Interpreting patient-reported outcomes in rheumatology

Thesis, University of Twente, 2008
ISBN 978-90-365-2659-3

Cover design by Arnold Veldhoen
Printed by Gildeprint Drukkerijen BV, Enschede, the Netherlands

The studies presented in this thesis were performed at the department of Psychology & Communication of Health & Risk (PCHR) of the University of Twente and the department of rheumatology of the Medisch Spectrum Twente hospital, both in Enschede, the Netherlands. The rheumatology research program of PCHR is financially supported by the Dutch Arthritis Association (Reumafonds).

INTERPRETING PATIENT-REPORTED OUTCOMES IN RHEUMATOLOGY

PROEFSCHRIFT

ter verkrijging van
de graad van doctor aan de Universiteit Twente,
op gezag van de rector magnificus,
prof. dr. W.H.M. Zijm,
volgens besluit van het College voor Promoties
in het openbaar te verdedigen
op vrijdag 6 juni 2008 om 16.45 uur

door

Peter Meindert ten Klooster
geboren op 3 juni 1976
te Hasselt

Promotor: prof. dr. M.A.F.J. van de Laar
Assistent-promotor: dr. E. Taal

Contents

Chapter 1	General introduction	7
Chapter 2	Changes in priorities for improvement in patients with rheumatoid arthritis during 1 year of anti-tumour necrosis factor treatment	21
Chapter 3	Patient-perceived satisfactory improvement (PPSI): interpreting meaningful change in pain from the patient's perspective	37
Chapter 4	Can we assess baseