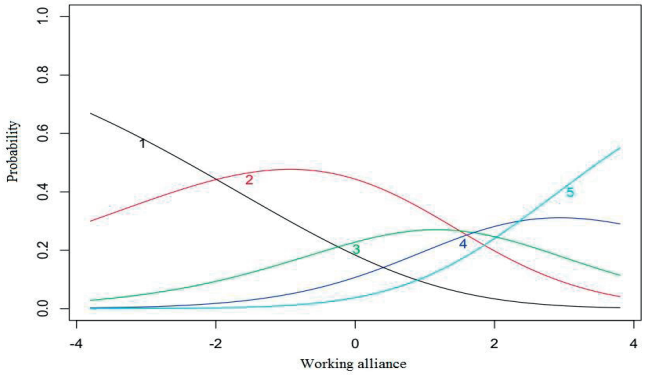
21. Tennant, A. and P.G. Conaghan, *The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper?* Arthritis Rheum, 2007. **57**(8): p. 1358-62.
22. Rizopoulos, D., *An R Package for Latent Variable Modeling and Item Response Theory Analyses.* Journal of Statistical Software 17, 2006.
23. Velozo, C., et al., *Maintaining instrument quality while reducing items: application of Rasch analysis to a self-report of visual function.* Journal of Outcome Measurement, 2001. **4**: p. 667-80.
24. Bachelor, A., *Clients' and therapists' views of the therapeutic alliance: similarities, differences and relationship to therapy outcome.* Clin Psychol Psychother, 2013. **20**(2): p. 118-35.
25. Huddy, V., et al., *The effect of working alliance on adherence and outcome in cognitive remediation therapy.* J Nerv Ment Dis, 2012. **200**(7): p. 614-9.
26. Horvath, A., *The Alliance.* Psychotherapy 2001. **38**: p. 365-372.
27. Barber, J.P., et al., *Alliance predicts patients' outcome beyond in-treatment change in symptoms.* J Consult Clin Psychol, 2000. **68**(6): p. 1027-32.
28. Anker, M.G., et al., *The alliance in couple therapy: Partner influence, early change, and alliance patterns in a naturalistic sample.* J Consult Clin Psychol, 2010. **78**(5): p. 635-45.
29. De Bolle, M., J.G. Johnson, and F. De Fruyt, *Patient and clinician perceptions of therapeutic alliance as predictors of improvement in depression.* Psychother Psychosom, 2010. **79**(6): p. 378-85.
30. Klein, D.N., et al., *Therapeutic alliance in depression treatment: controlling for prior change and patient characteristics.* J Consult Clin Psychol, 2003. **71**(6): p. 997-1006.
31. Barber, J., et al., *Therapeutic alliance as a predictor of outcome in treatment of cocaine dependance.* Psychotherapy Research 1999. **9**: p. 54-73.
32. DeRubeis, R. and M. Feeley, *Determinants of change in cognitive therpy for depression.* Cognitive Therapy and Research 1990. **14**: p. 469-482.
33. Puschner, B., M. Wolf, and S. Kraft, *Helping alliance and outcome in psychotherapy: What predicts what in routine outpatient treatment?* . Osychotherapy Research 2008. **18**: p. 167-178.

APPENDIX 1.

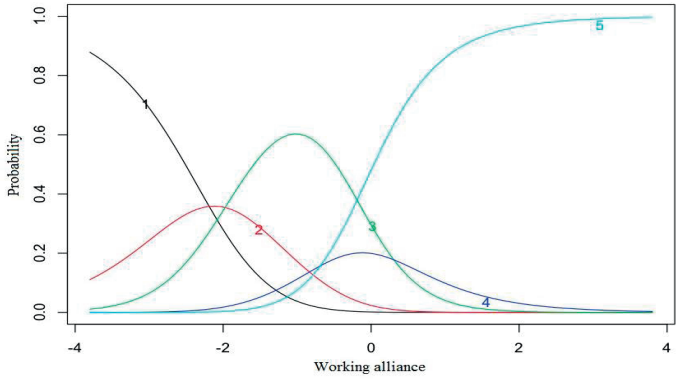
Item characteristic curves for the items in the WAV-12 questionnaire. Probability of working alliance score on the total construct for each response category of the item in different colours (1-5 likert scale).



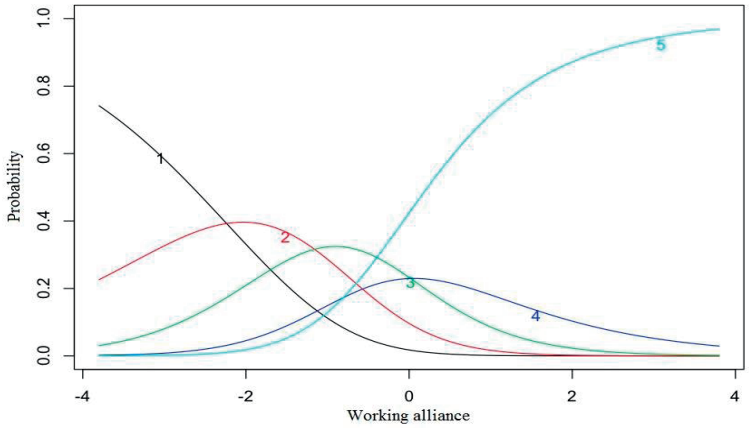
Item response category characteristic curve item 1



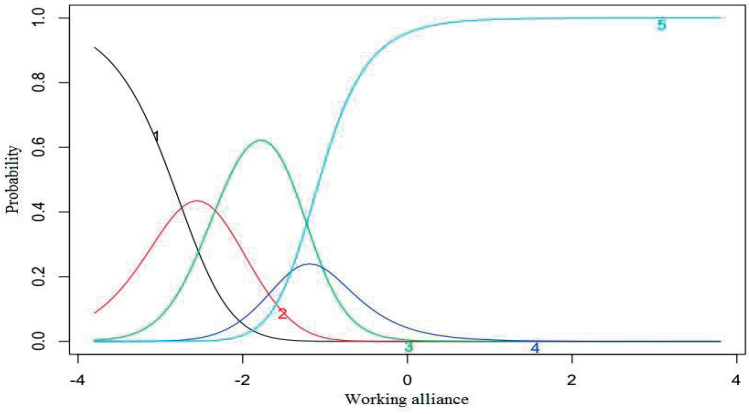
Item response category characteristic curve item 2



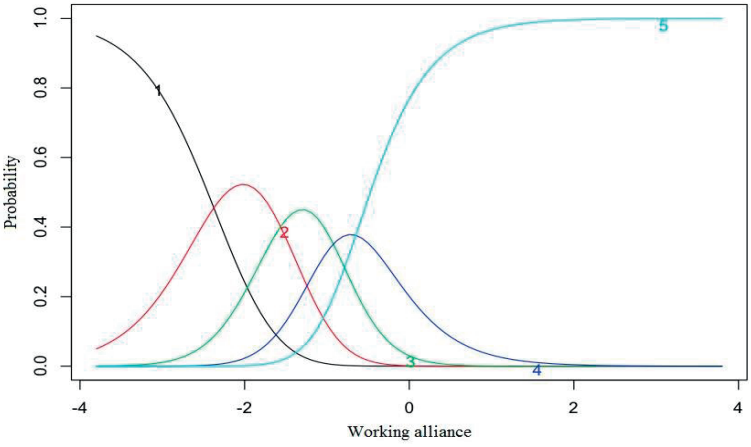
Item response category characteristic curve item 3



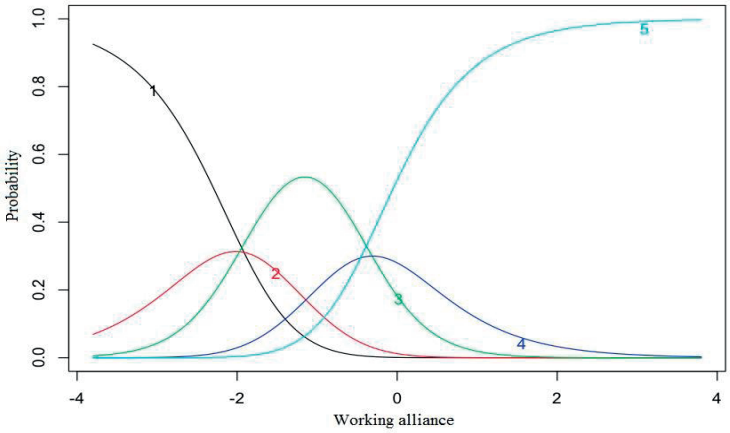
Item response category characteristic curve item 4



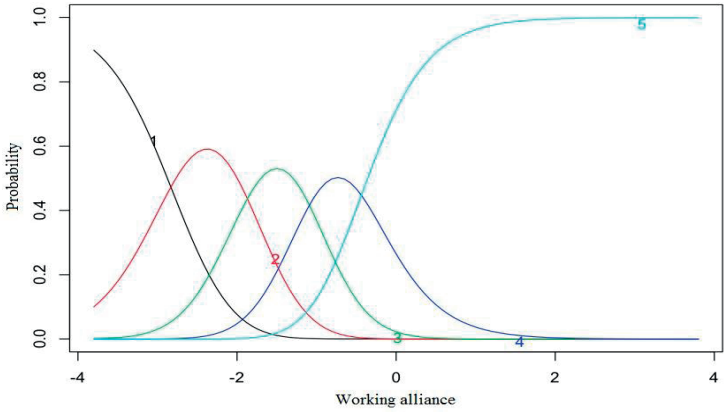
Item response category characteristic curve item 5



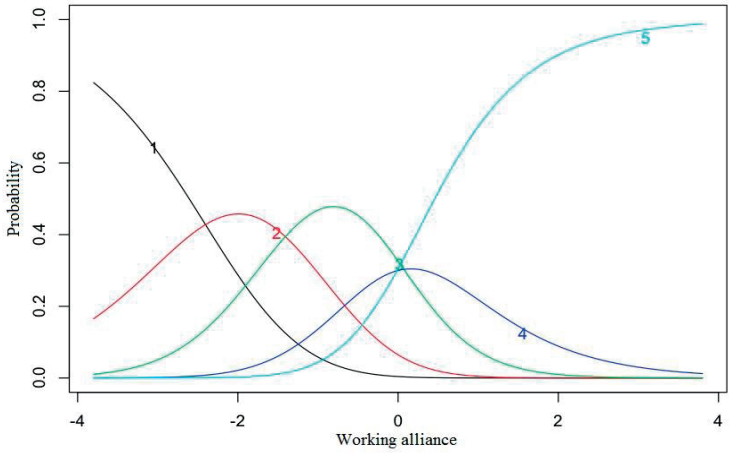
Item response category characteristic curve item 6



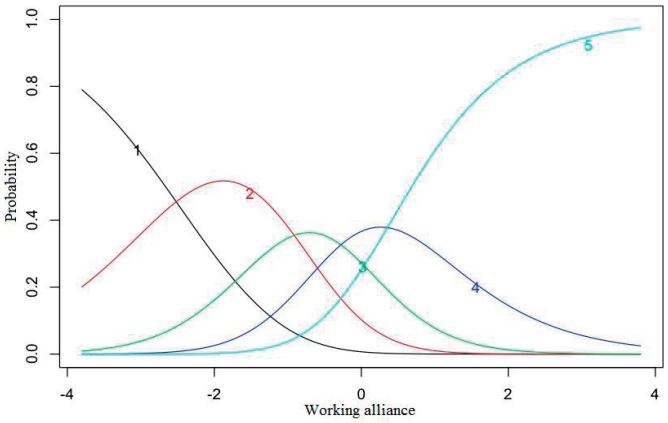
Item response category characteristic curve item 7



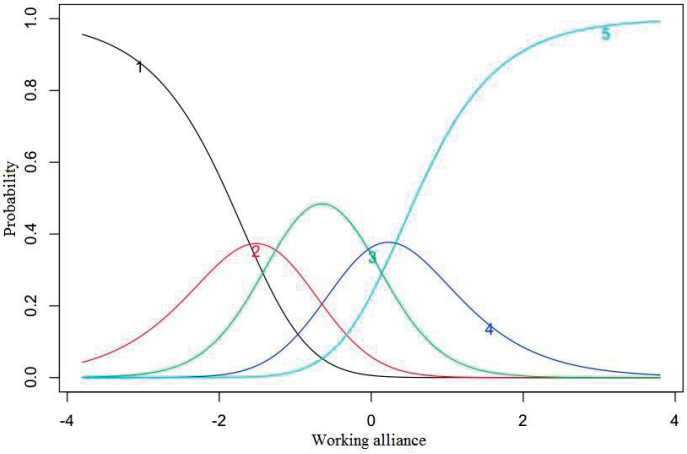
Item response category characteristic curve item 8



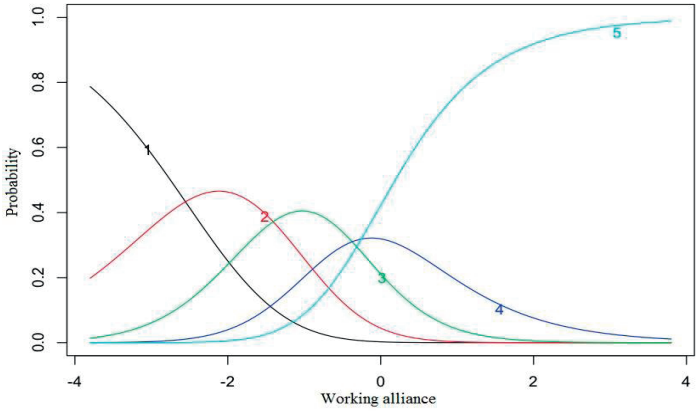
Item response category characteristic curve item 9



Item response category characteristic curve item 10



Item response category characteristic curve item 11



ategory characteristic curve item 12

APPENDIX 2

Physio Alliance Scale (PAS)

Naam .....
Datum: .....

Instructies:

Hieronder en op de volgende pagina worden een aantal omschrijvingen gegeven over de wijze waarop patiënten kunnen denken of voelen omtrent de relatie met de fysiotherapeut. Onder elke uitspraak bevinden zich vijf mogelijkheden om te antwoorden:
ZELDEN OF NOOIT / SOMS / VAAK / ZEER VAAK / ALTIJD
Indien de uitspraak aangeeft hoe u zich altijd voelt (of hoe u altijd denkt), omcirkelt u de antwoordmogelijkheid ALTIJD. Als ze nooit op u van toepassing is, omcirkelt u de antwoordmogelijkheid ZELDEN OF NOOIT. Gebruik de alternatieven tussenin om de variaties tussen deze extremen te beschrijven.
Geef een antwoord op alle uitspraken.

1. **Een resultaat van de therapie is dat ik weet hoe ik mijn klacht kan beïnvloeden.**
   ZELDEN OF NOOIT / SOMS / VAAK / ZEER VAAK / ALTIJD

2. **De therapie geeft mij een nieuwe kijk op mijn klacht.**
   ZELDEN OF NOOIT / SOMS / VAAK / ZEER VAAK / ALTIJD

3. **Ik geloof dat mijn fysiotherapeut(e) mij een prettig persoon vindt.**
   ZELDEN OF NOOIT / SOMS / VAAK / ZEER VAAK / ALTIJD

4. **De fysiotherapeut(e) betrekt mij bij het bepalen van de doelstellingen voor de therapie.**
   ZELDEN OF NOOIT / SOMS / VAAK / ZEER VAAK / ALTIJD

5. **Mijn fysiotherapeut(e) en ik respecteren elkaar .**
   ZELDEN OF NOOIT / SOMS / VAAK / ZEER VAAK / ALTIJD

6. **Mijn fysiotherapeut(e) en ik werken naar de doelstellingen toe waar we het beide over eens zijn.**
   ZELDEN OF NOOIT / SOMS / VAAK / ZEER VAAK / ALTIJD

7. **Ik heb het gevoel dat mijn fysiotherapeut(e) mij waardeert.**
   ZELDEN OF NOOIT / SOMS / VAAK / ZEER VAAK / ALTIJD

8. **Mijn fysiotherapeut(e) en ik zijn het erover eens wat voor mij belangrijk is om aan te werken.**
   ZELDEN OF NOOIT / SOMS / VAAK / ZEER VAAK / ALTIJD

9. **Ik heb het gevoel dat mijn fysiotherapeut(e) het beste met mij voor heeft, zelfs wanneer ik dingen doe waar hij/zij het niet mee eens is.**
   ZELDEN OF NOOIT / SOMS / VAAK / ZEER VAAK / ALTIJD

10. **Ik heb het gevoel dat de dingen die ik tijdens de therapie doe, mij zullen helpen om mijn doelen voor de therapie te bereiken.**
    ZELDEN OF NOOIT / SOMS / VAAK / ZEER VAAK / ALTIJD

11. **Mijn fysiotherapeut(e) en ik hebben dezelfde opvattingen over de veranderingen die goed zouden zijn voor mij.**
    ZELDEN OF NOOIT / SOMS / VAAK / ZEER VAAK / ALTIJD

12. **Ik geloof dat de manier waarop we aan mijn klacht werken, de juiste is.**
    ZELDEN OF NOOIT / SOMS / VAAK / ZEER VAAK / ALTIJD

Kijk of u alle antwoorden heeft ingevuld.

Hartelijk bedankt voor uw medewerking!

Thoomes-de Graaf M., Scholten-Peeters G.G., Duijn E., Karel Y.H., de Vet H.C.W., Koes B.W. , Verhagen A.P.

J Orthop Sports Phys Ther. 2017 Apr;47(4):278-286. doi: 10.2519/jospt.2017.7079. [Epub 2017 Feb 3].

THE RESPONSIVENESS AND INTERPRETABILITY OF THE SHOULDER PAIN AND DISABILITY INDEX

CHAPTER 7

ABSTRACT

Study Design: Clinical measurement study; prospective cohort design.

Background: Shoulder pain is a common disorder and treatment is most often focused on a reduction of pain and functional disabilities. Several reviews have encouraged the use of the Shoulder Pain and Disability Index (SPADI) to objectify functional disabilities. It is important to assess the responsiveness and interpretability of the SPADI when it is used by patients seeking help by a physiotherapist for their shoulder pain in primary care setting.

Objective: To assess the responsiveness and interpretability of the SPADI in patients with shoulder pain visiting a physiotherapist in primary care.

Methods: The target population consisted of patients consulting a physiotherapist for their shoulder pain. Patients received physiotherapy treatment completed the Dutch language version of the SPADI (SPADI-D) at baseline and the follow up of 26 weeks. To assess the interpretability of the assessing floor and ceiling effects and by calculating the minimal important change (MIC) using the ROC method including a visual anchor-based MIC distribution for several Global Perceived Effect scale (GPE) based anchors. The measurement error was calculated using the Smallest Detectable Change (SDC). For the responsiveness, the Area under the ROC curve (AUC) was used and correlations with the GPE and the change score of the Shoulder Disability Questionnaire (as this questionnaire measures the same construct) were assessed.

Results: In total 356 patients participated at baseline and 237 (67%) returned the SPADI after 26 weeks. The mean score at baseline of the SPADI was 46.7 points (on a 0-100 scale). The SPADI showed no signs of floor and ceiling effects. The SDC was 19.7 points. The MIC was 20 (43% of baseline value) and therefore a change of 43% or more in an individual patient was considered to be clinically relevant. The AUC was 0.81, the Spearman correlation between the SPADI change score and the GPE was 0.53 and the Pearson correlation between the SDQ and the SPADI change scores was 0.71.

Conclusion: The results of this study confirm the responsiveness of the SPADI, making it a useful instrument to assess functional disability in longitudinal studies; however, the measurement error should be taken into account when making decisions in individual patients.

Keywords: SPADI, responsiveness, shoulder, measurement error

INTRODUCTION

Shoulder pain is a common disorder in western society [1]. The point prevalence ranges from 7 to 27% [2], making it the second most reported musculoskeletal complaint in general practice [3]. Apart from pain, one of the main complaints of patients with shoulder pain is functional disability. Thus, treatment of shoulder disorders is usually aimed at reducing pain and functional disabilities [4].

Self-administered shoulder pain and disability questionnaires are designed to measure functional disability. These patient-reported outcome measures are often used in both clinical and research environments, to assess patient's perceived levels of disability and the impact of the disease on daily activities [5] and to evaluate functional status [4].

Several reviews have encouraged the use of the Shoulder Pain and Disability Index (SPADI) [6-9]. The SPADI is a disease specific instrument and is frequently used in primary care. The SPADI is easy to complete, convenient to use and is not time consuming to fill out [10]. It has been translated and validated (using hypothesis testing) into Danish, Norwegian, Tamil, German, Turkish and Slovene [11-16]. The Dutch SPADI (SPADI-D) has been recently validated (using hypothesis testing for known-group validity (high initial pain and work absence), divergent validity (Shoulder Disability Questionnaire (SDQ)), has shown to be reliable [17] and has been recommended in an evidence-based statement on shoulder pain, by the Royal Dutch physiotherapy association (KNGF) [18]. The responsiveness and interpretability of the SPADI-D has not been assessed before.

A systematic review showed there is moderate positive evidence for responsiveness of the English and Norwegian versions of the SPADI [9]. There was a variety in both setting and included patients that were part of the three primary studies included in this review, to assess the responsiveness of the SPADI. Only one study was performed in a physiotherapy setting with participating patients diagnosed with adhesive capsulitis [19]. Both other studies were performed in different settings, a general practitioner setting (patients with rotator cuff disease) [16] and a shoulder pain clinic setting (patients with mixed 'shoulder pain') [10]

It is important to assess the responsiveness and interpretability of the SPADI (-D) when it is used by patients seek treatment by a physiotherapist for their shoulder pain in primary care setting.

In the literature, interpretability is defined as: the degree to which one can assign qualitative meaning to an instrument's quantitative scores or changes in scores [20]. Therefore information about floor and ceiling effects and the minimal important change (MIC) should be provided [21]. The MIC is the smallest change in the score of an instrument that patients perceive as important [22]. The measurement error is the systematic

and random error of a patient's score that is not attributed to true changes in the construct to be measured [20]. Preferably, the measurement error should be smaller than the MIC [21, 23]. However, this is often not the case, which can be a consequence of the use of different mediators and calculations. Responsiveness is defined as: the ability of an instrument to detect changes over time in the construct to be measured [20].

Therefore, the aim of this study was to evaluate the measurement error, interpretability and responsiveness of the SPADI-D on patients with shoulder pain seeking help by a physiotherapist in primary care setting. We used the Global Perceived Effect (GPE)-scale as external criteria for improvement. To assess responsiveness, we hypothesized that the change score of the SPADI-D was highly correlated with a shoulder-specific instrument (the Shoulder Disability Questionnaire (SDQ)) and with the GPE-scale. A lower correlation was expected with a questionnaire with a different focus (EuroQol five-item quality of life questionnaire (EQ-5D-3L)).

METHODS

Design

This study is part of a prospective cohort study, including patients with shoulder complaints in primary care physiotherapy setting. Details of the design are presented elsewhere [24]. The Medical Ethics Committee of the Erasmus Medical Center in Rotterdam approved the study protocol (MEC-2011-414). All participants signed informed consent.

Study population

Patients were recruited from primary care physiotherapy clinics between November 2011 and December 2012. Patients with shoulder pain were eligible for inclusion if they were 18 years or over and adequately understood the Dutch language. Patients were excluded in the presence of serious pathology (infection, cancer of fracture), previous surgery or diagnostic imaging techniques of the shoulder, in the previous 3 months.

Therapists

Physiotherapists collected data at baseline and after 12 weeks on what kind of diagnostic label was used on patients, what type of treatment was used, and how many treatment sessions were given within the time frame.

Baseline measurement

Patients received a baseline assessment followed by usual physiotherapy care. Participating patients received an online questionnaire that included the SDQ, SPADI and the

EQ-5D-3L, all in Dutch. All three questionnaires have been reported to take approximately 3 minutes to complete [10, 25-27].

SPADI-D

The SPADI is a self-administered questionnaire designed to measure pain and disability associated with shoulder pain. It consists of 13 items (5 pain related items and 8 disability related items) [28]. However, factor analysis of the SPADI-D did not confirm the original factor structure and is based on one factor only [17]. Each question refers to the past week. Items can be scored on a visual analogue scale, ranging from 0 to 10, where 0 represents "no pain/no difficulty" and 10 "worst pain imaginable/so difficult it requires help" [10, 28]. The total score varies between 0 and 100, a higher score indicates a higher level of pain related disability [28].

SDQ

The SDQ is a pain-related disability questionnaire, which consists of 16 items. All items refer to pain related disability in the preceding 24 hours. Response options are "yes", "no" or "not applicable". The option "not applicable" indicates that the situation at issue has not occurred in the past 24 hours. The SDQ-score ranges from 0 to 100 with a higher score indicating more severe disability [4, 13]. The SDQ was originally designed and validated in Dutch [29, 30]. The SDQ shows acceptable content, divergent and construct validity [29] and is a responsive instrument [10, 31][30].

EQ-5D-3L

The EQ-5D-3L is a quality of life questionnaire covering 5 dimensions of health: mobility, self-care, usual activities, pain/discomfort and anxiety/depression [26, 32]. Each dimension has 3 levels (answer categories): no problems, some problems, extreme problems. Besides these five items, perceived health state is measured, using a scale from 0-100, with higher scores indicating better health status. The EQ-5D-3L has been used frequently, most often as part of cost-effectiveness studies [26, 33, 34]. The Dutch EQ-5D-3L is an official language version and has been validated [35].

Test-retest measurement

A randomly selected group of patients received a second SPADI-D after one week. The time interval was chosen to minimize recall bias as well as progression bias and is often considered appropriate [36]. A sample size of approximately 80 is considered acceptable [21]. The data collected from this test-retest measurement were used in a previously published study as well, in order to assess the reliability [17].

Follow up measurements

All patients received the SPADI-D, SDQ and the GPE-scale 26 weeks after initial presentation. Within this period the patient received physiotherapy treatment for one or more sessions.

GPE-scale

The GPE-scale is a 7-point Likert scale scoring whether the patient's condition has improved or deteriorated since their start of physiotherapy treatment ("Could you please state the amount of change concerning your recovery compared to when you first started treatment?")

The GPE-scale ranges from "worse than ever" to "completely recovered" (completely recovered, much improved, slightly improved, no change, slightly worse, much worse and worse than ever). The GPE-scale has good test-retest reliability and correlates well with changes in pain and disability.[37] Despite controversy about the role of global rating items, the GPE scale has frequently been used as an anchor and responsiveness studies [38-42]. All forms were available online, using LimeSurvey software (https://www.limesurvey.org/).

Analysis

All statistical analyses were performed with SPSS version 23 (IBM corporation, Armonk, NY). Regarding missing items, as described by the original authors [28, 29], patients were excluded from the analysis if there were more than two items missing per SPADI-subscale [28] or from the SDQ [29]. The total score for the included patients was calculated by adding up the item scores and dividing them only with the items that were deemed applicable to the subject [28, 29].

All data were checked on normality, using a Stem-and-leaf Plot, Q-Plot and whisker box. Nonparametric tests were used if data was not normally distributed. Descriptive statistics were used to calculate frequencies.

Interpretability

The distribution of scores in the patient population, floor and ceiling effects and interpretation of change scores are part of interpretability. Frequencies were presented as means and standard deviations (SD) for data that were normally distributed and interquartile range data for data that were not normally distributed.

If at baseline or at the 26-week follow-up more than 15% of the respondents achieved the highest or lowest possible scores, then we concluded that there were signs of floor or ceiling effects [22].

We calculated the amount of change between the baseline score and the SPADI-D score after 26 weeks, using the mean change and the SD per category of the GPE and

of all anchors. This provides information on how a change score on the SPADI-D corresponds to the magnitude of change, as perceived by patients.

The interpretation of change scores included calculating the minimal important change (MIC), which is the smallest change in score in the construct to be measured that patients perceive as important. We used the receiver operating characteristic (ROC) method, with the SPADI-D as the diagnostic test and the anchor (GPE scale) as the gold standard for calculating the MIC. The anchor distinguishes patients considered 'recovered' from patients who were 'not importantly changed'. The instrument's sensitivity is the proportion of 'recovered' patients according to the anchor that are correctly identified as such by the SPADI-D. Specificity is the proportion of patients with 'no important change' that is correctly identified as such by the SPADI-D. The MIC is defined as the optimal ROC cut-off point, which is the point on the ROC curve nearest to the upper left-hand corner [22].

On the GPE scale, a frequently used anchor, we considered patients to be recovered when they answered they were 'completely recovered' or 'much improved' and to be not importantly improved when they answered 'slightly improved', 'no change' or 'slightly worse' were classified [38, 39, 43].

We also created a visual anchor-based MIC distribution, which shows how well an instrument is able to distinguish between patients who are importantly improved from those that are not importantly changed [44].

The MIC can be influenced by the baseline score of patients (low or high); a percentage of the baseline score is more stable [45]. Therefore, we performed a subgroup analysis to assess the difference in MIC values with high and low baseline SPADI values (mean split).

As some researchers also included patients who were 'slightly improved' as importantly changed, making the MIC lower by definition, we also presented the MIC based on this anchor [43, 46].

The measurement error

The measurement error can be adequately expressed as the standard error of measurement (SEM). For this analysis we used the data of the test-retest set. The group of patients has been described in an earlier published study [17] and we therefore were aware of two extreme values in the test-retest data [17]. We, excluded these two extreme values to calculate the measurement error, but presented the results based on data including these extreme values as well, to assess its influence. We used the test-retest data to test whether or not there were systematic errors, using an analysis of variance. When there were no systematic errors, the ICC consistency was used to calculate the smallest error of measurement (SEM) and in all other cases the ICC agreement was used. The SEM was calculated as SD * $\sqrt{}$(1-ICC) [23] and the Smallest Detectable Change (SDC) was calculated as 1.96* $\sqrt{}$2* SEM [23] to assess the change beyond measurement error. We

presented a Bland and Altman plot to visually illustrate systematic errors. Ideally the MIC should be higher than the SDC [47].

Responsiveness

Responsiveness was assessed using the area under the ROC curve (AUC) and hypothesis testing. As the GPE has a high level of face validity and is considered to be a suitable criterion to measure change, we were able to use the AUC method [22]. However, doubt has been expressed about the reliability and validity of such measures of change [48] and we therefore chose to test specific hypotheses as well.

We calculated the AUC to assess the ability of the SPADI-D to discriminate between patients who are considered improved and not importantly changed according to the GPE, using an anchor similar to that described in the interpretability section [22]. A benchmark that has been previously used to establish that outcome measures are useful in discriminating improved and unimproved patients has been set at 0.70 AUC [23].

Hypothesis testing for responsiveness was based on the concept that the correlation between the change score of related constructs (GPE scale and SDQ) must be higher than with unrelated constructs (the depression and mobility items of the EQ-5D-3L). Hypothesis testing was quantified by the Pearson correlation coefficient in case of a normal distribution and by a Spearman correlation coefficient for all other distributions. Correlation coefficients between the SPADI-D change score and the change score of the SDQ and the GPE were expected to be above 0.50 and the EQ-5D-3L mobility and depression items were expected to be lower than 0.20 [22].

RESULTS

A total of 356 patients participated at baseline, 114 of whom did not return the SPADI-D follow up assessment at 26 weeks. In total, 242 patients returned the SPADI-D, of whom five were excluded due to the missing item criteria, resulting in 237 patients included in the analysis (66.6% of the baseline population). Some (22%) of the patients who were included in the test-retest measurement were not included in the responsiveness cohort, as they did not return the SPADI at 26 weeks or had missing item criteria. The mean age of the total baseline population was 49.5 (SD 13.1) years and 47% was male.

The physiotherapists used a variety of shoulder diagnoses to label the patients; however, the majority of patients were labelled as having subacromial impingement. The physiotherapists also used a variety of treatment techniques, mainly including advice, exercise, and mobilization/manipulation of the shoulder or thoracic spine. After 12 weeks, the majority of patients (59.5%) stopped therapy. Overall, the median number

of treatment sessions was six. The characteristics of the participants are presented in Table 1.

TABLE 1. Baseline characteristics of the participants per analysis

Population	Total cohort (n=356)	Follow up cohort (n=237)	Test-retest cohort complete (n=74)	Test-retest cohort without extreme values (n=72)
Gender (male) (%)	166 (47%)	109 (46%)	29 (39%)	29 (40%)
Age (SD)	49.5 (13.1)	50.0 (12.9)	51.4 (12.7)	51.5 (12.9)
SPADI-D score mean (SD)	46.7 (21.3)	47.0 (21.5)	50.8 (22.6)	50.2 (22.6)
Use of medication (%)	171 (48%)	117 (49%)	37 (50%)	37 (51%)
Pain intensity (NRS) median (IQR)	6 (4-7)	6 (4-7)	6 (4-8)	6 (4-8)
Number of treatment sessions of patients that stopped therapy within 12 weeks Median (IQR)*	-	6 (4-9)	-	-

Abbreviations: SD; Standard deviation, NRS; Numeric Rating Scale, IQR; Inter quartile range
* A total of 141 patients 59.5% stopped therapy sessions after 12 weeks.

The data of the SPADI-D at baseline and the change scores of both the SPADI-D and SDQ were considered to be normally distributed, in contrast to those of the EQ-5D-3L.

Interpretability

The mean score of the SPADI-D at baseline of the total population with shoulder pain was 46.7 (SD 21.3), and at 26 weeks 23.9 points (SD 24.2).-

At baseline, only one patient had a SPADI-D score of zero and none of the patients showed a score of 100; the highest score was 92 (0.3% of all patients). About 8.1% (n=29) of the patients scored in the lower part of the range of the scale (a score between 0 and 15), and only 2.2% (n=8) of the patients scored in the upper part of the range of the scale (between 85 and 100). After 6 months, 13.5%(n=32) of the patients had a score of zero and none (0%) had a score of 100; the highest score was 89 (0.4%). We therefore concluded there were no signs of floor and ceiling effects.

Table 2 shows the mean change per category on the GPE-scale. A total of 139 patients were considered recovered (with a change score between baseline and 26 weeks of -33.4, SD 19.5) and 95 as not importantly changed (with a change score between baseline and 26 weeks of -8.9, SD 21.4). The MIC was 20 points, resulting in a change of 42.8% of the baseline score. The sensitivity and specificity were both 0.75. Subgroup analysis resulted in similar results, the MIC for patients with a high baseline score was 43.0% (27.9 points),

with a sensitivity of 0.82 and specificity of 0.77 and for patients with a low baseline score 42.7% (12.2 points), with a sensitivity of 0.81 and specificity of 0.82.

TABLE 2. Mean change per category on the GPE-scale.

	Total		High baseline score		Low baseline score	
GPE scale	**Number of patients (N=237)**	**Mean change between SPADI-D baseline and after 26 weeks (SD)**	**Number of patients (N=120)**	**Mean change between SPADI-D baseline and after 26 weeks (SD)**	**Number of patients (N=117)**	**Mean change between SPADI-D baseline and after 26 weeks (SD)**
1. Completely recovered	43	-36.5 (22.1)	14	-61.4 (13.8)	29	-24.4 (13.4)
2. Much improved	96	-32.0 (18.1)	48	-42.2 (16.8)	48	-21.7 (12.8)
3. Slightly improved	61	-12.5 (21.5)	36	-21.2 (20.0)	25	-0.04 (17.4)
4. No change	28	-2.0 (16.5)	17	-7.7 (13.8)	11	6.8 (17.2)
5. Slightly worse	6	-4.4 (34.0)	3	-25.5 (37.3)	3	16.8 (12.6)
6. Much worse	3	7.1 (11.7)	2	4.0 (14.7)	1	13.11
7. Worse than ever	0		0			

Abbreviations: SD; Standard deviation

The visual anchor-based MIC distribution is presented in Figure 1. It shows the SPADI-D is capable in discriminating between patients that are importantly improved versus those who are not importantly changed.

For the alternative anchor, on which 'slightly improved' was considered to be importantly changed the MIC was 16 points.

Measurement error

The two patients with extreme values showed a change score between baseline and retest (≤ 7 days after baseline) of -31 and -30 points, respectively. These patients were no longer under physiotherapy treatment after 3 weeks and felt completely recovered after 6 weeks. The ANOVA analysis revealed there were systematic errors. With the outliers included, the mean difference was -4.1 (SD 10.7) between baseline and retest (50.8 versus 46.7). After exclusion of the two extreme values, the mean difference was -3.4 (SD 9.9) (50.2 versus 46.8). Figure 2 shows the Bland and Altman illustrating the systematic bias. The SEM was 7.1 and the SDC was 19.7.

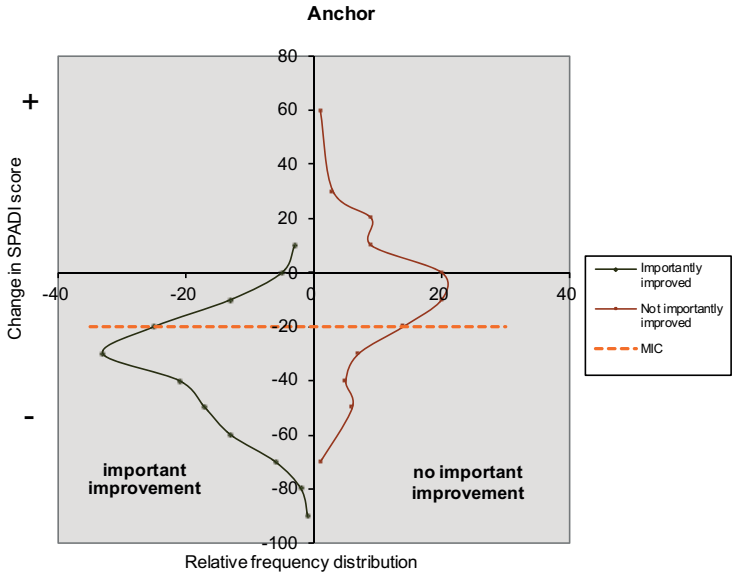


FIGURE 1. Visual anchor based MIC distribution
Abbreviations: MIC; Minimal Important Change.
Distribution of change scores on the SPADI-D of patients who reported an important improvement (n=139) compared with those with no important change (n=95) on the first anchor (GPE). The left quadrant above the line represents the misclassified patients that felt importantly improved but were not classified as such by their SPADI-D change score (23.7%). In the lower right quadrant, beneath the orange line are the patients that were misclassified, as they considered themselves as not importantly improved, but according to their SPADI-D change score they were classified as importantly improved (25.3%).

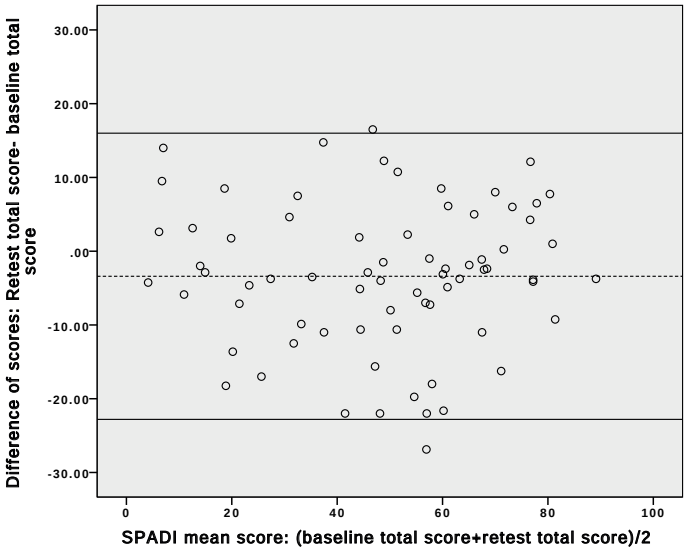


FIGURE 2. Bland and Altman plot, illustrating the mean difference between two measurements and the limits of agreement.

Responsiveness

The AUC was 0.81 with a 95% confidence interval ranging from 0.75 to 0.87. Figure 3 shows the ROC curve.

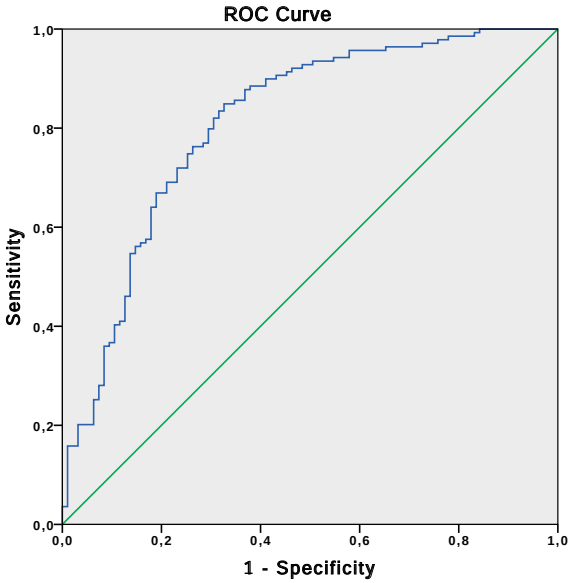


FIGURE 3. ROC curve.
Abbreviations: ROC; Receiver Operator Curve, AUC; Area Under the Curve
ROC curve based on anchor 1, resulting in an AUC of 0.81.

Hypothesis testing for responsiveness resulted in a Spearman correlation between the SPADI-D change score and the GPE-scale of 0.53. The Pearson correlation between the SPADI-D change score and the SDQ change score was 0.71. The Spearman correlation between the change score of the SPADI-D and the change score of the EQ-5D-3L depression was 0.06 and with mobility item 0.12. Based on the AUC values and with all hypotheses confirmed, we consider the SPADI-D to be a responsive measurement instrument.

DISCUSSION

This study shows that the SPADI-D is responsive, making it a useful evaluative instrument to assess functional disability in longitudinal studies in patients with shoulder pain visiting a physiotherapist. The SPADI-D can detect important changes. A change larger than 43% of the baseline score is considered to be a clinically relevant and important

change. However, the measurement error should be taken into account when used for decision-making in individual patients.

Comparison to the literature

Interpretability

Our study showed no signs of floor and ceiling effects, this was similar to earlier research [15, 49].

The MIC in our study was 20. One other study reported a MIC of 20.3 based on the ROC method, also using a quite similar global perceived effect scale (an 18-point Likert scale) as an anchor, with a 'similar' choice in dividing patients as 'recovered' and 'not importantly changed' [49]. That study population consisted of patients with rotator cuff disorders who were referred by their general practitioner to the Physical Medicine and Rehabilitation Department of a hospital. The patients in our study were comparable with in age, gender and work absence to those of the previous study; the baseline SPADI score in the previous study was approximately 5 points higher than that in our study population [49]. One study used a study population with upper extremity disorders, and calculated the MIC using mean change scores for patients with small but meaningful global change on a global disability rating scale they developed, resulting in a MIC of 13 points [50]. However, none of the above studies assessed whether or not the MIC varied between high or low baseline score.

Measurement error

Only a small number of studies assessed the measurement error of the SPADI [11, 15, 49, 51]. One study reported a SEM of 7.0 (95%CI 6.0-8.5) and an SDC of 19.4 [15]. The sample of that study showed a higher level of pain-related disability, as the SPADI baseline score was approximately 7 points higher than that of our study [15]. Another study reported an SDC of 19.7 [49] and one study reported a smallest detectable difference of 17 points [51]. A study using a different study population (patients that had undergone total shoulder or hemi-arthroplasty), reported an SDC of 18 points [11]. All these SDC values are comparable to those of our study, when the results of the analysis that excluded outliers were used. We feel the most appropriate analysis is the one that excluded the outliers, resulting in an SDC of 19.7. However, the analysis with the outliers included, resulted in an SDC of 22.5. The MIC was higher than the SDC when the outliers were excluded.

Responsiveness

The AUC in our study (0.81) was comparable with that of other studies (range, 0.80-0.92), despite using different GPE-scales (5-point and an 18-point Likert scale) [10, 19, 49].

The Spearman-correlation with the GPE-scale found in our study, was comparable with a previous study [10]. No other studies used the SDQ change score as a comparator, although the construct of this questionnaire is comparable with the SPADI. One study used correlations between the SPADI and other pain-related disability questionnaires (CROFT index; Disabilities of the Arm, Shoulder and Hand Questionnaire (DASH); Problem Elicitation Technique (PET), and Health Assessment Questionnaire (HAQ)) and perceived improvement, and they were all above 0.49, except for the HAQ [19]. Range of motion was also used as a comparator [13, 28, 51]; however, we feel this measures a different construct and is therefore not appropriate.

Strength and limitations

This study has some limitations. We did not use the GPE-scale to check whether patients were indeed stable within 7 days between the test and re-test, which could have influenced the measurement error. However, the 7-day time frame we used is commonly accepted [36]. Moreover, the median duration of shoulder pain at the start of inclusion was 16 weeks in our study population. Physiotherapists usually treat patients with shoulder pain for about 11 weeks (SD 11.3) [52]. It is therefore unlikely that patients would have been recovered within one week. We checked data for patients with extreme change scores, as there is always the chance that a patient's condition will improve or worsen within this time frame. There was a systematic error, with a mean difference of -3.4 points between test and retest, suggesting a very small and minimal improvement. The two patients with extreme values were no longer under treatment after three weeks, and it is therefore likely that these patients are an exception and have indeed changed substantially. We reported the results for both the population with extreme values and without extreme values, so clinicians can take this into consideration.

One of the strengths of this study is that our population consisted of patients visiting a physiotherapist, as the SPADI is frequently used by physiotherapists and pain/activity limitations are important outcome measures, thus it is important to assess the measurement properties in this study population. Moreover, this study consists of a relatively large sample size. Another strength of this study is that we assessed whether the MIC would vary over different parts of the complete range of SPADI scores (e.g. High versus low baseline SPADI scores). This is important for clinical as well as research purposes, as it reflects that when symptoms are severe they can change more dramatically (in absolute terms) to be of importance to patients than when patients have a lower baseline score.

Implications for clinical practice

Patients with a change score 43% or more of their baseline SPADI-D score considered themselves to be importantly improved; therefore, a change score of 43% in individual patients may be regarded as clinically relevant. A change score of less than 20 points

could be due to measurement error. An example for clinicians: if a patient had a baseline SPADI-D score of 50 and SPADI-D score of 20 at follow-up, one could consider this to be a real change, as it is greater than the measurement error and ~~as~~ clinically relevant, ~~as~~ the change score being greater than the MIC (43%). However, when a patient has a baseline of 35 points and scores 20 points at follow-up, this could be considered ~~as~~ clinically relevant, as it is a change above 43%, but this change could still be a measurement error. A change score of 15 points is beneath 19.7 and could be due to measurement error. Clinicians have to take the measurement error into account when they use the SPADI-D for evaluative purposes in individual patients.

The present study found the SPADI-D to be a responsive instrument for assessing patients who seek physiotherapy care for shoulder pain and functional disability. The SPADI-D was able to detect changes larger than 43% of the baseline score, which is considered to be clinically relevant and important change. However, when making decisions based on SPADI-D scores in individual patients, measurement error must be taken into account.

REFERENCES

1. Feleus, A., et al., *Management in non-traumatic arm, neck and shoulder complaints: differences between diagnostic groups.* Eur Spine J, 2008. **17**(9): p. 1218-29.
2. Luime, J.J., et al., *Prevalence and incidence of shoulder pain in the general population; a systematic review.* Scand J Rheumatol, 2004. **33**(2): p. 73-81.
3. Picavet, H.S. and J.S. Schouten, *Musculoskeletal pain in the Netherlands: prevalences, consequences and risk groups, the DMC(3)-study.* Pain, 2003. **102**(1-2): p. 167-78.
4. van der Windt, D.A., et al., *The responsiveness of the Shoulder Disability Questionnaire.* Ann Rheum Dis, 1998. **57**(2): p. 82-7.
5. Mintken, P.E., P. Glynn, and J.A. Cleland, *Psychometric properties of the shortened disabilities of the Arm, Shoulder, and Hand Questionnaire (QuickDASH) and Numeric Pain Rating Scale in patients with shoulder pain.* J Shoulder Elbow Surg, 2009. **18**(6): p. 920-6.
6. Bot, S.D., et al., *Clinimetric evaluation of shoulder disability questionnaires: a systematic review of the literature.* Ann Rheum Dis, 2004. **63**(4): p. 335-41.
7. Roy, J.S., J.C. MacDermid, and L.J. Woodhouse, *Measuring shoulder function: a systematic review of four questionnaires.* Arthritis Rheum, 2009. **61**(5): p. 623-32.
8. Breckenridge, J.D. and J.H. McAuley, *Shoulder Pain and Disability Index (SPADI).* J Physiother, 2011. **57**(3): p. 197.
9. Thoomes-de Graaf, M.S.-P., G.G.M. ; Schellingerhout, J.M. ; Bourne, A.M.; Buchbinder, R.; Koehorst, M.; Terwee, C.B.; and A.P. Verhagen, *Evaluation of measurement properties of self-administered PROMs aimed at patients with non-specific shoulder pain and "activity limitations": a systematic review.* Qual Life Res, 2016.
10. Paul, A., et al., *A comparison of four shoulder-specific questionnaires in primary care.* Ann Rheum Dis, 2004. **63**(10): p. 1293-9.
11. Angst, F., et al., *Cross-cultural adaptation, reliability and validity of the German Shoulder Pain and Disability Index (SPADI).* Rheumatology (Oxford), 2007. **46**(1): p. 87-92.
12. Bicer, A. and H. Ankarali, *Shoulder Pain and Disability Index: A validation study in Turkish women.* Singapore Med J, 2010. **51**(11): p. 865-70.
13. Jamnik, H. and M.K. Spevak, *Shoulder pain and disability Index: Validation of slovene version.* Int J Rehabil Res, 2008. **31**(4): p. 337-41.
14. Jeldi, A.J., et al., *Cross-cultural adaptation, reliability and validity of the Indian (Tamil) version of the Shoulder Pain and Disability Index.* Hong Kong Physiotherapy Journal, 2012.
15. Christiansen, D.H., J.H. Andersen, and J.P. Haahr, *Cross-cultural adaption and measurement properties of the Danish version of the Shoulder Pain and Disability Index.* Clin Rehabil, 2013. **27**(4): p. 355-60.
16. Ekeberg, O.M., et al., *Agreement, reliability and validity in 3 shoulder questionnaires in patients with rotator cuff disease.* BMC Musculoskelet Disord, 2008. **9**: p. 68.
17. Thoomes-de Graaf, M., et al., *The Dutch Shoulder Pain and Disability Index (SPADI): a reliability and validation study.* Qual Life Res, 2014.
18. Jansen, M.J., et al., *KNGF Evidence Statement Subacromiale klachten.* Nederlands Tijdschrift voor Fysiotherapie, 2011. **121**(1).
19. Staples, M.P., et al., *Shoulder-specific disability measures showed acceptable construct validity and responsiveness.* J Clin Epidemiol, 2010. **63**(2): p. 163-70.

20. Mokkink, L.B., et al., *The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes.* J Clin Epidemiol, 2010. **63**(7): p. 737-45.
21. Mokkink, L.B., et al., *The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content.* BMC Med Res Methodol, 2010. **10**: p. 22.
22. de Vet HC, T.C., Mokkink LB, Knol DL, *Practical guides to biostatistics and epidemiology. Measurement in medicine.* . UK; Cambridge, 2011.
23. Terwee, C.B., et al., *Quality criteria were proposed for measurement properties of health status questionnaires.* J Clin Epidemiol, 2007. **60**(1): p. 34-42.
24. Karel, Y.H., et al., *Current management and prognostic factors in physiotherapy practice for patients with shoulder pain: design of a prospective cohort study.* BMC Musculoskelet Disord, 2013. **14**(1): p. 62.
25. Dogu, B., et al., *Which questionnaire is more effective for follow-up diagnosed subacromial impingement syndrome? A comparison of the responsiveness of SDQ, SPADI and WORC index.* J Back Musculoskelet Rehabil, 2013. **26**(1): p. 1-7.
26. Rabin R, O.M., Oppe M, Janssen B, Herdman M. , *EQ-5D User Guide; Basic information on how to use the EQ-5D-5L instrument.* The EuroQol Group, 2011.
27. Canaway, A.G. and E.J. Frew, *Measuring preference-based quality of life in children aged 6-7 years: a comparison of the performance of the CHU-9D and EQ-5D-Y--the WAVES pilot study.* Qual Life Res, 2013. **22**(1): p. 173-83.
28. Roach, K.E., et al., *Development of a shoulder pain and disability index.* Arthritis Care Res, 1991. **4**(4): p. 143-9.
29. de Winter, A.F., et al., *The Shoulder Disability Questionnaire differentiated well between high and low disability levels in patients in primary care, in a cross-sectional study.* J Clin Epidemiol, 2007. **60**(11): p. 1156-63.
30. van der Heijden, G.J., P. Leffers, and L.M. Bouter, *Shoulder disability questionnaire design and responsiveness of a functional status measure.* J Clin Epidemiol, 2000. **53**(1): p. 29-38.
31. van der Windt, D.A., et al., *Shoulder disorders in general practice: prognostic indicators of outcome.* Br J Gen Pract, 1996. **46**(410): p. 519-23.
32. EuroQolGroup, *EuroQol – a new facility for the mea- surement of health-related quality of life.* Health Policy, 1990. **16: 199–208**.
33. Szende, A., M. Oppe, and N. Devlin, *EQ5D value sets- inventory, comparative review and user guide.* 2007, Dordrecht: Springer.
34. Johnson, J.A. and A.S. Pickard, *Comparison of the EQ-5D and SF-12 health surveys in a general population survey in Alberta, Canada.* Med Care, 2000. **38**(1): p. 115-21.
35. Lamers, L.M., et al., *The Dutch tariff: results and arguments for an effective design for national EQ-5D valuation studies.* Health Econ, 2006. **15**(10): p. 1121-32.
36. Streiner, D.L. and G.R. Norman, *Health measurement scales a practical guide to the development and use.* . 2008: Oxford university press.
37. Kamper, S.J., Ostelo, R.W., Knol, D.L., Maher, C.G., de Vet, H.C. & Hancock, M.J., *Global Perceived Effect scales provided reliable assessments of health transition in people with musculoskeletal disorders, but ratings are strongly influenced by current status.* Journal of Clinical Epidemiology, 2010. **63**(7): p. 760-766.
38. Weenink, J.W., J. Braspenning, and M. Wensing, *Patient reported outcome measures (PROMs) in primary care: an observational pilot study of seven generic instruments.* BMC Fam Pract, 2014. **15**: p. 88.

39. Luijsterburg, P.A., et al., *Physical therapy plus general practitioners' care versus general practitioners' care alone for sciatica: a randomised clinical trial with a 12-month follow-up.* Eur Spine J, 2008. **17**(4): p. 509-17.
40. Jorritsma, W., et al., *Detecting relevant changes and responsiveness of Neck Pain and Disability Scale and Neck Disability Index.* Eur Spine J, 2012. **21**(12): p. 2550-7.
41. Soer, R., et al., *Responsiveness and minimal clinically important change of the Pain Disability Index in patients with chronic back pain.* Spine (Phila Pa 1976), 2012. **37**(8): p. 711-5.
42. Demoulin, C., et al., *What factors influence the measurement properties of the Roland-Morris disability questionnaire?* Eur J Pain, 2010. **14**(2): p. 200-6.
43. Farrar, J.T., et al., *Clinical importance of changes in chronic pain intensity measured on an 11-point numerical pain rating scale.* Pain, 2001. **94**(2): p. 149-58.
44. de Vet, H.C., et al., *Minimally important change determined by a visual method integrating an anchor-based and a distribution-based approach.* Qual Life Res, 2007. **16**(1): p. 131-42.
45. de Vet, H.C., et al., *Minimally important change values of a measurement instrument depend more on baseline values than on the type of intervention.* J Clin Epidemiol, 2014.
46. Strand, L.I., et al., *The Short-Form McGill Pain Questionnaire as an outcome measure: test-retest reliability and responsiveness to change.* Eur J Pain, 2008. **12**(7): p. 917-25.
47. de Vet, H.C., et al., *When to use agreement versus reliability measures.* J Clin Epidemiol, 2006. **59**(10): p. 1033-9.
48. Norman, G.R., P. Stratford, and G. Regehr, *Methodological problems in the retrospective computation of responsiveness to change: the lesson of Cronbach.* J Clin Epidemiol, 1997. **50**(8): p. 869-79.
49. Ekeberg, O.M., et al., *A questionnaire found disease-specific WORC index is not more responsive than SPADI and OSS in rotator cuff disease.* J Clin Epidemiol, 2010. **63**(5): p. 575-84.
50. Schmitt, J.S. and R.P. Di Fabio, *Reliable change and minimum important difference (MID) proportions facilitated group responsiveness comparisons using individual threshold criteria.* J Clin Epidemiol, 2004. **57**(10): p. 1008-18.
51. Tveita, E.K., et al., *Responsiveness of the shoulder pain and disability index in patients with adhesive capsulitis.* BMC Musculoskelet Disord, 2008. **9**: p. 161.
52. Kooijman, M., et al., *Jaarcijfers 2010 en trendcijfers 2006-2010 fysiotherapie.* Landelijke Informatievoorziening Paramedische Zorg. Utrecht: NIVEL, http://www.nivel.nl/lipz.

Thoomes-de Graaf M., Scholten-Peeters G.G., Karel Y.H., Verwoerd A.., Koes B.W. , Verhagen A.P.

Qual Life Res. 2018 Feb;27(2):401-410. doi: 10.1007/s11136-017-1698-y. [Epub 2017 Sep 7].

ONE QUESTION MIGHT BE CAPABLE OF REPLACING THE SHOULDER PAIN AND DISABILITY INDEX (SPADI) WHEN MEASURING DISABILITY: A PROSPECTIVE COHORT STUDY

CHAPTER 8

ABSTRACT

Questions: Is it possible to replace the Shoulder Pain and Disability Index (SPADI) with a single substitute question for people with shoulder pain, when measuring disability and how well does this substitute question perform as a predictor for recovery.

Design: A prospective cohort study.

Participants: A total of 356 patients with shoulder pain in primary care.

Analyses: Convergent, divergent and "known" groups validity were assessed by using hypotheses testing. Responsiveness was assessed using the Receiver Operating Curve and hypothesis testing. In addition, we performed multivariate regression to assess if the substitute question showed similar properties as the SPADI and if it affected the model itself, using recovery as an outcome.

Results: The Spearman correlation coefficient between the total SPADI score and the substitute question was high, and moderate with the Shoulder Disability Questionnaire. The correlation between the substitute question and the EQ-5D-3L was low and the responsiveness was acceptable. The substitute question did not significantly contribute to both prognostic prediction models as opposed to the SPADI. Regardless all models showed poor to fair discrimination.

Conclusion: The single question is a reasonable substitute for the SPADI and can be used as a screening instrument for shoulder disability in primary clinical practice. It has slightly poorer predictive power and should therefore not be used for prognosis.

Keywords: SPADI, single question, disability, shoulder, questionnaire

INTRODUCTION

Activity limitations are one of the most important health consequences for patients with shoulder pain [1]. Activity limitations can range from difficulties with opening a jar and getting dressed, to impeding sleep [2]. Shoulder pain presents an economic burden on society due to costs of sick leave and health care and also impacts patient's quality of life [3]. As such, health related patient reported outcome measures (PROMs) that assess perceived activity limitations are useful in terms of assessing the physical impairment in patients with shoulder pain [4, 5].

Both the Shoulder Pain and Disability Index (SPADI) as the Shoulder Disability Questionnaire (SDQ) are PROMs focusing on activity limitations. Several (systematic) reviews have encouraged the use of the SPADI in both clinical and research settings [6-8].

A survey among physiotherapists (PTs) concluded that PROMs are most often used to ensure quality of care, to communicate with other health care providers, and to determine progress (outcomes) of individual patients [9]. These findings are consistent among other health care professionals [10]. Apart from this, a PROM can be used to predict recovery. For example, there is consistent evidence that a high level of disability is one of the predictors of poor recovery for patients with shoulder pain [11].

Nevertheless, PROMs are not (fully) integrated into clinical practice yet. A survey among nearly 500 PTs concluded that only half of them regularly used a PROM during their work [9]; this is consistent with other health care providers [12]. The most common reasons for not using PROMs is that it is too time consuming for patients to complete (43%) and for clinicians to analyse, calculate, and score (30%); moreover, several PROMs are too difficult for patients to complete independently (29.1%) [9]. Even the PTs that do use PROMs during their work, agreed (more than 75%) with the problems described by the non-users and also stated that PROMs are often confusing to patients.

Several initiatives have been started as a response to these concerns to facilitate the integration of PROMs in clinical care. Clinicians prefer PROMs that can be completed quickly (70%) [9]. Therefore, modifications and abbreviations of several PROMs have been developed and validated [13, 14]. Recently, the Patient-Reported Outcomes Measurement Information System (PROMIS) was developed using sample qualitative input from patients and specific analysing methods (item response theory), to construct and evaluate a preliminary item bank to measure physical functioning [15]. Computer-adaptive testing has tremendous potential for a quick and precise PROM assessment, with significantly reduced burden for patients and clinicians [16]. Another initiative is the development of single substitute questions; recently a study concluded that it may be feasible to replace the Tampa Scale for Kinesiophobia by a single substitute question for predicting outcome in people with sciatica in primary care [17].

We therefore aimed to develop and evaluate the validity, responsiveness and predictive power of a single substitute question for the SPADI as this might be helpful to integrate a PROM into clinical practice.

METHODS

Design

This is a secondary analysis of a prospective cohort study (ShoCoDiP-study), including patients with shoulder pain in physiotherapy setting. Aims of the ShoCoDiP-study were e.g. to evaluate physiotherapy care and prognostic factors in patients with shoulder pain and investigate whether Musculoskeletal ultrasound and the working alliance are related to patient recovery. Details of the design are presented elsewhere [18]. The Medical Ethics Committee of the Erasmus Medical Center in Rotterdam approved the study (MEC-2011-414). Informed consent was obtained from all patients.

Study population

Patients were recruited from primary care physiotherapy clinics between November 2011 and December 2012. Patients with shoulder pain were eligible for inclusion if they were at least 18 years old and adequately understood the Dutch language. Patients with serious pathology (infection, cancer or fracture), previous surgery or diagnostic imaging techniques of the shoulder, such as Magnetic Resonance Imaging or Ultrasound in the previous 3 months, were excluded [18].

Development of the substitute question

In a focus meeting with the ShoCoDiP-project team (consisting of physiotherapists, manual therapists, general practitioners, a radiologist, an orthopaedic surgeon and epidemiologists) various items were discussed that could act as a substitute question to cover the entire domain of the SPADI questionnaire. The final substitute question was chosen based on consensus within the research team: "Please state the amount of limitation in daily activity you experience due to your shoulder pain". This question could be answered on an 11-point scale, where: 0 = no limitation at all and 10 = completely disabled".

Baseline measurement

Participating patients received an online questionnaire that included items focused on demographic characteristics, pain intensity (Numeric Rating Scale (NRS)), disability (the SDQ, SPADI and substitute question) and health related quality of life (EQ-5D-3L).

Pain intensity

The 11-point NRS was used to capture the patient's pain intensity. The scale is anchored from "no pain" to "worst imaginable pain". Patients rate their current level of pain and their worst and least amount of pain in the last 24 hours. The NRS has shown to be valid, reliable and responsive in patients with shoulder pain [5].

Activity limitations

The SPADI is a self-administered questionnaire designed to measure pain and disability associated with shoulder pain. It consists of 13 items and each question refers to the past week. Five items measure severity/intensity of pain and 8 items measure disability. Items can be scored on a scale ranging from 0 to 10, where 0 represents "no pain/no difficulty" and 10 "worst pain imaginable/ so difficult it requires help" [19, 20]. The total score varies between 0 and 100, a higher score indicates a higher level of pain related disability [19]. The Dutch SPADI (SPADI-D) has shown to be valid (hypothesis testing, factor structure), reliable (internal consistency and test-retest), interpretable (measurement error, floor and ceiling effects) and responsive, in patients with shoulder pain in primary care [21, 22].

The SDQ is a pain-related disability questionnaire developed in Dutch, which consists of 16 items [4, 23]. All items refer to the preceding 24 hours. Response options are "yes", "no" or "not applicable". The option "not applicable" indicates the situation at issue has not occurred in the past 24 hours. The SDQ-score can range from 0 to 100 with a higher score indicating more severe disability [4, 23]. The SDQ is a valid and responsive measure [1, 24].

Quality of life

The EQ-5D-3L is a health-related quality of life questionnaire covering 5 dimensions of health: mobility, self-care, usual activities, pain/discomfort and anxiety/depression [25]. Response options are "no problems", "some problems", "extreme problems". The Dutch version is an official language version [25].

Follow up

All patients received the SPADI-D, the SDQ, the substitute question and the Global Perceived Effect (GPE)-scale 26 weeks after initial presentation. Within this period, the patient received individualized physiotherapy treatment for one or more sessions. Outcome measure was perceived recovery by the patient, measured with the GPE-scale. The GPE-scale is a 7-point scale scoring whether the patient's condition has improved or deteriorated. This scale ranges from "completely recovered" to "worse than ever". The GPE-scale has good test-retest reliability and correlates well with changes in pain and disability [26].

Analysis

All statistical analyses were performed with SPSS 23. For this study, all patients that did not answer the substitute question were excluded. Handling of missing items for the SPADI and SDQ was performed as described by the original authors [19, 27]. This means that patients were excluded from the analysis if there were more than two items missing per SPADI-subscale [19] or when more than two items were missing from the SDQ [27]. The total score of the questionnaires for the included patients were calculated by adding up the item scores and dividing them only by the number of items that were answered and deemed applicable to the subject [19, 27].

All data were checked on normality, using a Stem-and-leaf Plot, Q-Plot and Whisker box. Non-parametric tests were used if data was not normally distributed. Descriptive statistics were used to calculate frequencies.

Validity

Correlations and hypotheses

Correlations were calculated using the Pearson correlation coefficient in case of a normal distribution of the data, otherwise a Spearman correlation coefficient was used. Correlations were rated as follows: $r < 0.30$ as low (a negligible correlation) ; $0.30 \leq r < 0.45$ as moderate; $0.45 \leq r < 0.60$ as substantial and $r \geq 0.60$ as high [28].

Convergent validity relates to the extent to which a particular instrument corresponds to the construct (theoretical concept) of shoulder pain and function [29]. As the substitute question is designed to possibly replace the SPADI, we hypothesize that the correlation between substitute question and the total score of the SPADI is high ($r \geq 0.60$). We also measured the correlation between the substitute question and the SDQ, as the instruments are based on a similar construct, we expected a high correlation as well, but lower than the correlation with the SPADI (as the substitute question is designed to replace the SPADI). The SDQ has a different type of answering option and the focus of the SDQ lies on "pain during an activity", as opposed to the SPADI of which the majority of questions is focussed on "difficulties with performing an activity due to pain". We therefore expected the substitute question to be highly correlated ($r>0.60$) with the SPADI and substantially correlated (r between 0.45 and 0.60) with the SDQ [29].

Divergent validity relates to the extent to which a particular instrument does not correspond to the construct (theoretical concept) of shoulder pain and function. As two items of the EQ-5D-3L and the substitute question are based on different constructs (the mobility-item and the item anxiety/depression) we expect the correlation coefficient between both to be low ($r < 0.30$) [29].

Known groups validity We assumed that patients with high initial pain (>7 on the Numeric Rating Scale in the preceding 24 h) and work absence would have a higher level

of perceived disability. Both groups had been chosen a priory. The independent sample Mann Whitney U test was used to test the difference between known groups.

Responsiveness

Responsiveness was assessed using the area under the ROC curve (AUC) and hypothesis testing. Patients were selected if they completed the SPADI-D and the substitute question at baseline and follow up and the GPE-scale at follow-up at 26 weeks.

AUC method

We calculated the AUC to assess the ability of the substitute question to discriminate between patients who are considered improved and not importantly changed according to the GPE, using a frequently used anchor and considered patients as recovered when they answered they were 'completely recovered' or 'much improved' and as not importantly improved when they answered 'slightly improved', 'no change' or 'slightly worse' [30-32].

A benchmark that has been previously used to establish that outcome measures are useful in discriminating improved and unimproved patients has been set at 0.70 AUC [33].

Hypothesis testing

Hypothesis testing for responsiveness was based on the concept that the correlation between the change score of related constructs (SPADI) must be high. Hypothesis testing was quantified by the Pearson correlation coefficient in case of a normal distribution of the data and otherwise a Spearman correlation coefficient was used. Correlation coefficients between the substitute change score and the SPADI change score were expected to be above 0.50 [34]. A substantial correlation (r between 0.45 and 0.60) was also expected between the change score of substitute question and the change score of the SPDQ and the GPE scale. Correlations between the change score of the substitute question and the change score of EQ-5D-3L mobility as well as the anxiety/depression item, were expected to be low ($r < 0.30$).

Predictive power

Multivariate logistic regression analysis was used to predict recovery after 26 weeks. All assumptions (linearity between independent variables and log odds and multicollinearity (>0.80) for continuous variables) were checked before model building. We included no more than one independent variable per 10 events (for the smallest outcome group) in the multivariable analysis [35].

Basic model

A systematic review concluded that there was moderate to strong evidence that high pain intensity, increasing age, a longer duration of complaints, and high disability at baseline predict a poorer outcome in patients with shoulder pain [11]. Another review concluded that higher age, a longer duration of shoulder pain and high disability, were associated with poor recovery [36].

Patients were selected if they completed the GPE-scale at follow-up at 26 weeks and all items of interest at baseline (age, duration of complaints, pain intensity, the substitute question and the SPADI). We checked if there were significant differences in the relevant characteristics between the patients selected in this analysis and those excluded.

Initially, three different models were built. The first model included all predictors (age, duration of complaints and pain intensity) retrieved from the systematic reviews [11, 36]. In the second model, we added the SPADI and in model 3 we added the substitute question to model 1.

Sensitivity analysis

A sensitivity analysis (model 4) was performed by adding relevant prognostic factors as found in our own analysis in the total cohort [37] and not in systematic reviews (no depression or anxiety, a paid job and good working alliance (measured with the working alliance inventory (WAI)). We chose to exclude the WAI, as the total score of the WAI was only available for 64 patients. We added the SPADI to the basic sensitivity model in model 5 and added the substitute question in model 6.

We assessed the prognostic power (Nagelkerke R^2), the discriminative ability (AUC) and the reliability of the models (Hosmer and Lemeshow). We considered a comparable (<15% difference) overall correct percentage and Nagelkerke R^2 in model 2 and 3, as an indication that it might be valid to replace the questionnaire by its substitute question in predicting outcome. An AUC can be categorized into four categories: poor discrimination (between 0.5 and 0.7), fair discrimination (between 0.7 and 0.8), acceptable discrimination (AUC > 0.8) whereas an AUC of 1.0 indicates perfect discrimination [38]. Hosmer and Lemeshow goodness of fit tests were used to assess whether or not the observed event rates match the expected event rates in subgroups of the model population, a good model fit is indicated by a non-significant result. The -2loglikelihood is the equivalent of the residuals; a lower value is a better fit.

Furthermore, we checked whether or not the total score from the SPADI and the substitute question contributed significantly to the original model (model 1), using the Chi-Square test.

We repeated this process for the sensitivity analysis with different predictors (model 4-6).

RESULTS

Patient characteristics

A total of 389 patients responded in our cohort study, 19 of them did not return the SPADI at baseline. We excluded another 14 patients due to too many missing data on the SPADI or SDQ. Of these 356 patients all answered the substitute question and were therefore included in this study. Demographic characteristics are presented in Table 1, the mean age of the patients was 49.5 (SD 13) years and 47% was male. Of these 356 patients, 250 completed the GPE after 26 weeks and answered all items of interest at baseline (age, duration of complaints, NRS and the SPADI according to the missing item criteria and the substitute question). Responsiveness was based on 237 patients answering the substitute question at baseline and follow up and the GPE-scale.

TABLE 1. Baseline characteristics

Population	Total cohort (n=356)	Cohort "Follow up" (n=250)	Not included in the predictive study (n=106)	P-value
Gender (male) (%)	166 (47%)	116 (46%)	50 (47%)	0.894
Age Mean (SD)	49.5 (13.1)	50.2 (13.0)	47.8 (13.1)	0.118
SPADI score (0-100) Mean (SD)	46.7 (21.3)	47.5 (21.2)	45.0 (21.7)	0.310
Substitute question (0-10) Median (IQR)	4 (2-6)	4 (2-6)	3.5 (1-6)	0.549
Duration of complaints in weeks Median (IQR)	12 (6-26)	12 (6-26)	12 (6-24)	0.502
Use of medication (%)	171 (49%)	129 (52%)	42 (40%)	0.055
Pain intensity (NRS) (0-10) Median (IQR)	6 (4-7)	6 (4-7)	5 (4-7)	0.068

The data of the substitute question was not normally distributed. The median score of the substitute question was 4 points with an interquartile range (IQR) from 2 to 6. The SPADI was normally distributed and had a mean of 46.7 (21.3).

As it is unusual to compare data presented in different ways, we also presented the median of the SPADI (median; 48.7, IQR: 28.8-65.0) in order to facilitate a swift visual inspection of the score of the question of interest (the substitute question) and the score of the total SPADI.

Validity

Convergent validity

The Spearman correlation coefficient between the substitute question and the total SPADI score was 0.74 and with the SDQ 0.59. Our hypotheses were confirmed as the substitute question showed a high correlation with the SPADI and a substantial correlation with the SDQ.

Divergent validity

The spearman correlation between the substitute question and the mobility-item of the EQ-5D-3L was 0.23 and with the item anxiety/depression 0.20. Our hypotheses were hereby confirmed as the correlation was low between the instruments that measure a different construct and the substitute question.

Known groups validity

Differences between "known groups" were statistically significant (table 2).

TABLE 2. Known groups validity

	Group	Median score substitute question	P-value
Pain (n=356)	High initial pain >7	6 (4-7)	0.000
	Low initial pain <7	3 (1-5)	
Work absence (n=318)	Work absence due to shoulder pain	6 (5-7	0.000
	No work absence due to shoulder pain	3 (1-5.25)	

Responsiveness

The AUC was 0.76 with a 95% confidence interval ranging from 0.70 to 0.83. Figure 1 shows the ROC curve based upon the GPE.

Hypothesis testing for responsiveness resulted in a Spearman correlation between the SPADI-D change score and the substitute change score of 0.71 and 0.60 with the SDQ change score. The spearman correlation between the GPE and the substitute question was 0.47. The Spearman correlation between the substitute question and both the mobility as the anxiety/depression item of the EQ-5D-3L was 0.10.

Based on the AUC values and confirmation of the hypothesis, we consider the substitute question to be a responsive measurement instrument.

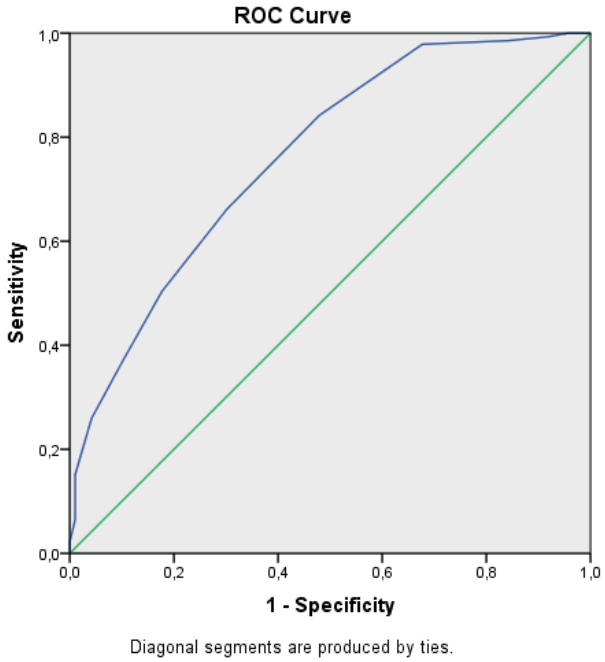


FIGURE 1. ROC curve based upon the GPE

Prediction model

There were no significant differences in the relevant characteristics between the patients selected in this analysis (n=250) and those excluded (n= 106) (Table 1).

Out of 250 patients, 150 patients were labelled as recovered after 26 weeks. For all variables included in the model the variance inflation factors were < 1.5 and correlation coefficients <0.8, suggesting that no linearity and multicollinearity was present.

Table 3 shows the predictive models. Model 1 consisted of the following variables: age, pain and duration of complaints. The correct overall percentage was 64.8% and the Nagelkerke R^2 was 0.90.

Model 2 consisted of the following variables: age, pain, duration of complaints, and the SPADI. The Chi Square test for adding the SPADI was significant (p=0.029).

Model 3 consisted of the following variables: age, pain, duration of complaints, and the substitute question. The Chi Square test for adding the substitute question was not significant (p=0.193).

All three models showed poor discrimination and the AUC values were within the 95%CI intervals of each other. Differences between both models were small (Table 3). The largest differences were found between the Hosmer and Lemeshow goodness of fit of model 2 and 3; however, both were non- significant. The odds of the SPADI and the

substitute question were quite exchangeable; however, the confidence interval of the substitute question was wider.

TABLE 3. Predictive value

	Model 1 (n=250)	Model 2 (n=250)	Model 3 (n=250)
Predictors for recovery	OR (95%CI)	OR (95%CI)	OR (95%CI)
Age (younger)	0.98 (0.96-1.00)	0.98 (0.96-1.01)	0.98 (0.96-1.00)
Duration of complaints (in weeks)(shorter)	0.99 (0.99-1.00)	0.99 (0.99-1.00)	0.99 (0.99-1.00)
Pain using an NRS (lower levels of pain)	0.92 (0.80-1.05)	1.02 (0.87-1.21)	0.97 (0.83-1.13)
Disability using the total SPADI score (lower level of functional disability)		0.98 (0.97-1.00)	
Disability using the substitute question (lower level of functional disability)			0.92 (0.81-1.04)
Performance of the model			
Correct overall percentage	64.8%	65.6%	65.2%
Nagelkerke R^2	0.090	0.114	0.098
AUC (95%CI)	0.64 (0.57-0.72)	0.66 (0.59-0.73)	0.65 (0.58-0.72)
Hosmer and Lemeshow	0.757	0.875	0.553
-2 Log likelihood	319.286	314.534	317.594

Model 1: age, duration of complaints and pain; Model 2: age, duration of complaints, pain and the SPADI; Model 3: age, duration of complaints, pain and the substitute question

Sensitivity analysis

The basic model (model 4) consisting of age, duration of complaints, pain, employment and not being depressed and was based on 241 patients, as nine patients had a missing value regarding employment or depression. The correct overall percentage was 63.9% and the Nagelkerke R^2 was 0.127.

Model 5 included all predictors plus the SPADI. The Chi Square Omnibus test for adding the SPADI was significant (p= 0.039).

Model 6 included all predictors plus the substitute question. The Chi Square test for adding the substitute question was not significant (p=0.501) Table 4.

All models showed poor discrimination, with small differences. The largest differences were found between the Hosmer and Lemeshow goodness of fit of model 4 and 5; however, both were non- significant. The odds of the SPADI and the substitute question were again quite exchangeable; however, the confidence interval of the substitute question was wider.

TABLE 4. Sensitivity analysis

	Model 4 (n=241)	Model 5 (n=241)	Model 6 (n=241)
Predictors for recovery	OR (95%CI)	OR (95%CI) p-value	OR (95%CI) p-value
Having a job	1.77 (0.87-3.62)	1.80 (0.88-3.68)	1.75 (0.85-3.57)
Being depressed (not being depressed helps to recover)	0.41 (0.20-0.85)	0.42 (0.21-0.88)	0.43 (0.21-0.89)
Age (younger)	0.99 (0.97-1.02)	0.99 (0.96-1.02)	0.99 (0.97-1.02)
Duration of complaints (in weeks) (shorter)	0.99 (0.99-1.00)	0.99 (0.99-1.00)	0.99 (0.99-1.00)
Pain using an NRS (lower levels of pain)	0.95 (0.83-1.09)	1.06 (0.89-1.27)	0.98 (0.83-1.14)
Lower disability (SPADI total score)		0.98 (0.97-1.00)	
Lower disability (substitute question)			0.96 (0.84-1.09)
Performance of the model			
Correct overall percentage	63.9%	66.0%	66.8%
Nagelkerke R^2	0.127	0.149	0.130
AUC (95%CI)	0.67 (0.60-0.74)	0.69 (0.62-0.75)	0.68 (0.61-0.74)
Hosmer and Lemeshow	0.310	0.853	0.051
-2 Log likelihood	301.001	296.753	300.547

Model 4: age, duration of complaints, pain, depression and being employed; Model 5: age, duration of complaints, pain, depression, being employed, the SPADI; Model 6: age, duration of complaints, pain, depression, being employed, the substitute question.

DISCUSSION

Measurement with the single question can be completed in a shorter amount of time as compared with the SPADI, which takes about three minutes to complete. This could have impact on the use of the instrument in clinical practice and increase the integration of patient -reported outcome measures (PROMs), as the most common reasons for not using them are that they are too time consuming for patients to complete and too time consuming for clinicians to analyse. Quality of life research revealed that both single questions and multi-item scales have a high potential as well as some disadvantages at the same time [39]. They stated that the two types of indices are not mutually exclusive and can be used together in a single research study or in a clinical setting. Single items have the advantage of simplicity at the cost of detail [39]. Multiple-item indices have the advantage of providing a complete profile of quality of life component constructs at the cost of increased burden and of asking potentially irrelevant questions [39].

However, the predictive power of the substitute question is not entirely equal to the SPADI as the substitute question did not significantly contribute to both models according to the Chi- Square test, as opposed to the SPADI. Regardless, switching between the SPADI and the substitute question did not have a great impact on the AUC, as all

models (with the SPADI and the substitute question) showed poor discrimination. The predictive power of the model including the substitute question for predicting recovery was slightly lower compared to the model with the SPADI, both were poor. As these prediction models should be used carefully, this especially applies to using the substitute question as a predictor.

Comparison to the literature

Not many studies have been published regarding a substitute question. One study reported that a single self-reported question to assess habitual physical activity is valid and responsive to change and thus useful for epidemiological research in community-dwelling older people, also in follow-up studies. They found correlations between self-reported habitual physical activity and mobility and accelerometer-based physical activity variables [40]. Another study assessed the reliability, the specificity and sensitivity of a single question (with a dichotomized answering option) regarding hearing impairment in elder people. The reliability of the single question was lower than the reliability of the complete questionnaire. Their conclusion was that the entire instrument was more effective in assessing the impact of a hearing impairment on quality of life than the single question [41]. A third study assessed if the use of single items of a depression questionnaire were a reasonable alternative to the total scale in chiropractic patients with low back pain. They analysed the association between the single candidate items and outcome, as well as the predictive capacity of both the total questionnaire as the single items. The conclusion of the authors was that a single item (no. 1 or 3) was a reasonable substitute for the entire scale when screening for depression as a prognostic factor [42]. The first study that assessed validity, responsiveness and predictive power of a substitute question compared to a complete questionnaire, found a similar result with regards to the Tampa Scale for Kinesiophobia [17]. The conclusion of this manuscript was that the unique single substitute question might be able to replace the Tampa Scale.

Strengths and limitations

This is a new type of research, which is focused on a very pragmatic solution regarding the disuse of PROMs. The population consisted of patients from primary care, a population that is very important within the health care system and where pain-related disability is a relevant issue. We had a relatively high number of included patients, although this could have been higher if we had chosen to use imputation techniques instead of excluding patients due to the missing item criteria. We chose to respect these criteria, as our aim was to assess whether or not the substitute question might be feasible to replace the SPADI, and the criteria of the PROMs itself are therefore more important than to use imputation techniques, in order to make a more steady prediction model due to the higher number of included patients. As the demographic characteristics of the in-

cluded and excluded patients did not differ, it seems unlikely that there is selection bias regarding the inclusion of patients in the responsiveness and predictive power analyses. There were no remarkable deviations with regards to the patient characteristics of the complete study population compared to the target population (patients with shoulder pain in primary care) as far as we could discern, e.g. the number of participating females was higher than the number of participating males, which is in line with the gender specific incidence [43], as was the average age [44].

Patients were asked to answer if their shoulder pain had changed since the beginning of treatment. The time between baseline and follow up was 26 weeks, which might have influenced their recollection of their shoulder problem at the beginning. Although this is common practice, this could have an impact on the results.

Although the SPADI is designed as if it consists of two parts (pain and disability), we chose to only formulate one substitute question and to assess the correlation with the total SPADI. The theoretical deviation into two separate parts has not been confirmed in our earlier study [21]. As the majority of the SPADI questions focuses on difficulties with performing an activity due to pain we formulated the substitute question with a similar focus (difficulties with performing an activity due to shoulder pain).

Future research

It is important to test the content validity of the substitute question, with patients, clinicians and experts together. Besides, the reliability, validity, responsiveness and predictive value should be further assessed before this question can be used in clinical practice.

Conclusion

The correlation between the substitute question and the full SPADI was relatively high. Combined with acceptable responsiveness, the substitute question can potentially be used as a screening instrument for shoulder disability in primary clinical practice. The single question has slightly poorer predictive power than the complete SPADI, and should therefore not be used for prognosis at this moment.

REFERENCES

1. Van Der Windt, D.A.W.M., et al., *The responsiveness of the Shoulder Disability Questionnaire.* Ann Rheum Dis, 1998. **57**(2): p. 82-7.
2. Feleus, A., et al., *Management in non-traumatic arm, neck and shoulder complaints: differences between diagnostic groups.* Eur Spine J, 2008. **17**(9): p. 1218-29.
3. Huisstede, B.M., et al., *Incidence and prevalence of upper-extremity musculoskeletal disorders. A systematic appraisal of the literature.* BMC Musculoskelet Disord, 2006. **7**: p. 7.
4. van der Windt, D.A., et al., *The responsiveness of the Shoulder Disability Questionnaire.* Ann Rheum Dis, 1998. **57**(2): p. 82-7.
5. Mintken, P.E., P. Glynn, and J.A. Cleland, *Psychometric properties of the shortened disabilities of the Arm, Shoulder, and Hand Questionnaire (QuickDASH) and Numeric Pain Rating Scale in patients with shoulder pain.* J Shoulder Elbow Surg, 2009. **18**(6): p. 920-6.
6. Bot, S.D., et al., *Clinimetric evaluation of shoulder disability questionnaires: a systematic review of the literature.* Ann Rheum Dis, 2004. **63**(4): p. 335-41.
7. Roy, J.S., J.C. MacDermid, and L.J. Woodhouse, *Measuring shoulder function: a systematic review of four questionnaires.* Arthritis Rheum, 2009. **61**(5): p. 623-32.
8. Breckenridge, J.D. and J.H. McAuley, *Shoulder Pain and Disability Index (SPADI).* J Physiother, 2011. **57**(3): p. 197.
9. Jette, D.U., et al., *Use of standardized outcome measures in physical therapist practice: perceptions and applications.* Phys Ther, 2009. **89**(2): p. 125-35.
10. Snyder, C.F., et al., *Implementing patient-reported outcomes assessment in clinical practice: a review of the options and considerations.* Qual Life Res, 2012. **21**(8): p. 1305-14.
11. Kuijpers, T., et al., *Systematic review of prognostic cohort studies on shoulder disorders.* Pain, 2004. **109**(3): p. 420-31.
12. Russak, S.M., et al., *The use of rheumatoid arthritis health-related quality of life patient questionnaires in clinical practice: lessons learned.* Arthritis Rheum, 2003. **49**(4): p. 574-84.
13. Stratford, P.W. and J.M. Binkley, *Measurement properties of the RM-18. A modified version of the Roland-Morris Disability Scale.* Spine (Phila Pa 1976), 1997. **22**(20): p. 2416-21.
14. Beaton, D.E., et al., *Development of the QuickDASH: comparison of three item-reduction approaches.* J Bone Joint Surg Am, 2005. **87**(5): p. 1038-46.
15. Rose, M., et al., *Evaluation of a preliminary physical function item bank supported the expected advantages of the Patient-Reported Outcomes Measurement Information System (PROMIS).* J Clin Epidemiol, 2008. **61**(1): p. 17-33.
16. Turner, R.R., et al., *Patient-reported outcomes: instrument development and selection issues.* Value Health, 2007. **10 Suppl 2**: p. S86-93.
17. Verwoerd, A.J., et al., *A single question was as predictive of outcome as the Tampa Scale for Kinesiophobia in people with sciatica: an observational study.* J Physiother, 2012. **58**(4): p. 249-54.
18. Karel, Y.H., et al., *Current management and prognostic factors in physiotherapy practice for patients with shoulder pain: design of a prospective cohort study.* BMC Musculoskelet Disord, 2013. **14**(1): p. 62.
19. Roach, K.E., et al., *Development of a shoulder pain and disability index.* Arthritis Care Res, 1991. **4**(4): p. 143-9.
20. Paul, A., et al., *A comparison of four shoulder-specific questionnaires in primary care.* Ann Rheum Dis, 2004. **63**(10): p. 1293-9.

21. Thoomes-de Graaf, M., et al., *The Dutch Shoulder Pain and Disability Index (SPADI): a reliability and validation study.* Qual Life Res, 2014.
22. Thoomes-de Graaf, M., et al., *The Responsiveness and Interpretability of the Shoulder Pain and Disability Index.* J Orthop Sports Phys Ther, 2017: p. 1-21.
23. Jamnik, H. and M.K. Spevak, *Shoulder pain and disability Index: Validation of slovene version.* Int J Rehabil Res, 2008. **31**(4): p. 337-41.
24. de Winter, A.F., et al., *The Shoulder Disability Questionnaire differentiated well between high and low disability levels in patients in primary care, in a cross-sectional study.* J Clin Epidemiol, 2007. **60**(11): p. 1156-63.
25. Lamers, L.M., et al., *The Dutch tariff: results and arguments for an effective design for national EQ-5D valuation studies.* Health Econ, 2006. **15**(10): p. 1121-32.
26. Kamper, S.J., Ostelo, R.W., Knol, D.L., Maher, C.G., de Vet, H.C. & Hancock, M.J., *Global Perceived Effect scales provided reliable assessments of health transition in people with musculoskeletal disorders, but ratings are strongly influenced by current status.* Journal of Clinical Epidemiology, 2010. **63**(7): p. 760-766.
27. de Winter, A.F., et al., *The Shoulder Disability Questionnaire differentiated well between high and low disability levels in patients in primary care, in a cross-sectional study.* J Clin Epidemiol, 2007. **60**(11): p. 1156-63.
28. Burnand, B., W.N. Kernan, and A.R. Feinstein, *Indexes and boundaries for "quantitative significance" in statistical decisions.* J Clin Epidemiol, 1990. **43**(12): p. 1273-84.
29. Mokkink, L.B., et al., *The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content.* BMC Med Res Methodol, 2010. **10**: p. 22.
30. Weenink, J.W., J. Braspenning, and M. Wensing, *Patient reported outcome measures (PROMs) in primary care: an observational pilot study of seven generic instruments.* BMC Fam Pract, 2014. **15**: p. 88.
31. Luijsterburg, P.A., et al., *Physical therapy plus general practitioners' care versus general practitioners' care alone for sciatica: a randomised clinical trial with a 12-month follow-up.* Eur Spine J, 2008. **17**(4): p. 509-17.
32. Farrar, J.T., et al., *Clinical importance of changes in chronic pain intensity measured on an 11-point numerical pain rating scale.* Pain, 2001. **94**(2): p. 149-58.
33. Terwee, C.B., et al., *Quality criteria were proposed for measurement properties of health status questionnaires.* J Clin Epidemiol, 2007. **60**(1): p. 34-42.
34. de Vet HC, T.C., Mokkink LB, Knol DL, *Practical guides to biostatistics and epidemiology. Measurement in medicine.* . UK; Cambridge, 2011.
35. Harrell, F.E., Jr., K.L. Lee, and D.B. Mark, *Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors.* Stat Med, 1996. **15**(4): p. 361-87.
36. Chester, R., et al., *Predicting response to physiotherapy treatment for musculoskeletal shoulder pain: a systematic review.* BMC Musculoskelet Disord, 2013. **14**: p. 203.
37. Karel, Y.H., et al., *Development of a Prognostic Model for Patients With Shoulder Complaints in Physiotherapy.* Phys Ther, 2016.
38. Hosmer, D.W.J., S. Lemeshow, and R.X. Sturdivant, *Applied Logistic Regression*. 2013: John Wiley & Sons.
39. Sloan, J.A., et al., *Assessing the clinical significance of single items relative to summated scores.* Mayo Clin Proc, 2002. **77**(5): p. 479-87.

40. Portegijs, E., et al., *Validity of a single question to assess habitual physical activity of community-dwelling older people.* Scand J Med Sci Sports, 2016.
41. Tomioka, K., et al., *The Hearing Handicap Inventory for Elderly-Screening (HHIE-S) versus a single question: reliability, validity, and relations with quality of life measures in the elderly community, Japan.* Qual Life Res, 2013. **22**(5): p. 1151-9.
42. Kongsted, A., et al., *Brief screening questions for depression in chiropractic patients with low back pain: identification of potentially useful questions and test of their predictive capacity.* Chiropr Man Therap, 2014. **22**(1): p. 4.
43. Picavet, H.S. and J.S. Schouten, *Musculoskeletal pain in the Netherlands: prevalences, consequences and risk groups, the DMC(3)-study.* Pain, 2003. **102**(1-2): p. 167-78.
44. Kooijman, M., et al., *Jaarcijfers 2010 en trendcijfers 2006-2010 fysiotherapie.* Landelijke Informatievoorziening Paramedische Zorg. Utrecht: NIVEL, http://www.nivel.nl/lipz.

GENERAL DISCUSSION

CHAPTER 9

The aim of this thesis was to evaluate the diagnostic process of (Dutch) physiotherapists (PTs) in patients with shoulder pain, mainly with regards to Patient Reported Outcome Measures (PROMs) and the use of diagnostic musculoskeletal ultrasound (DMUS). Main findings and their limitations are discussed per topic. All studies involving patients were part of the 'Shoulder Complaints and Diagnostic Ultrasound in Physiotherapy' (ShoCo-DiP) project.

Patient reported outcome measures (PROMS)

Objectifying Functional Disability with PROMs; reviewing the existing evidence

A number of reviews have been performed on shoulder-specific PROMs focusing on 'activity limitations' [1-3]. Nevertheless, we performed a new systematic review ourselves, because of the following reasons:

Firstly, there is a great variety in PROMs available to measure 'activity limitations' in patients with shoulder pain, e.g. there are self-administered PROMs and PROMs including a physical examination component. Moreover, the study population of interest can have impact on e.g. the responsiveness. Some PROMs are assessed in a mixed study population, such as in patients with upper extremity complaints, which is not comparable with patients with shoulder pain only. We decided to narrow our research question in order to provide a more solid statement for the specific patient group of interest. We focused on self- administered PROMs with a main goal to measure 'activity limitations' due to shoulder pain and to limit our study-population to patients with non-specific shoulder pain.

Secondly, the methodological quality of studies investigating the measurement properties of PROMs should be assessed in a standardized way [4]. We decided to use the recently developed COSMIN-checklist for this purpose. The COSMIN-checklist was unavailable when previous reviews had been performed [1, 2] or reviews have summarized the characteristics and measurement properties of a limited number of PROMs, but did not assess the methodological quality of the included studies or the impact of methodological quality on the results, and consequently their conclusions have several limitations [5-7]. One review however examined the psychometric properties of PROMs using the COSMIN on patients with rotator cuff disorders [8], however they included studies using a mixed population despite their specific aim of the study.

Thirdly, we aimed to present the results per PROM per language version. Other reviews have neglected differences in cultural context and translations of the original version, while this may influence the psychometric properties [9-11].

The results of our systematic review (Chapter 2) indicate that the Shoulder Pain and Disability Index (SPADI) for English users was rated best. We found moderate evidence

for construct hypothesis testing and responsiveness, and strong evidence for internal consistency, however no statements with regards to reliability, measurement error or content validity could be made. PROMs in other languages than English, Dutch or Norwegian only received an 'unknown', 'poor' or 'limited' evidence score on one or more measurement properties [12]. As the SPADI is the most widely used PROM and has the best ratings in several languages, it would be useful to assess the Dutch version for both clinical and research purposes.

Clinicians have to consider the quality of measurement properties combined with the intended purpose of using a PROM. For instance, Dutch users could either choose between the Simple Shoulder Test (SST) and the Shoulder Disability Questionnaire (SDQ), however no measurement properties are available with regards to responsiveness for the SST. If a clinician aims to evaluate the response to (physiotherapy) treatment over time, this should be taken into account.

Due to differences in the choice of study populations and the inclusion of the types of PROMs, our results were different from other systematic reviews. One review stated the DASH received the best ratings, however they included studies evaluating the DASH that did not report their results for shoulder pain patients separately [1]. Two reviews stated the measurement properties of the PROMs assessed were acceptable [2, 3]. Both reviews did not perform an evidence synthesis; the psychometric properties per PROM were presented but without the methodological quality per study. They also included PROMs that we excluded (the American Shoulder and Elbow Surgeons (ASES), RC-Quality Of Life (RC-QOL), Western Ontario Rotator Cuff Index (WORC), Extremity Functional Index (UEFI), Upper Extremity Functional Scale (UEFS), Upper Limb Functional Index (ULFI), American Shoulder and Elbow Surgeon questionnaire (ASES), Oxford Shoulder Score (OSS), Penn Shoulder Score (PSS)). The review using the COSMIN on a population with rotator cuff disorders included different PROMs as well and they stated the WORC showed the best overall quality. Their conclusion with regards to the SPADI were reasonably comparable to ours, although they included studies using a mixed population despite their specific aim of the study [8].

Despite our recommendations with regards to the use of the available self-administered shoulder specific PROMs, we feel there is a need for a different approach. As all of the evaluated instruments were developed in the '90s, none of these PROMs showed strong or moderate evidence for all measurement properties after twenty years of research. Meanwhile, knowledge regarding the development of a PROM has increased and recommendations have been made to instrument-developers; articulate how a particular conceptual framework guided their construct selection, item development (including e.g. in-depth interviews and focus groups with patients and experts in the field) and

psychometric testing [13]. A PROM should have evidence supporting its content validity, including evidence that patients and experts consider the content of the PROM relevant and comprehensive for the concept, population, and aim of the measurement application [14]. We found that the content validity of most PROMs for patients with shoulder pain is still unknown and could only rate the SDQ and the SDQ-UK on content validity, as some development studies did not involve patients with shoulder pain (e.g. present their results separately for patients with shoulder pain), or did not present the process and results of e.g. the importance of a question according to patients and the consequences of it [15-20].

Also, important issues concerning the limitation of functional activities have changed over time, e.g. computer use is nowadays completely integrated into everyday life, but this is not included in most PROMs. Not only have relevant items been changed, but also the available methodology and technology has reached a new level of sophistication, including "modern" psychometric techniques of item banking, item response theory (IRT) and computer-adaptive testing (CAT) [13]. It has been suggested that CAT has tremendous potential for yielding precise PROM assessment quickly and with a reduced burden for the patient/respondent [13]. Recently, the National Institute of Health has developed a Patient-Reported Outcomes Measurement Information System (PROMIS) using sample qualitative input from patients and IRT methods, to construct and evaluate a preliminary item bank for measuring physical function [21] and upper-extremity and mobility subdomain scores were constructed [22]. No studies have been performed on patients with non-specific shoulder pain so far. However, the PROMIS Physical Functioning CAT (PROMIS PF CAT) showed a high correlation with the ASES (0.67) and a substantial correlation with the Western Ontario Shoulder Instability Index (WOSI) (0.49) in patients with shoulder instability, and did not demonstrate ceiling effects [23]. The PROMIS PF upper extremity CAT (PROMIS PFUE CAT) has been validated on patients with shoulder arthritis (as it had a high correlation with the SST of 0.64 and a substantial correlation with the ASES of 0.57) [24]. The time to complete the PROMIS PFUE CAT was significantly less than the time to complete the two other PROMS (SST and ASES) (62.6 ± 22.8 seconds versus 96.9 ± 25.1 and 160.6 ± 51.5 seconds) [24]. The SPADI takes approximately 2 to 3 minutes to complete [25, 26]. No floor or ceiling effects of the PROMIS PFUE CAT were observed [24]. All studies have been performed on small study samples and not in our population of interest so far.

In conclusion, we advised to develop a new PROM focusing on 'activity limitations' using new techniques and methods for patients with shoulder pain. Meanwhile the PROMIS PFUE CAT has become available and was assessed on patients with shoulder instability and shoulder arthritis. Although the PROMIS PFUE CAT is based upon the upper extremity, the wider spectrum of items is not a problem when using computer adaptive

techniques, as only relevant questions are assessed. Therefore, the PROMIS PFUE CAT could be a promising option to assess 'activity limitations' in patients with non-specific shoulder pain in a complete and timesaving way. This should be assessed in further primary studies. We therefore propose to use the SPADI until there is a better alternative. As the SPADI is the most widely used PROM and has the best ratings in several languages, it would be useful to assess the Dutch version for both clinical and research purposes.

Assessment of clinimetrics of the SPADI-D

Currently the Royal Dutch Society for Physical Therapy (KNGF) recommends the implementation of the Dutch SPADI (SPADI-D) in their shoulder evidence based statement [27]. This is based on the fact that several reviews encouraged the use of the SPADI in clinical and research settings [1, 2, 28]. Moreover, functional limitations, as assessed by the SPADI, have been described as a predictive factor by several reviews [29-31]. Despite the SPADI is frequently used internationally (in research as well as in clinical practice), the SPADI had not been validated and tested for reliability in Dutch.

We found (chapter 3) that the SPADI-D can be considered a valid and reliable PROM. It discriminates well between extreme groups, correlates well with the SDQ and internal consistency and test-retest reliability are high [12]. Besides, the SPADI-D is responsive, making it a useful tool to evaluate change in functional disability in longitudinal studies in patients with shoulder pain visiting a PT (Chapter 7) [32]. The total SPADI score varies between 0 and 100, a higher score indicates a higher level of pain related disability [16]. A change larger than 43% of the baseline score is considered as a clinically relevant and important change for individual patients. However, a change beyond the measurement error (Smallest Detectable Change (SDC) of 19.7) should be taken into account when used for decision-making in individual patients. The measurement error is of less importance when the SPADI-D is used for research purposes in groups of patients when a mean score is used and therefore the SDC is much smaller.

Our study had some limitations. Firstly, the translation process of the SPADI-D from English to Dutch was not published and it is therefore unknown whether it was performed according to international guidelines [33]. Nevertheless, the online published SPADI-D is commonly used in clinical practice and in research and is also integrated in multiple patient-management software programs. On the other hand, a (possibly) poor translation process does not necessarily mean that the instrument has a poor cross-cultural validity [34].

Secondly, we did not use the GPE (Global Perceived Effect) scale to check if patients were indeed stable within the period of 7 days between the test and the re-test. The median duration of shoulder pain at the start of inclusion was 17 weeks in our study population, and PTs usually treat patients with shoulder pain for about 11 weeks (SD

11.3) [35]. It is therefore unlikely that patients would have been importantly recovered within one week. Although, the time frame we have chosen is commonly accepted we cannot be completely sure all included patients had not importantly changed during these 7 days [36]. Therefore, we checked data for patients with extreme change scores, as there is always the chance of an improvement or deterioration within this timeframe and we presented both the measurement error based upon the population with and without the two outliers. Clinicians and researchers can make their own choice in implementing our results; the more conservative approach would be to use the measurement error including the outliers (22.5 instead of 19.7).

Our results were reasonably comparable to other studies assessing the SPADI. Our main finding deviating from other studies is that the Dutch SPADI consists of one factor only. However, only one study reported a factor structure as originally described by Roach [37] and this study was rated as good in our review. The majority of studies could not confirm this loading pattern or reported a one-factor structure [38-41], however if included in our review, these studies were rated as fair [38, 40] or poor [39]. One study indicated the wording of the SPADI items might influence this outcome. The disability items ask respondents to indicate the amount of difficulty they have with specified functions. If difficulty in performing an activity is reported, patients might consider pain to be part of what makes the activity difficult [41]. The total SPADI-D is considered to be a pain and disability questionnaire, focusing on 'activity limitations'.

Could one question replace the SPADI to objectify and evaluate functional disability as well as in a predictive sense?

Guidelines recommend the use of PROMs and a large proportion of clinicians feel there are advantages in using PROMs. Nevertheless, PROMs are not fully integrated into clinical practice [42]. The implementation and use of PROMs are time-consuming and clinicians prefer PROMs that can be completed quickly (70%) [42, 43]. As a response, several initiatives have been started to facilitate the integration of PROMs in clinical care. Therefore, modifications and abbreviations of several PROMs have been developed and validated [44, 45]. Another initiative is the development of single substitute questions. Recently a study concluded that it may be feasible to replace the Tampa Scale for Kinesiophobia by a single substitute question for predicting outcome in people with sciatica in primary care [46]. We chose a final substitute question for the SPADI based on consensus within the research team: "Please state the amount of limitation in daily activity you experience due to your shoulder pain". This question could be answered on an 11-point scale, where: "0 = no limitation at all and 10 = completely disabled". Our results showed that the substitute question of the SPADI could possibly replace the SPADI in clinical practice, as the correlation between the substitute question and the total SPADI was relatively

high (Spearman correlation of 0.74) and showed acceptable responsiveness (AUC 0.76) [12] (Chapter 8).

However, the substitute question cannot be used as a predictive factor yet. The single question was not a significant contributor in our predictive models. On the other hand, using the SPADI or the substitute question did not have a great impact on both models either, as the discriminative ability remained poor and the explained variance was low in both models. We did not use the complete cohort in this study (as we only included patients answering all items of interest of the basic model and the SPADI and the substitute question) opposed to another ShoCoDip- paper with the main aim to develop a prognostic model. However, both in our study (N= 250) as well as the "prognostic modelling study" (N= 389) [47] the (final) model showed poor performance (discriminative ability (AUC <0.7) and explained variance (R^2 <0.15).

Nevertheless, we conclude that it is premature to state whether or not the substitute question is a predictive factor, as there is no strong evidence to support or refute this. Therefore, it would be useful to assess the predictive power of the substitute question in another study population. At present, more research needs to be done to definitely conclude whether this substitute question can replace the SPADI. It may be very practical for patients as well as clinicians to use this one question. We did not assess the content validity, although this is of great importance. We considered the substitute question to be a derivative of the total SPADI and have developed it with a focus group of PTs, General Practitioners (GPs), a radiologist, an orthopaedic surgeon and epidemiologists, however patients with shoulder pain were not part of this focus group. In order to have a complete outline, information regarding the reliability and the measurement error (plus content validity) should be available. We suggest that this substitute question could be a worthwhile alternative for clinicians to objectify and evaluate limitations in activity of patients with shoulder pain.

Objectifying alliance

Working alliance might be a predictor for improvement [48-50] and the Working Alliance Inventory (WAV-12) is one of the most commonly used questionnaires to measure working alliance, but has not been assessed in a physiotherapy setting or in Dutch.

Chapter 7 describes the results of a study assessing the validity of the Flemish Working alliance inventory short-form (WAV-12), which was a translation of the (English) short-form version of the Working Alliance Inventory (WAI) in terms of the construct and discriminative abilities for a population of patients with shoulder pain in physiotherapy care. In chapter 8 we described the WAV-12 as part of the assessment of the predictive ability of the substitute question of the SPADI.

We found that a large number of patients did not fill out the complete WAV-12; only 22% of patients answered all items of the WAV-12, therefore multiple imputation

techniques were used to assess Cronbach's alpha. The unidimensionality of the WAV-12 indicates that all items measure the same concept. The WAV-12 appears to have good discriminative abilities in the lower end of the construct, however ceiling-effects were found in 10 out of 12 items [12]. The low response rate might indicate that the measurement instrument is not appropriate either in terms of language, setting, or that patients had other specific reasons not to complete the questionnaire. In comparison, a large study, including 1871 patients following psychotherapy, showed a complete response rate of 94% (not missing a single item) [51]. This study also revealed strong evidence for ceiling effects and the authors concluded that the Working Alliance Inventory (WAI) lacks sensitivity to distinguish patients in the highest ranges [51]. Research studying the measurement properties of the Brazilian WAI on patients with low back pain resulted in similar problems (high ceiling effects) [52]. The findings regarding ceiling effects is in accordance with a recently published scoping review of the literature of the WAI [53]. The review states the WAI needs re-contextualization for suitable use in musculoskeletal practice [53].

Based on our results, we made adjustments and reworded items of the WAV-12 using Delphi rounds including both experts and patients to assess their opinion regarding the adjusted items, in order to respond to the inappropriateness to incorporate the WAV-12 into a physiotherapy setting. The WAV-12 is yet, in its current form, not implementable in Dutch physiotherapy setting and we aim to evaluate the adjusted version in the future. To our knowledge, no other PROMs are available in Dutch to measure working alliance (especially in physiotherapy setting).

The working alliance seems to have potential as a predictor of outcome [47]. Nevertheless, the adjusted version of the WAV-12 should be tested in a new group of patients, to assess if it impacts the amount of complete answering and its clinimetric properties. In order to include the adjusted WAV-12 in future studies and to use it in clinical setting, more research is needed with regards to the impact on the factor structure, reliability and construct validity.

Diagnostic musculoskeletal ultrasound (DMUS)

In chapter 4 we present the assessment of the inter-professional agreement of DMUS between PTs and radiologists in patients with shoulder pain for full thickness tears, partial thickness tear, calcification and subacromial bursitis. Also, we explored the influence of experience or training of the PTs with regards to the overall agreement.

We found substantial agreement (Kappa = 0.63) between PTs and radiologists in diagnosing a full thickness tear. The overall kappa of all four categories was 0.36, indicating fair agreement. Subgroup analysis regarding experience and education level showed the agreement between the more experienced and higher trained PTs and the radiologist was higher compared to the less experienced and basic trained PTs. Nevertheless, we

would not recommend to implement DMUS in every day clinical care, or in the educational program of physiotherapy yet. However, as education and experience impacts the agreement, perhaps in time DMUS could well be placed into the daily clinical pathway of PTs.

Can, based upon the literature, DMUS be used by PTs to distinguish between patients that a) need referral to secondary care (potentially specific or serious pathology), b) could benefit from physiotherapy management and c) those that should just be monitored? In chapter 5 we describe an explorative study, dividing patients with shoulder pain in these three treatment related categories.

The results from this study indicate the overall kappa between PTs and radiologists using these new treatment related categories was moderate (Kappa = 0.60). There was substantial agreement within the category 'referral to secondary care' (k=0.74) and both 'possible indication for physical therapy management' (k=0.57) and 'watchful waiting' (k=0.46) showed moderate agreement. Although it is too soon to implement these new treatment related categories into clinical care, it seems to be a promising avenue. The explorative study might be an opening into considering patients and diagnostic modalities in a new way in the future.

Both our agreement studies have some limitations, especially the explorative study. The most important limitation regards the instructions of both PTs and radiologists. The instructions could have been stricter in the original study, and it would have been more ideal if we would have instructed the professionals with the new treatment related categories in reality instead of recoding old data in the explorative study.

More research is needed before a conclusive statement can be made with regards to the use of DMUS in physiotherapy. If PTs want to use DMUS as a diagnostic tool to provide them with a specific diagnosis, it is important to assess the agreement between PTs and radiologists in using DMUS, with stricter diagnostic criteria. However, it would be interesting to further assess whether categorizing patients according to treatment strategy might be more valid and reliable, as this can have a great impact on daily clinical care.

Overall recommendations

Recommendations for clinical practice

A few guidelines are available at the moment to advice clinicians which diagnostic tools or PROMs should be used (general shoulder pain [54], subacromial pain syndrome [27, 55]). However, PROMs focussed on 'activity limitations' should be used to objectify the amount of disability, as this is often one of the main goals of treatment. If Dutch clini-

cians want to use a PROM to assess therapy effects on 'activity limitations' the SDQ could be used according to our review. The SPADI-D is a good alternative to measure activity in limitations as it is reliable, valid and responsive in patients with shoulder pain in primary care. Technicians and software developers could potentially facilitate clinicians in using the SPADI-D (and other PROMs), by letting the administrative system automatically calculate both a sum score and the minimal important change (while also addressing the implications of the measurement error).

Clinicians not using a PROM in their daily practice due to time-problems could use the substitute question in order to objectify, as well as to evaluate treatment success, although single items have the advantage of simplicity at the cost of detail [56].

DMUS at this time is not a recommended option as part of the diagnostic strategy for every day clinical care. Subgroup-analysis indicated both experience as the level of training impact the agreement, therefore clinicians interested in DMUS should invest in their clinical progression (additional courses, discussing the results with peers and radiologists). It is however of great importance to all PTs using DMUS to be aware of the (high) likelihood that the diagnostic results found by the individual PT do not match the conclusion of the radiologist in secondary care. Being critical about implementing tests is part of Evidence Based Practice and is advised by national bodies. In accordance, it is essential that the results of DMUS are not considered to be the absolute truth, as the agreement between the PT and radiologist is not high and the validity of DMUS when used by a PT is still unknown. However, as DMUS is not invasive and when used critically, it could be of value to the individual clinician interested in DMUS.

DMUS might, in the future, be used in a slightly different way; i.e. as a tool to facilitate the PT in determining if a patient possibly needs referral to secondary care or not. In case physiotherapy is indicated according to DMUS, it seems to be appropriate to assess functional limitations (range of motion, strength etc.) related to the needs of the patient, as both mobilization and exercises seem to be the main interventions in the entire physical therapy group. However, more research is needed to make solid statements with regards to DMUS used in the traditional and the suggested new way.

Recommendations for research

For research purposes, Core Outcome Sets (COS) are being developed at the moment to assure that future research at least includes a minimum set of PROMs in order to compare results. At the moment a group of international scientists, clinicians and patients is working on a COS for trials including patients with shoulder pain [57-60]. Until a COS for shoulder pain is published, we recommend using the SPADI. Although the SDQ is recommended in our review for Dutch users, the SPADI-D would be a good alternative. Mainly because the SPADI is validated in multiple languages and could therefore be

used internationally. The English SPADI is not only rated best in our review but is also frequently used in scientific research at this moment, making the SPADI-D a compatible choice in order to compare scientific results (e.g. baseline-score or outcome). At this time, DMUS is not advised to be part of a COS, as the inter-professional agreement is low.

Dutch researchers focussing on prognostic factors should use the SPADI-D (instead of the SDQ) as this has been proven to be a prognostic factor. It could however, be of value to assess the substitute question of the SPADI as well in further prognostic studies, as this would be of value to clinicians for practical reasons.

Research is most valuable when acknowledged by clinicians, as they are the ones needing to implement new evidence into their daily clinical care. It would be of interest to assess if the implementation of existing PROMs (such as the SPADI) could be influenced, when targeting the issues described by clinicians (e.g. a lack of knowledge) responsible for not implementing PROMs in their daily clinical care (e.g. providing them with an administrative system that helps them to interpret the (change)scores).

Moreover, based upon earlier research, creating a more individualized and time saving alternative for PROMs would assist in addressing some of these implementation-problems. New techniques such as CAT could be a great way to overcome e.g. the problem of time-burden. The PROMIS PFUE CAT could be a promising option to assess 'activity limitations' in patients with non-specific shoulder pain in a complete and timesaving way. This should be assessed a study including patients with non-specific shoulder pain visiting a PT in primary care.

With regards to DMUS a lot of research still needs to be done regarding its usefulness in primary physiotherapy care. For instance, it would be of interest to assess if the alternative way of using DMUS (using treatment related categories) impacts the agreement between the PT and radiologist (reliability study) with the concordant instructions for both the PT as the radiologist. Moreover, it would be interesting to assess whether this stratification impacts the outcome of the therapeutic process, the actual recovery and the cost-effectiveness.

Our conclusions regarding DMUS in the traditional way were based on one agreement study only, more research is needed to solidly confirm these findings before we can make definitive statements.

REFERENCES

1. Bot, S.D., et al., *Clinimetric evaluation of shoulder disability questionnaires: a systematic review of the literature.* Ann Rheum Dis, 2004. **63**(4): p. 335-41.
2. Roy, J.S., J.C. MacDermid, and L.J. Woodhouse, *Measuring shoulder function: a systematic review of four questionnaires.* Arthritis Rheum, 2009. **61**(5): p. 623-32.
3. St-Pierre, C., et al., *Psychometric properties of self-reported questionnaires for the evaluation of symptoms and functional limitations in individuals with rotator cuff disorders: a systematic review.* Disabil Rehabil, 2015: p. 1-20.
4. Mokkink, L.B., et al., *The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content.* BMC Med Res Methodol, 2010. **10**: p. 22.
5. Angst, F., et al., *Measures of adult shoulder function: Disabilities of the Arm, Shoulder, and Hand Questionnaire (DASH) and Its Short Version (QuickDASH), Shoulder Pain and Disability Index (SPADI), American Shoulder and Elbow Surgeons (ASES) Society Standardized Shoulder Assessment Form, Constant (Murley) Score (CS), Simple Shoulder Test (SST), Oxford Shoulder Score (OSS), Shoulder Disability Questionnaire.* Arthritis Care Res, 2011. **63**(SUPPL. 11): p. S174-S88.
6. Desai, A.S., A. Dramis, and A.J. Hearnden, *Critical appraisal of subjective outcome measures used in the assessment of shoulder disability.* Ann R Coll Surg Engl, 2010. **92**(1): p. 9-13.
7. Fayad, F., Y. Mace, and M.M. Lefevre-Colau, *[Shoulder disability questionnaires: a systematic review].* Ann Readapt Med Phys, 2005. **48**(6): p. 298-306.
8. Huang, H., et al., *A Systematic Review of the Psychometric Properties of Patient-Reported Outcome Instruments for Use in Patients With Rotator Cuff Disease.* Am J Sports Med, 2015. **43**(10): p. 2572-82.
9. Beaton, D.E., et al., *Guidelines for the process of cross-cultural adaptation of self-report measures.* Spine (Phila Pa 1976), 2000. **25**(24): p. 3186-91.
10. Wang, W.L., H.L. Lee, and S.J. Fetzer, *Challenges and strategies of instrument translation.* West J Nurs Res, 2006. **28**(3): p. 310-21.
11. Schellingerhout, J.M., et al., *Measurement properties of translated versions of neck-specific questionnaires: a systematic review.* BMC Med Res Methodol, 2011. **11**: p. 87.
12. Ropero Pelaez, F.J. and S. Taniguchi, *The Gate Theory of Pain Revisited: Modeling Different Pain Conditions with a Parsimonious Neurocomputational Model.* Neural Plast, 2016. **2016**: p. 4131395.
13. Turner, R.R., et al., *Patient-reported outcomes: instrument development and selection issues.* Value Health, 2007. **10 Suppl 2**: p. S86-93.
14. Reeve, B.B., et al., *ISOQOL recommends minimum standards for patient-reported outcome measures used in patient-centered outcomes and comparative effectiveness research.* Qual Life Res, 2013. **22**(8): p. 1889-905.
15. Beaton, D.E., J.G. Wright, and J.N. Katz, *Development of the QuickDASH: comparison of three item-reduction approaches.* J Bone Joint Surg Am, 2005. **87**(5): p. 1038-46.
16. Roach, K.E., et al., *Development of a shoulder pain and disability index.* Arthritis Care Res, 1991. **4**(4): p. 143-9.
17. L'Insalata, J.C., et al., *A self-administered questionnaire for assessment of symptoms and function of the shoulder.* J BONE JT SURG SER A, 1997. **79**(5): p. 738-48.
18. Lippitt, S.B., D.T. Harryman, and F.A. Matsen, *A practical tool for evaluation of function: the Simple Shoulder Test. In The Shoulder: a Balance of Mobility and Stability.* The American Academy of Orthopaedic Surgeons, 1993: p. pp. 501-518. .
19. Hudak, P.L., P.C. Amadio, and C. Bombardier, *Development of an upper extremity outcome measure: The DASH (disabilities of the arm, shoulder, and head).* AM J IND MED, 1996. **29**(6): p. 602-8.

20. Beaton, D.E., J.G. Wright, and J.N. Katz, *Development of the QuickDASH: comparison of three item-reduction approaches.* J Bone Joint Surg (Am), 2005. **87A**(5): p. 1038-1046.
21. Rose, M., et al., *Evaluation of a preliminary physical function item bank supported the expected advantages of the Patient-Reported Outcomes Measurement Information System (PROMIS).* J Clin Epidemiol, 2008. **61**(1): p. 17-33.
22. Hays, R.D., et al., *Upper-extremity and mobility subdomains from the Patient-Reported Outcomes Measurement Information System (PROMIS) adult physical functioning item bank.* Arch Phys Med Rehabil, 2013. **94**(11): p. 2291-6.
23. Anthony, C.A., et al., *Performance of PROMIS Instruments in Patients With Shoulder Instability.* Am J Sports Med, 2017. **45**(2): p. 449-453.
24. Minoughan, C.E., et al., *Correlation of PROMIS Physical Function Upper Extremity Computer Adaptive Test with American Shoulder and Elbow Surgeons shoulder assessment form and Simple Shoulder Test in patients with shoulder arthritis.* J Shoulder Elbow Surg, 2017.
25. Dogu, B., et al., *Which questionnaire is more effective for follow-up diagnosed subacromial impingement syndrome? A comparison of the responsiveness of SDQ, SPADI and WORC index.* J Back Musculoskelet Rehabil, 2013. **26**(1): p. 1-7.
26. Paul, A., et al., *A comparison of four shoulder-specific questionnaires in primary care.* Ann Rheum Dis, 2004. **63**(10): p. 1293-9.
27. Jansen, M.J., et al., *KNGF Evidence Statement Subacromiale klachten.* Nederlands Tijdschrift voor Fysiotherapie, 2011. **121**(1).
28. Breckenridge, J.D. and J.H. McAuley, *Shoulder Pain and Disability Index (SPADI).* J Physiother, 2011. **57**(3): p. 197.
29. Struyf, F., et al., *A Multivariable Prediction Model for the Chronification of Non-traumatic Shoulder Pain: A Systematic Review.* Pain Physician, 2016. **19**(2): p. 1-10.
30. Kuijpers, T., et al., *Systematic review of prognostic cohort studies on shoulder disorders.* Pain, 2004. **109**(3): p. 420-31.
31. Chester, R., et al., *Predicting response to physiotherapy treatment for musculoskeletal shoulder pain: a systematic review.* BMC Musculoskelet Disord, 2013. **14**: p. 203.
32. Thoomes-de Graaf, M., et al., *The Responsiveness and Interpretability of the Shoulder Pain and Disability Index.* J Orthop Sports Phys Ther, 2017: p. 1-21.
33. Wild, D., et al., *Principles of Good Practice for the Translation and Cultural Adaptation Process for Patient-Reported Outcomes (PRO) Measures: report of the ISPOR Task Force for Translation and Cultural Adaptation.* Value Health, 2005. **8**(2): p. 94-104.
34. Mokkink, L.B., et al., *COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures.* Qual Life Res, 2017.
35. Kooijman, M., et al., *Jaarcijfers 2010 en trendcijfers 2006-2010 fysiotherapie.* Landelijke Informatievoorziening Paramedische Zorg. Utrecht: NIVEL, http://www.nivel.nl/lipz.
36. Streiner, D.L. and G.R. Norman, *Health measurement scales a practical guide to the development and use.* . 2008: Oxford university press.
37. Hill, C.L., et al., *Factor structure and validity of the shoulder pain and disability index in a population-based study of people with shoulder symptoms.* BMC Musculoskelet Disord, 2011. **12**: p. 8.
38. Tveita, E.K., et al., *Factor structure of the Shoulder Pain and Disability Index in patients with adhesive capsulitis.* BMC Musculoskelet Disord, 2008. **9**.
39. Jamnik, H. and M.K. Spevak, *Shoulder pain and disability Index: Validation of slovene version.* Int J Rehabil Res, 2008. **31**(4): p. 337-41.

40. MacDermid, J.C., P. Solomon, and K. Prkachin, *The Shoulder Pain and Disability Index demonstrates factor, construct and longitudinal validity.* BMC Musculoskelet Disord, 2006. **7**: p. 12.
41. Roddey, T.S., et al., *Comparison of the University of California - Los Angeles Shoulder Scale and the Simple Shoulder Test with the Shoulder Pain and Disability Index: Single-administration reliability and validity.* Phys Ther, 2000. **80**(8): p. 759-68.
42. Jette, D.U., et al., *Use of standardized outcome measures in physical therapist practice: perceptions and applications.* Phys Ther, 2009. **89**(2): p. 125-35.
43. Swinkels, R.A., et al., *Current use and barriers and facilitators for implementation of standardised measures in physical therapy in the Netherlands.* BMC Musculoskelet Disord, 2011. **12**: p. 106.
44. Stratford, P.W. and J.M. Binkley, *Measurement properties of the RM-18. A modified version of the Roland-Morris Disability Scale.* Spine (Phila Pa 1976), 1997. **22**(20): p. 2416-21.
45. Beaton, D.E., et al., *Development of the QuickDASH: comparison of three item-reduction approaches.* J Bone Joint Surg Am, 2005. **87**(5): p. 1038-46.
46. Verwoerd, A.J., et al., *A single question was as predictive of outcome as the Tampa Scale for Kinesiophobia in people with sciatica: an observational study.* J Physiother, 2012. **58**(4): p. 249-54.
47. Karel, Y.H., et al., *Development of a Prognostic Model for Patients With Shoulder Complaints in Physiotherapy.* Phys Ther, 2016.
48. Martin, D.J., J.P. Garske, and M.K. Davis, *Relation of the therapeutic alliance with outcome and other variables: a meta-analytic review.* J Consult Clin Psychol, 2000. **68**(3): p. 438-50.
49. Welmers-van de Poll, M.J., et al., *Alliance and Treatment Outcome in Family-Involved Treatment for Youth Problems: A Three-Level Meta-analysis.* Clin Child Fam Psychol Rev, 2017.
50. Ferreira, P.H., et al., *The therapeutic alliance between clinicians and patients predicts outcome in chronic low back pain.* Phys Ther, 2013. **93**(4): p. 470-8.
51. Mallinckrodt, B. and Y.T. Tekie, *Item response theory analysis of Working Alliance Inventory, revised response format, and new Brief Alliance Inventory.* Psychother Res, 2016. **26**(6): p. 694-718.
52. Araujo, A.C., et al., *Measurement properties of the Brazilian version of the Working Alliance Inventory (patient and therapist short-forms) and Session Rating Scale for low back pain.* J Back Musculoskelet Rehabil, 2017. **30**(4): p. 879-887.
53. Babatunde, F., J. MacDermid, and N. MacIntyre, *Characteristics of therapeutic alliance in musculoskeletal physiotherapy and occupational therapy practice: a scoping review of the literature.* BMC Health Serv Res, 2017. **17**(1): p. 375.
54. Winters, J.C., et al., *NHG-Standaard Schouderklachten (Tweede herziening).* Huisarts Wet 2008(2008:51(11):555-565).
55. Diercks, R., et al., *Guideline for diagnosis and treatment of subacromial pain syndrome: a multidisciplinary review by the Dutch Orthopaedic Association.* Acta Orthop, 2014. **85**(3): p. 314-22.
56. Sloan, J.A., et al., *Assessing the clinical significance of single items relative to summated scores.* Mayo Clin Proc, 2002. **77**(5): p. 479-87.
57. Page, M.J., et al., *Identifying a core set of outcome domains to measure in clinical trials for shoulder disorders: a modified Delphi study.* RMD Open, 2016. **2**(2): p. e000380.
58. Buchbinder, R., et al., *A Preliminary Core Domain Set for Clinical Trials of Shoulder Disorders: A Report from the OMERACT 2016 Shoulder Core Outcome Set Special Interest Group.* J Rheumatol, 2017. **44**(12): p. 1880-1883.
59. Gagnier, J.J., et al., *Creation of a core outcome set for clinical trials of people with shoulder pain: a study protocol.* Trials, 2017. **18**(1): p. 336.

60. Page, M.J., et al., *Outcome Reporting in Randomized Trials for Shoulder Disorders: Literature Review to Inform the Development of a Core Outcome Set.* Arthritis Care Res (Hoboken), 2018. **70**(2): p. 252-259.

SUMMARY

CHAPTER 10.1

Chapter 1 is an introduction of shoulder pain and its diagnostics and presents the aims of this thesis.

Chapter 2 describes a systematic review evaluating the measurement properties of both the original versions as well as the translated versions of self-administered Patient Reported Outcome Measures (PROMs) focusing on the shoulder assessing "activity limitations" for patients with nonspecific shoulder pain, using the COSMIN checklist.
The search strategy resulted in a total of 3421 hits. We included 31 articles, evaluating 7 different Patient Reported Outcome Measures (PROMs). The Shoulder Pain and Disability Index (SPADI) was the most frequently evaluated PROM. None of the self-administered shoulder specific PROMs received strong or moderate evidence for all measurement properties in any language. PROMs in other languages than English, Dutch or Norwegian only received an 'unknown', 'poor' or 'limited' evidence score on one or more measurement properties.

The SPADI was recommended for English, Norwegian and Turkish users, although caution is advised for Turkish users, due to the limited available measurement properties. For Dutch users, the Shoulder Disability Questionnaire (SDQ) and Simple shoulder test (SST) were recommended. The Dutch SST showed strong evidence for the internal consistency and construct validity, moderate evidence for hypothesis testing and limited evidence for the reliability. The Dutch SDQ showed limited evidence for construct hypothesis testing and moderate evidence for responsiveness. We recommend choosing between either the SST or the SDQ depending on the purpose of its use. As the SPADI is the most widely used PROM and has the best ratings in several languages, it would be useful to further assess the Dutch version for both clinical and research purposes.

Chapter 3 describes the validation process of the Dutch SPADI (SPADI-D) and the assessment of reliability. Patients with shoulder pain were recruited from primary care physiotherapy clinics. At baseline patients received an online questionnaire that included the SPADI-D, SDQ and EuroQol five-item quality of life questionnaire (EQ-5D-3L). A randomly selected group of patients received a second SPADI-D after 1 week.

A total of 356 patients and a randomly selected group of 74 subjects for the reliability analysis were included. There was a significant difference between extreme groups (a high/low level of pain and work absence/presence) in SPADI score. This means that the SPADI-D is able to differentiate between different groups. The convergent validity of the SPADI-D was good, as the Spearman correlation between the SPADI-D and the SDQ was 0.69. Divergent testing resulted in a Spearman correlation of 0.25 with the EQ5D mobility-item and 0.14 with the depression-item, indicating the SPADI-D and EQ-5D-3L measure a different construct. We considered the SPADI-D to consist of one factor according to principal component factor analysis. As parallel analysis revealed that the

eigenvalue of the first factor should be above 1.44 and of the second factor above 1.33 to be extracted. Only one factor was extracted, the eigenvalue of the second factor was 0.97. A one-factor solution explained 57.9% of the variance and the second factor added only 7%. Findings were consistent with all five analyses based on two random subsamples. The internal consistency was high (Cronbach's alpha = 0.94 for the total score), and the test–retest reliability was good (ICC = 0.89-0.90).

We therefore considered the SPADI-D as a valid and reliable questionnaire for patients in primary care for assessing functional disability.

Chapter 4 evaluates the inter-professional agreement of diagnostic ultrasound between physiotherapists and radiologists in patients with shoulder pain for full thickness tears, partial thickness tear, calcification and subacromial bursitis. Next it describes if experience or training of the physiotherapist influences the overall agreement. A priori, substantial or high agreement was considered to be an appropriate norm. Patients were recruited from primary physiotherapy care and were excluded when diagnostic imaging had been performed in the previous three months.
In this study, a total of 13 physiotherapists trained and experienced in the use of diagnostic ultrasound and 9 experienced musculoskeletal radiologists participated. Patients were assessed in a usual physiotherapeutic manner, of which diagnostic ultrasound could be a part. A total of 65 patients participated and received a diagnostic ultrasound of their physiotherapist and a second diagnostic ultrasound of the radiologist within one week. The overall kappa of all four main diagnostic categories was 0.36 (95%CI 0.29-0.43), indicating fair agreement. The overall observed agreement, based on these four categories, was 80%, the specific positive agreement was 51% and the specific negative agreement was 86%. The kappa for the full thickness tear category was 0.63, indicating borderline substantial agreement. We found moderate agreement (0.54) for bursitis, fair agreement (0.28) for calcification, and slight agreement (0.10) for partial thickness tears. Subgroup analysis showed an overall kappa in the more experienced group of 0.43 (moderate) compared to a kappa of 0.17 (slight) in the less experienced group. Furthermore, we found a kappa of 0.43 (moderate) in the advanced course group compared to of 0.09 (slight) in the basic course group.

We concluded diagnostic ultrasound should not be recommended to be integrated into diagnostic clinical practice yet. However, there might be a possible added value of diagnostic ultrasound in the future of the physiotherapy profession, based upon subgroup analysis. At this moment, however, conclusions based on the results of the diagnostic ultrasound of the physiotherapist only, should be interpreted with caution.

Chapter 5 describes the exploration of the inter-professional agreement of diagnostic ultrasound between physiotherapists and radiologists using new treatment related categories using previous data. These new diagnostic labels were developed based on effective treatment strategies. A priori, a kappa higher than 0.70 was considered to be appropriate.

A literature search was performed to assess which traditional diagnostic labels could be recoded into new treatment related categories, resulting in 32 useful articles. The 'full thickness tear', 'biceps tendon tear' and 'SLAP- lesion' were labelled to: 'referral to secondary care'. Here, it is important that the patient is referred to a medical doctor to perform additional diagnostic tests and/or to discuss operative possibilities. 'Calcification', 'tendinopathy' and 'partial tear' of the rotator cuff, 'subacromial impingement' and 'bursitis' were labelled as an 'indication for physiotherapy'. All others ('arthritis/ arthrosis of the AC-joint', 'calcification' and 'tendinopathy' of the biceps and 'no pathology') were labelled as 'watchful waiting'. The overall kappa was 0.60 (95%CI 0.43-0.76), indicating these new treatment related categories showed moderate agreement between physiotherapists and radiologists. There was substantial agreement regarding the new diagnostic label 'referral to secondary care' (k=0.74) and both 'possible indication for physiotherapy management' (k=0.57) and 'watchful waiting' (k=0.46) showed moderate agreement.

Although the agreement did not reach a kappa value of 0.70, we considered this approach to be a promising option to further assess in future research. Diagnostic ultrasound might be used at first consultation to facilitate physiotherapists in making decisions regarding the appropriateness of the consultation in the future.

Chapter 6 evaluates if the working alliance inventory short-form (WAV-12) is a valid measurement instrument in terms of the construct and discriminative abilities for a population of patients with shoulder pain in physiotherapy care.

A total of 389 patients were enrolled by 66 physiotherapists, of which 274 patients filled in one or more items of the WAV-12. A large number of patients only completed a limited number of items and just a small percentage (22%; 78 patients) filled in the complete WAV-12. Compared to the response rate of other questionnaires send at 6 weeks in our 'Shoulder Complaints and Diagnostic Ultrasound in Physiotherapy' (ShoCoDiP)- cohort study, this was remarkable. Ceiling effects were observed in ten out of twelve items. A partial credit RASCH analysis revealed the items have good discriminative abilities in the lower end of the construct, as the item information curve showed the amount of information given by the questionnaire is highest between an ability of −2 and 0. A pooled Cronbach's alpha coefficient based upon five imputed datasets was high (0.89), indicating items are highly correlated and measure the same explanatory concept. No differential item functioning was found on gender and age. Validity for the items

in the questionnaire appears to be sound, but due to the difference in the percentage of missing data among the items and observed ceiling effects, we advised linguistic (Dutch) and contextual (physiotherapeutic setting) adjustments. Therefore, a Delphi study, using a two- round survey, including 11 panel members (6 experts and 5 patients) was performed. The panels opinion on the adjustments in the WAV-12 resulted in a new specific version, the Physio Alliance Scale (PAS). We concluded the WAV-12 is not appropriate to be implemented into a Dutch physiotherapy setting, and the adjusted form has not been tested yet.

Chapter 7 evaluates the measurement error, interpretability and responsiveness of the SPADI-D on patients with shoulder pain seeking help by a physiotherapist in primary care setting.

A total of total of 356 patients participated at baseline and 237 were included in the analysis using data at 26 weeks. Participating physiotherapists used a variety of shoulder diagnoses to label the patients; however, the majority of patients were labelled as having subacromial impingement. The physiotherapists also used a variety of treatment techniques, mainly including advice, exercise, and mobilization/manipulation of the shoulder or thoracic spine and the majority of patients (59.5%) completed therapy sessions after 12 weeks. A total of 139 patients were considered recovered. The SPADI showed no signs of floor and ceiling effects. The minimal important change (MIC) was 20 points, resulting in a change of 42.8% of the baseline score. The sensitivity and specificity were both 0.75. Subgroup analysis resulted in similar results: the MIC for patients with a high baseline score was 43.0% (27.9 points), with a sensitivity of 0.82 and specificity of 0.77, and for patients with a low baseline score was 42.7% (12.2 points), with a sensitivity of 0.81 and specificity of 0.82. The smallest detectable change (SDC) was 19.7. The responsiveness was good, as the Area Under the Curve (AUC) was 0.81 (with a 95% confidence interval ranging from 0.75 to 0.87) and hypothesis testing for responsiveness resulted in a Spearman correlation between the SPADI-D change score and the Global Perceived Effect scale (GPE) scale of 0.53. The Pearson correlation between the SPADI-D change score and the SDQ change score was 0.71. The Spearman correlation between the change score of the SPADI-D and the EQ-5D-3L depression item was 0.06 and the EQ-5D-3L mobility item was 0.12.

This study shows that the SPADI-D is responsive, making it a useful evaluative instrument to assess functional disability in longitudinal studies in patients with shoulder pain visiting a physical therapist. The SPADI-D can detect important changes. A change larger than 43% from the baseline score is considered to be a clinically relevant and important change. However, the measurement error should be taken into account when used for decision making in individual patients.

Chapter 8 describes the development of a single substitute question for the SPADI and the evaluation of its convergent/divergent validity, responsiveness and predictive power as this might be helpful to integrate a PROM into clinical practice.

In a meeting with the ShoCoDiP-project team (consisting of physiotherapists, manual therapists, general practitioners, a radiologist, an orthopedic surgeon and epidemiologists) various items were discussed that could act as a substitute question to cover the entire domain of the SPADI questionnaire. The final substitute question was chosen based on consensus within the research team: "Please state the amount of limitation in daily activity you experience due to your shoulder pain". This question could be answered on an 11-point scale, where: 0 = no limitation at all and 10 = completely disabled". The predictive power was assessed using predictive factors from the literature.

A total of 356 patients were included and 250 were included in the predictive power analysis as they completed the GPE after 26 weeks and answered all items of interest at baseline (age, duration of complaints, Numeric Rating Scale (NRS) and the SPADI according to the missing item criteria and the substitute question). Responsiveness was based on 237 patients answering the substitute question at baseline and follow up and the GPE-scale.

Convergent validity was confirmed, as the Spearman correlation coefficient was high between the substitute question and the total SPADI (0.74) and substantial with the SDQ (0.59). The spearman correlation between the substitute question and the mobility-item of the EQ-5D-3L was 0.23 and with the item anxiety/depression 0.20, indicating the instruments measure a different construct than the substitute question. Differences between "known groups" were statistically significant (a high/low level of pain and work absence/presence). Responsiveness was considered to be good, as the AUC was 0.76 (CI 95% 0.70 to 0.83) and hypothesis testing was confirmed. The Spearman correlation between the SPADI-D change score and the substitute change score was 0.71 and 0.60 with the SDQ change score. The Spearman correlation between the GPE and the substitute question was 0.47. A low correlation was found between the substitute question and both the mobility as the anxiety/depression item of the EQ-5D-3L (0.10). The predictive power of the substitute question is not comparable to the complete SPADI as the Chi Square test for adding the substitute question was not significant in the model based on the literature (p=0.193) as well as for the model including factors based upon our prospective cohort study (p=0.501) as opposed to the SPADI. The odds of the SPADI and the substitute question were quite exchangeable, however the confidence interval of the substitute question was wider. All models showed poor discrimination and the AUC values were within the 95%CI intervals of each other.

We concluded the substitute question might be an appropriate tool to replace the SPADI, especially for clinicians not using a PROM due to time burden reasons. However, more research is needed.

Chapter 9 reflects on the findings of this thesis and recommendations for clinicians and researchers are presented.

SAMENVATTING

CHAPTER 10.2

Hoofdstuk 1 leidt het thema in, schouderpijn en het diagnostische proces, en omschrijft de rationale ten aanzien van de doelstellingen van dit proefschrift.

Hoofdstuk 2 beschrijft een systematische review gericht op de evaluatie van meeteigenschappen m.b.v. de COSMIN, van zowel originele als vertaalde versies van zelf-gerapporteerde schouder-gerelateerde vragenlijsten gericht op "beperkingen in activiteiten" bij patiënten met a- specifieke schouderpijn. Het literatuuronderzoek resulteerde in een totaal van 3421 unieke artikelen, waarvan er 31 zijn geïncludeerd welke 7 verschillende vragenlijsten evalueren. De Shoulder Pain and Disability Index (SPADI) was de meest geëvalueerde vragenlijst. Geen van de zelf-gerapporteerde schouder specifieke vragenlijsten (in welke taal dan ook) ontving sterk of redelijk bewijs voor alle meeteigenschappen. Vragenlijsten in een andere taal dan het Engels, Nederlands of Noors ontvingen een score 'onbekend' 'slecht' of 'beperkt' op één of meerdere meeteigenschappen.

De SPADI wordt aanbevolen voor Engels, Noorse of Turkse gebruikers, maar voorzichtigheid is geboden voor Turkse gebruikers, aangezien er weinig bekend is t.a.v. de meeteigenschappen. Nederlandse gebruikers kunnen gebruik maken van de Shoulder Disability Questionnaire (SDQ) en de Simple shoulder test (SST). Er is sterk bewijs voor de interne consistentie en de construct validiteit van de Nederlandse SST, redelijk bewijs voor construct validiteit m.b.v. het testen van hypotheses en beperkt bewijs voor de betrouwbaarheid. Er is beperkt bewijs voor de construct validiteit (hypothese testen) van de Nederlandse SDQ en redelijk bewijs voor de responsiviteit. We raden aan om tussen de SST en de SDQ te kiezen op basis van het beoogde doel. Het zou zinnig zijn om de Nederlandse versie van de SPADI te onderzoeken, aangezien de SPADI de meest gebruikte vragenlijst is en wordt aanbevolen in meerdere talen voor zowel de kliniek als voor onderzoek.

Hoofdstuk 3 beschrijft het validatieproces en het testen van de betrouwbaarheid van de Nederlandse SPADI (SPADI-D). Patiënten met schouderpijn werden gerekruteerd in eerstelijns fysiotherapiepraktijken. Alle patiënten ontvingen bij aanvang een online vragenlijst, waar de SPADI-D, SDQ en de EuroQol five-item quality of life questionnaire (EQ-5D-3L) onderdeel van waren. Een random geselecteerde groep patiënten ontving een tweede SPADI-D na een week. In totaal konden 356 patiënten worden meegenomen in de analyse en 74 patiënten in de betrouwbaarheidsanalyse. Er was een significant verschil in de SPADI- score tussen de extreme groepen (een hoog/laag pijnniveau en wel/geen werkverzuim). Dit betekent dat de SPADI-D in staat is te differentiëren tussen verschillende groepen. De convergent validiteit van de SPADI-D is goed, aangezien de Spearman correlatie tussen de SPADI-D en de SDQ 0.69 is. Het testen van de divergent validiteit resulteerde in een Spearman correlatie van 0.25 met het EQ5D mobiliteits-item en 0.14 met het depressie item, dit betekent dat de SPADI-D en de items van de EQ5D

verschillende constructen meten. De SPADI-D wordt door ons gezien als een vragenlijst bestaande uit één factor, gebaseerd op principal component factor analyse. Parallel analyse liet zijn dat de eigenvalue van de eerste factor boven de 1.44 moest zijn en de tweede factor moest boven de 1.33 zijn om te worden geëxtraheerd. Slechts één factor kon worden geëxtraheerd, aangezien de tweede factor 0.97 was. Een "één factor- verklaring" verklaarde 57.9% van de variatie en de tweede factor voegde slechts 7% toe. De bevindingen waren consistent bij alle vijf de analyses gebaseerd op twee random subsamples. De interne consistentie was hoog (Cronbach's alpha= 0.94 voor de hele schaal) en de test-hertest betrouwbaarheid was goed (ICC= 0.89-0.90). Op basis van onze gevonden resultaten concluderen wij dat de SPADI-D een valide en betrouwbare vragenlijst is voor patiënten met schouderklachten in de eerste lijn waarbij functionele beperkingen in kaart worden gebracht.

Hoofdstuk 4 beschrijft de interprofessionele overeenstemming van echografische diagnostiek tussen fysiotherapeuten en radiologen over een volledige ruptuur, partiele ruptuur, calcificaties en subacromiale bursitis bij patiënten met schouderpijn. Vervolgens beschrijft het of ervaring of opleiding bij fysiotherapeuten de overall overeenstemming beïnvloed. A-priori werd gesteld dat substantiële of hoge overeenstemming adequaat was. Patiënten werden gerekruteerd uit eerstelijns fysiotherapiepraktijken en werden geëxcludeerd als er in de afgelopen drie maanden al eerder echografische diagnostiek had plaatsgevonden.

In totaal namen in deze studie 13, in het gebruik van diagnostische musculoskeletale echografie geschoolde en ervaren, fysiotherapeuten en 9 ervaren musculoskeletale radiologen deel. Patiënten werden op een normale manier fysiotherapeutisch onderzocht, waarvan musculoskeletale echografische diagnostiek een onderdeel kon uitmaken. In totaal namen 65 patiënten deel en kregen een diagnostische echo bij de fysiotherapeut en binnen een week een tweede diagnostische echo bij de radioloog.

De overall kappa van de vier diagnostische hoofdcategorieën was 0.36 (95%CI 0.29-0.43), hetgeen matige overeenstemming betekent. De overall geobserveerde overeenkomst, gebaseerd op deze 4 categorieën was 80%, de specifieke positieve overeenstemming was 51% en de specifieke negatieve overeenstemming was 86%. De kappa voor de categorie 'volledige ruptuur' was 0.63 ofwel voldoende tot goede overeenstemming. Wij vonden redelijke overeenstemming (0.54) voor bursitis, matige overeenstemming (0.28) voor calcificatie en geringe overeenstemming (0.10) voor partiële rupturen. Subgroep analyse liet een overall kappa van 0.43 (redelijk) in de meer ervaren groep zien in vergelijking met een kappa van 0.17 (matig) in de minder ervaren groep. Daarnaast vonden wij een kappa van 0.43 (redelijk) bij de groep fysiotherapeuten die vervolgcursussen hadden gedaan ten opzichte van 0.09 (matig) in de groep die alleen een basiscursus hadden gedaan. Wij concludeerden dat het nog niet aangeraden wordt

om diagnostische musculoskeletale echografie in de klinische praktijk toe te passen. Echter, op basis van subgroep analyse zou er zou in de toekomst mogelijk toegevoegde waarde van diagnostische musculoskeletale echografie in de fysiotherapie kunnen zijn. Op dit moment echter, moeten de conclusies van de diagnostische musculoskeletale echografie van de fysiotherapeut met voorzichtigheid worden geïnterpreteerd.

Hoofdstuk 5 beschrijft de interprofessionele overeenstemming van diagnostische echografie tussen fysiotherapeuten en radiologen waarbij nieuwe behandel gerelateerde categorieën zijn gebruikt. Deze nieuwe diagnostische labels werden gebaseerd op de effectieve behandelstrategie. Voor deze studie is de data van hoofdstuk 3 gebruikt. A-priori werd een kappa hoger dan 0.70 als passend beschouwd.

Er werd een literatuurstudie gedaan waardoor de traditionele labels in nieuwe behandel gerelateerde categorieën konden worden her-labeld, welke resulteerde in 32 bruikbare artikelen. De 'volledige ruptuur', 'bicepspees ruptuur' en 'SLAP-laesie' werden gelabeld als: 'doorverwijzing naar tweedelijns zorg'. Hierbij is het van belang dat de patiënt doorverwezen wordt naar een arts voor toegevoegde (beeldvormende) diagnostiek en/of om chirurgische ingrepen te overwegen. 'Calcificatie", 'tendinopathie' en 'partiële ruptuur' van de rotator cuff werden gelabeld als 'indicatie voor fysiotherapie'. Alle anderen ('artritis/ artrose van het AC-gewricht', 'calcificatie' en 'tendinopathie' van de biceps en 'geen pathologie') werden gelabeld als 'afwachtend beleid'. De overall kappa was 0.60 (95%CI 0.43-0.76), hetgeen aangeeft dat deze nieuwe diagnostische labels redelijke overeenstemming liet zien tussen fysiotherapeuten en radiologen. Er was voldoende tot goede overeenstemming (k=0.74) binnen het nieuwe label 'doorverwijzing naar tweedelijns zorg' en zowel 'indicatie voor fysiotherapie' (k=0.57) als 'afwachtend beleid' (k=0.46) lieten redelijke overeenstemming zien. Alhoewel de overeenstemming niet de kappa waarde van 0.70 bereikte in de gebruikte dataset, menen wij toch dat deze aanpak een veelbelovende optie voor toekomstig onderzoek kan zijn. Diagnostische echografie zou bij een eerste onderzoek gebruikt kunnen worden om de fysiotherapeut te ondersteunen bij de indicatiestelling van toekomstige consulten.

Hoofdstuk 6 beschrijft dat de 'working alliance inventory short-form' (WAV-12) een valide instrument is in termen van construct en onderscheidende mogelijkheden voor groepen van patiënten met schouderpijn in de fysiotherapeutische praktijk.

In totaal werden 389 patiënten door 66 fysiotherapeuten geïncludeerd, waarvan 274 patiënten één of meer items van de WAV-12 invulden. Een groot aantal patiënten vulde maar een beperkt aantal items in en slechts een klein percentage (22%, 78 patiënten) vulde de WAV-a12 volledig in. In vergelijking met het responspercentage van andere vragenlijsten die na 6 weken werden verzonden in onze 'Shoulder Complaints and Diagnostic Ultrasound in Physiotherapy' (ShoCoDiP)- cohort studie, was dit opmerkelijk.

Plafond effecten werden waargenomen in tien van de twaalf items. Een partiële RASCH analyse onthulde dat de items een goede onderscheidende mogelijkheid in het laagste einde van het construct hebben, aangezien de item informatiecurve liet zien dat de hoeveelheid informatie gegeven door de vragenlijst het hoogst is tussen een mogelijkheid van -2 en 0. Een gepoolde Cronbach's alpha coëfficiënt gebaseerd op vijf geïmputeerde datasets was hoog (0.89), wat aangeeft dat items hoog gecorreleerd zijn en hetzelfde verkennende/ informatieve concept meten. Er werd geen 'differential item functioning' gevonden op leeftijd en geslacht. De validiteit van de items in de vragenlijst lijkt solide, maar door het verschil in het percentage ontbrekende data onder de items en de waargenomen plafond effecten, adviseren wij taalkundige (Nederlandse) en contextuele (fysiotherapiepraktijk) aanpassingen. Daartoe werd een Delphi studie uitgevoerd, middels een survey van twee rondes met gebruikmaking van 11 panelleden (6 experts en 5 patiënten). De mening van het panel over de aanpassingen op de WAV-12 resulteerde in een nieuwe specifieke versie, de Physio Alliance Scale (PAS). Wij concludeerden dat de WAV-12 niet toegepast kan worden in een Nederlandse fysiotherapie setting. De aangepaste versie (PAS) is nog niet getest op methodologische kwaliteit en bruikbaarheid waardoor geschiktheid van deze vragenlijst binnen de fysiotherapie onduidelijk is.

Hoofdstuk 7 beschrijft de meetfout, interpreteerbaarheid en responsiviteit van de SPADI-D bij patiënten met schouderpijn die hulp zoeken bij fysiotherapeuten in een eerstelijnspraktijk.

In totaal werden 356 patiënten geïncludeerd op baseline en 237 patiënten geïncludeerd in de analyse met gebruikmaking van de data op 26 weken. Deelnemende fysiotherapeuten gebruikten een variëteit van schouderdiagnoses om de patiënten te labelen; de meeste patiënten werden echter gelabeld als een 'subacromiale impingement' hebbend. De fysiotherapeuten gebruikten ook een variëteit aan behandeltechnieken, die vooral adviezen, oefeningen en mobilisatie/manipulatie van de schouder of thoracale wervelkolom bevatten en het merendeel van de patiënten (59.5%) sloten hun therapiesessies af na 12 weken.

In totaal werden 139 patiënten als hersteld beschouwd. De SPADI liet geen tekenen van 'floor en ceiling effects' zien. De 'minimal important change' (MIC) was 20 punten, wat resulteerde in een verandering van 42.8% van de baseline score. De sensitiviteit en specificiteit waren allebei 0.75. Subgroep analyse liet vergelijkbare resultaten zien: de MIC voor patiënten met een hoge baseline score was 43.0% (27.9 punten), met een sensitiviteit van 0.82 en specificiteit van 0.77en voor patiënten met een lage baseline score was de MIC 42.7% (12.2 punten) met een sensitiviteit van 0.81 en specificiteit van 0.82. De 'smallest detectable change' (SDC) was 19.7. De responsiviteit was goed, aangezien de 'Area Under the Curve' (AUC) 0.81 was (met een 95% betrouwbaarheidsinterval tussen 0.75 en 0.87) en hypothese testen voor de responsiviteit resulteerde in een Spearman

correlatie tussen de SPADI-D veranderscore en de 'Global Perceived Effect schaal' (GPE) van 0.53. De Pearson correlatie tussen de SPADI-D veranderscore en de SDQ veranderscore was 0.71. De Spearman correlatie tussen de veranderscore van de SPADI-D en de EQ-5D-3L depressie item was 0.06 en de EQ-5D-3L mobiliteit item was 0.12.

Deze studie toont aan dat de ~SPADI-D responsief is, wat het een bruikbaar evaluatief instrument maakt om functionele beperkingen te evalueren bij patiënten met schouderpijn die een fysiotherapeut consulteren. De SPADI-D kan belangrijke veranderingen waarnemen. Een verandering groter dan 43% vanaf de baseline wordt beschouwd als een klinisch relevante en belangrijke verandering. De meetfout moet echter wel in beschouwing worden genomen wanneer de SPADI-D gebruikt wordt bij beslissingen bij individuele patiënten.

Hoofdstuk 8 beschrijft de ontwikkeling van een enkele vraag ter vervanging van de SPADI en de evaluatie van zijn convergente/ divergente validiteit, responsiviteit en voorspellende waarde aangezien dit behulpzaam kan zijn om een PROM in de klinische praktijk te integreren.

Tijdens een bijeenkomst van het ShoCoDiP-project team (bestaande uit fysiotherapeuten, manueel therapeuten, huisartsen, een radioloog, een orthopedisch chirurg en epidemiologen) werden verschillende items besproken die als een vervangende vraag zouden kunnen dienen om het gehele domein, dat de SPADI vragenlijst bestrijkt, afdekt. De uiteindelijke vervangende vraag werd gekozen gebaseerd op consensus binnen het onderzoeksteam: "*Geef aan in welke mate u beperkt bent in uw dagelijks functioneren door uw schouderklacht".*

Deze vraag kon worden beantwoord op een 11-puntsschaal, waarbij 0 = 'geen enkele beperking' en 10 = 'volledig beperkt'. De voorspellende waarde werd beoordeeld met gebruikmaking van voorspellende waardes uit de literatuur.

In totaal werden 356 patiënten geïncludeerd en werden 250 patiënten geïncludeerd in de voorspellende waarde analyses, aangezien zij de GPE na 26 weken ingevuld hadden en zij alle benodigde items (leeftijd, klachtenduur, Numerieke Rating Scale (NRS) en de SPADI overeenkomstig de missing item criteria en vervangvraag) bij aanvang beantwoord hadden.

De responsiviteit was gebaseerd op 237 patiënten die de vervangvraag beantwoord hadden bij aanvang en bij follow-up en de GPE-schaal

Convergente validiteit werd bevestigd aangezien de Spearman correlatiecoëfficiënt hoog was tussen de vervangvraag en de totale SPADI (0.74) en substantieel met de SDQ (0.59)

De Spearman correlatie tussen de vervangvraag en het mobiliteitsitem van de EQ-5D-3L was 0.23 en met het item angst/ depressie 0.20, wat aangeeft dat deze instrumenten een ander construct meten dan de vervangvraag. Verschillen tussen "bekende groepen"

waren statistisch significant (een hoog/laag niveau van pijn en werkverzuim). De responsiviteit werd als goed beoordeeld, aangezien de AUC 0.76 (CI 95% 0.70 tot 0.83) was en hypothese testen werd bevestigd. De Spearman correlatie tussen de SPADI-D veranderscore en de vervangvraag score was 0.71 en 0.60 met de SDQ veranderscore. De Spearman correlatie tussen de GPE en de vervangvraag was 0.47. Een lage correlatie werd gevonden tussen de vervangvraag en zowel de mobiliteits- als de angst/depressie items van de EQ-5D-3L (0.10)

De voorspellende waarde van de vervangvraag is niet vergelijkbaar met de complete SPADI, aangezien de Chi Square test om de vervangvraag toe te voegen niet significant was in zowel het model gebaseerd op de literatuur (p=0.139) als ook in het model met factoren uit onze eigen prospectieve cohortstudie (p=0.501) tegenover de SPADI. De odds van de SPADI en de vervangvraag waren uitwisselbaar, alleen was het betrouwbaarheidsinterval van de vervangvraag breder. Alle modellen lieten matige onderscheidenheid zien en de AUC-waarden waren binnen de 95% betrouwbaarheidsintervallen van elkaar.

Wij concluderen dat de vervangvraag een geschikt instrument kan zijn om de SPADI te vervangen, vooral voor clinici die geen PROM gebruiken vanwege redenen van tijdsgebrek. Meer onderzoek is echter nog wel nodig.

Hoofdstuk 9 is een reflectie op de bevindingen in dit proefschrift, waarin tevens aanbevelingen worden gedaan voor zowel clinici als onderzoekers**.**

DANKWOORD

CHAPTER 11

Het laatste gedeelte van het boekje, zowel gedrukt als in procesmatige zin. Het voelt dan ook een beetje gek om het woord "dankwoord" op te schrijven, omdat dit betekent dat het einde van dit proces nu echt in zicht is. Er zijn veel mensen die ik zou willen bedanken; collega's, familie en vrienden, waarvan ik er een aantal specifiek wil benoemen.

Graag zou ik willen beginnen bij Arianne, mijn co-promotor. Voor mij was de wetenschap een hele nieuwe wereld en jij hebt me door dit proces heen 'geloodst'. Ik wil je heel graag bedanken voor je steun, je betrokkenheid, je gedrevenheid en je altijd rake feedback. Gedurende dit proces is mijn leven veranderd en jij hebt daar flexibel en prettig op geanticipeerd en altijd laten blijken dat je achter me stond. We hebben onze samenwerking gelukkig continue kunnen voortzetten, niet alleen toen het lectoraat ophield te bestaan maar ook nu je helemaal naar Australië bent geëmigreerd. Ik ben blij dat je er bent.

Wendy, je was mijn tutor op de SOMT en hebt me destijds gestimuleerd om voor het eerst hoorcollege te geven. Jij bent daarnaast degene die mij benaderd heeft voor het ShoCoDiP-project. Ook daar werd je mijn begeleider en heb ik allerlei nieuwe ervaringen opgedaan en daar wil ik je heel graag voor bedanken. De periode bij Avans was bijzonder, een heel nieuw lectoraat met nieuwe projecten. Het was een hele drukke, energieke tijd helemaal passend bij jouw karakter ;). Je hebt altijd snel tijd vrij gemaakt om me van feedback te voorzien. Bedankt voor je begeleiding en bevlogenheid.

Bart, bedankt voor je rust en kalmte, je positivisme en je overstijgende commentaar. Je hebt me inzicht gegeven in het promotie-traject en me verzekerd dat het echt minder eng zou zijn dan ik vooraf dacht.

Uiteraard wil ik ook de voltallige commissie bedanken. Ik zou graag het hele projectteam willen bedanken, die betrokken zijn geweest bij de start van het ShoCoDiP-project (Maaike, Annechien, Eric, Marcel, Joost, Ad, Geert-Jan, Ramon, Bart, Arianne, Wendy, Yasmaine en Edwin). We hebben het voorrecht gehad om met verschillende partijen samen te werken aan dit project (Amphia ziekenhuis, verschillende fysiotherapienetwerken, de Universtiteit van Maastricht, het Erasmus Medisch Centrum en Avans hogeschool), waardoor we gebruik hebben kunnen maken van de diversiteit in achtergronden. Graag zou ik ook de subsidieverstrekker SIA RAAK willen bedanken. Uiteraard wil ik alle andere betrokkenen ook graag bedanken, de participerende fysiotherapeuten, radiologen en patiënten. Maar ook mijn co-auteurs wil ik heel graag bedanken, ik voel me vereerd dat jullie met mij hebben meegewerkt aan de artikelen en wil jullie heel graag bedanken voor de geleverde inspanningen en de meegebrachte expertise.

Graag wil ik van deze gelegenheid gebruik maken om ook Maarten en Emiel te bedanken die mij hebben geënthousiasmeerd voor de mogelijkheden die er zijn naast het werk in de klinische praktijk en specifiek m.b.t. de wetenschap.

Ik wil daarnaast mijn collega's bedanken, vanuit alle drie de hoeken die de afgelopen jaren voor mij van belang zijn geweest in mijn werk. De combinatie van onderzoek, lesgeven en de klinische praktijk heeft het voor mij tot een hele boeiende en mooie periode gemaakt. Yasmaine, met jou ben ik aan het ShoCoDiP -project begonnen en ik vind het enorm bijzonder dat we het ook echt tegelijk afronden. We hebben veel geleerd samen, o.a. dat de actieve bereidheid van collega's om patiënten te rekruteren wordt beïnvloed door het opzetten van gratis symposia. We gingen langs de ziekenhuizen met worstenbroodjes, chocolaatjes met ons logootje en hebben meerdere symposia opgezet. We hebben ook heel erg veel samen voor het eerst gedaan; een protocol schrijven, online vragenlijsten uitzetten, een database bouwen en opschonen enz. Fantastisch dat we nu ook voor het eerst promoveren, niet samen, maar wel tegelijk.
Mijn overige collega's op het lectoraat (Lieke, Dennis, Bert en Guus) wil ik bedanken voor een leuke enerverende periode. Maar ook mijn collega's van het docententeam en dan met name Nienke, die altijd geïnteresseerd is geweest en met wie ik heel erg prettig heb samengewerkt. Ook Jasper wil ik graag bedanken, het was zeer prettig dat jij vlak voor mij ging promoveren. Bedankt voor je vriendelijke woorden en je hulpvaardigheid. Uiteraard kan ik mijn collega's in de praktijk niet vergeten (Willem, Erik, Laurie, Roos, Annelies, Patricia en Milou) met wie ik met veel plezier iedere dag patiëntenzorg lever en waar we wetenschap een plekje (en handen en voeten) proberen te geven. De afgelopen jaren hebben we behoorlijk veel opgezet en geïnnoveerd en dat heeft ertoe geleid dat ik over dit proefschrift iets langer gedaan heb dan vooraf gepland, bedankt daarvoor ;). Zonder gekheid, ik ben trots op wat we hebben neergezet.

Uiteraard zijn er meer mensen die mij hebben beïnvloed (/geïnspireerd), hebben gesteund, me hebben laten lachen, hebben geholpen te relativeren, enz. Familie en vrienden, ik wil jullie daar heel graag voor bedanken. Graag wil ik mijn moeder in het bijzonder bedanken die regelmatig is bijgesprongen om Loena op te vangen. Dat heeft mij enorme rust en blijdschap gegeven, wetende dat zij op een hele fijne plek bivakkeerde als ik me even moest afzonderen. Uiteraard wil ik mijn paranimfen Loes en Erik graag bedanken, heel fijn dat jullie naast me staan. Doycke, ook jou wil ik graag bedanken voor het vastleggen van een speciaal moment. Als laatste wil ik graag Erik en Loena bedanken. Het was soms best lastig om tijd te vinden naast alle projecten en het runnen van een praktijk om te werken aan mijn promotie.

Erik, zonder jou had ik alle projecten tegelijk niet kunnen volbrengen. Je hebt me gesteund en je bent er altijd voor me geweest. Je hebt me geholpen om ook 'nee' te

kunnen zeggen tegen nieuwe projecten die me erg leuk leken en hebt me geholpen te laten zien dat het alleen maar écht leuk blijft als het behapbaar blijft. Jij en Loena zijn enorm begripvol geweest. Loena, lief klein meisje, wat ben je al groot in je doen en laten. Je hebt hele wijze en grappige dingen tegen me gezegd en je kwam me heel vaak even wat te drinken brengen als ik op kantoor aan het werk was. Lieverd, het wordt nu weer tijd om voorlopig op zondag lekker alleen maar dingen samen te doen!

CURRICULUM VITAE

Marloes Thoomes-de Graaf is geboren op 23 november 1984 in Alphen aan den Rijn. Marloes ging in 2004 fysiotherapie studeren in Leiden. Direct na het afronden van de opleiding fysiotherapie combineerde zij een nieuwe studie met het werken in twee particuliere praktijken. Marloes volgde het excellente-traject bij de SOMT tijdens haar master-opleiding manuele therapie. Dit resulteerde in haar eerste internationale publicatie. Zij startte in 2010 met de opleiding Evidence Based Practice aan de Universtiteit van Amsterdam en combineerde dit met een baan als docent op de opleiding fysiotherapie op Avans in Breda. Marloes werkte op deze opleiding als docent in de minor wervelkolom en heeft meegewerkt aan het herschrijven van het lesmateriaal voor deze minor. Naast haar baan als fysio- manueel therapeut werkte zij destijds niet alleen als docent bij Avans, maar was zij tevens werkzaam op het lectoraat. Vanuit het lectoraat diagnostiek is het wetenschappelijk onderzoek, waar dit proefschrift uit voortkomt opgestart in samenwerking met het Erasmus Medisch Centrum en de Universiteit van Maastricht. In 2012 heeft Marloes de graad "Master of Science" behaald door de opleiding klinische epidemiologie aan de UVA af te ronden. In 2014 is Marloes gestopt als docent aan de opleiding fysiotherapie, en heeft zij haar promotieonderzoek vervolgd aan de Erasmus Universiteit. Zij heeft sindsdien aan diverse projecten deelgenomen, van onder andere de Nederlandse Vereniging Manuele Therapie en schrijft onderwijsmateriaal voor bij- en nascholingen. Sinds 2012 is Marloes ook praktijkhouder van een gespecialiseerde fysiotherapiepraktijk in Hazerswoude-Rijndijk en combineerde ze dit werk met haar promotieonderzoek.

PORTFOLIO

Courses

"Systematic review of clinimetrical instruments (using the COSMIN)" VU University Amsterdam (VUMC), dept. Epidemiology & Biostatistics EMGO Institute for Healthcare Research, 2013	1 ECT
"Systematic searching in PubMed and other electronic databases" Erasmus University, 2013	1 ECT
"Advanced use of Endnote as an electronic library reference system" Erasmus University, 2013	1 ECT
Clinical epidemiology, MSc, University of Amsterdam, 2010-2012	98 ECTS
Didactical training, AVANS university and SOMT University, 2011	2 ECTS
	103 ECTS
Mike Stewart: known pain workshop, Denkfysio, Nijmegen, 2018	16 hours
Masterclasses duizeligheid, kenniscentrum duizeligheid, Gelre Ziekenhuis Apeldoorn, 2017-2018	21 hours
Explain pain supercharged, Lorimer Moseley, Denkfysio, Amsterdam, 2017	16 hours
Extended scope practitioner, SOMT university, Amersfoort, 2015	114 hours
Ultrasound of the shoulder, Nationaal Trainingscentrum Echografie, Amersfoort, 2013	7 hours
	174 hours

Conferences

Oral presentation at 'dag van de fysiotherapeut' (KNGF), Utrecht, 2016	6 hours
Oral presentations at a conference ICMSU 2015, International Congress of Musculoskeletal Ultrasound, AMC, Amsterdam, 2015	6 hours
Poster presentation at a conference KNGF congres, 2012	1 hour
Organizing NVMT congress, Hilversum, the Netherlands, 2018	8 hours
Organizing NVMT congress, Papendal, the Netherlands, 2017	8 hours
Organizing NVMT congress, Ede, the Netherlands, 2016	8 hours
Organizing NVMT congress, Oegstgeest, the Netherlands 2015	8 hours
Organizing NVMT congress, Rotterdam, the Netherlands 2013	8 hours
Attending NHG wetenschapsdag 2013, Leiden, the Netherlands	8 hours
Attending Kennisnetwerk KNGF, Amersfoort, the Netherlands 2015	8 hours
Attending Oostendorp lezing, ervaringen uit de orofaciale MT, Amersfoort, the Netherlands 2015	3 hours
Attending IFOMPT 2012, Quebec, Canada	40 hours
Attending Dokters over schouders, 2012, Leiden, the Netherlands	4 hours
Attending ISOQOL 2012, Amsterdam, the Netherlands	8 hours
Attending WCPT 2011, Amsterdam, the Netherlands	32 hours
Attending IFOMPT 2008, Rotterdam, the Netherlands	16 hours

	172 hours

Teaching

Supervising Master's thesis 2015	56 hours
Supervising Bachelors's theses 2011-2014	350 hours
Physical therapy education 2011-2014	1200 hours
Developing educational material /post courses 2014-2018	300 hours
	1906 hours

Workload: 103 ECTS + 2080 hours (74 ECTS) = 171 ECTS

LIST OF PUBLICATIONS

In international peer reviewed journals

Karel Y, Thoomes-De Graaf M, Scholten-Peeters G, Ferreira P, Rizopoulos D, Koes BW, Verhagen AP. Validity of the Flemish working alliance inventory in a Dutch physiotherapy setting in patients with shoulder pain. Physiother Theory Pract. 2017 Nov 9:1-9. doi: 10.1080/09593985.2017.1400141. [Epub 2017 Nov 9].

Thoomes-de Graaf M., Scholten-Peeters G.G.W., Karel Y.H., Verwoerd A.., Koes B.W. , Verhagen A.P. One question might be capable of replacing the Shoulder Pain and Disability Index (SPADI) when measuring disability: a prospective cohort study. Qual Life Res. 2018 Feb;27(2):401-410. doi: 10.1007/s11136-017-1698-y. [Epub 2017 Sep 7].

Thoomes EJ, van Geest S, van der Windt DA, Falla D, Verhagen AP, Koes BW, Thoomes-de Graaf M, Kuijper B, Scholten-Peeters WGM, Vleggeert-Lankamp CL. Value of physical tests in diagnosing cervical radiculopathy: a systematic review. Spine J. 2018 Jan;18(1):179-189. doi: 10.1016/j.spinee.2017.08.241. [Epub 2017 Aug 31].

Thoomes-de Graaf M., Thoomes E., Carlesso L., Kerry R., Rushton A. Adverse effects as a consequence of being the subject of orthopaedic manual therapy training, a worldwide retrospective survey. Musculoskelet Sci Pract. 2017 Jun;29:20-27. doi: 10.1016/j.msksp.2017.02.009. [Epub 2017 March 6].

Thoomes-de Graaf M., Scholten-Peeters G.G., Duijn E., Karel Y.H., de Vet H.C.W., Koes B.W. , Verhagen A.P. The responsiveness and interpretability of the Shoulder Pain and Disability Index. J Orthop Sports Phys Ther. 2017 Apr;47(4):278-286. doi: 10.2519/jospt.2017.7079. [Epub 2017 Feb 3].

Karel YHJM, Verhagen AP, Thoomes-de Graaf M, Duijn E, van den Borne MPJ, Beumer A, Ottenheijm RPG, Dinant GJ, Koes BW, Scholten-Peeters GGM. Development of a Prognostic Model for Patients With Shoulder Complaints in Physical Therapist Practice. Phys Ther. 2017 Jan 1;97(1):72-80. doi: 10.2522/ptj.20150649.

Karel YHJM, Scholten-Peeters GGM, Thoomes-de Graaf M, Duijn E, van Broekhoven JB, Koes BW, Verhagen AP. Physiotherapy for patients with shoulder pain in primary care: a descriptive study of diagnostic- and therapeutic management. Physiotherapy. 2017 Dec;103(4):369-378. doi: 10.1016/j.physio.2016.11.003. [Epub 2016 Nov 28].

Thoomes-de Graaf M., Scholten-Peeters GGM., Schellingerhout JM, Bourne AM, Buchbinder R, Koehorst M, Terwee CB., Verhagen AP. Evaluation of measurement properties of self-administered PROMs aimed at patients with non-specific shoulder pain and "activity limitations": a systematic review. Qual Life Res. 2016 Sep;25(9):2141-60. doi: 10.1007/s11136-016-1277-7. [Epub 2016 Apr 2].

Thoomes-de Graaf M, Thoomes E.J. Ist zervikogener Schwindel eine eigenständige muskuloskeletale Entität? Manuelletherapie. July 2016 20(03):109-115. doi: 10.1055/s-0042-108661

Thoomes-de Graaf M, Thoomes E. A novel way of functional retraining of cervical motor control in a water polo player with combined cervicogenic and tension type headaches. J Man Manip Ther. 2016 Feb;24(1):26-33. doi:10.1179/2042618614Y.0000000067.

Thoomes-de Graaf M, Scholten-Peeters GG, Duijn E, Karel Y, Koes BW, Verhagen AP. The Dutch Shoulder Pain and Disability Index (SPADI): a reliability and validation study. Qual Life Res. 2015 Jun;24(6):1515-9. doi: 10.1007/s11136-014-0879-1. [Epub 2014 Dec 4].

Marloes Thoomes-de Graaf, Wendy GM Scholten-Peeters, Edwin Duijn, Yasmaine HJM Karel, Maaike PJ van den Borne, Annechien Beumer, Ramon PG Ottenheijm, Geert Jan Dinant, Eric Tetteroo, Cees Lucas, Bart W Koes, Arianne P Verhagen. Inter-professional agreement of ultrasound-based diagnoses in patients with shoulder pain between physical therapists and radiologists in the Netherlands. Man Ther. 2014 Oct;19(5):478-83. doi: 10.1016/j.math.2014.04.018. [Epub 2014 May 14].

Karel YH, Scholten-Peeters WG, Thoomes-de Graaf M, Duijn E, Ottenheijm RP, van den Borne MP, Koes BW, Verhagen AP; ShoCoDiP (Shoulder Complaints and using Diagnostic ultrasound in Physiotherapy practice) study group, Dinant GJ, Tetteroo E, Beumer A, van Broekhoven JB, Heijmans M. Current management and prognostic factors in physiotherapy practice for patients with shoulder pain: design of a prospective cohort study. BMC Musculoskelet Disord. 2013 Feb 11;14:62. doi:10.1186/1471-2474-14-62.

Thoomes-de Graaf M, Schmitt MS. The effect of training the deep cervical flexors on neck pain, neck mobility, and dizziness in a patient with chronic nonspecific neck pain after prolonged bed rest: a case report. J Orthop Sports Phys Ther. 2012 Oct;42(10):853-60. doi: 10.2519/jospt.2012.4056. [Epub 2012 Jul 26].

In national journals

Thoomes E.J., Thoomes-de Graaf M. Hoofdpijn vanuit een fysiotherapeutische perspectief bekeken. Physios. December 2015.

Thoomes-de Graaf M, Scholten-Peeters GG, Duijn E, Karel Y, Koes BW, Verhagen AP. The Dutch Shoulder Pain and Disability Index (SPADI): evaluatie validiteit en betrouwbaarheid Nederlandse SPADI. Fysiopraxis. Juli 2015.

Thoomes-de Graaf M, Thoomes E.J. Neurodynamica is essentieel in moderne musculoskeletale fysiotherapie. Physios. Maart 2015.

Books

Manuelletherapie Expertenwissen; Die besten Schwerpunkt-Artikel 2012 – 2016.
488 S. , 100 Abb. , gebunden (FH). Thieme, 2017. ISBN: 9783132419094

De pijnpuzzel. Marloes Thoomes-de Graaf, Michiel Trouw. Boekengilde, 2017 ISBN: 78-94-6323-2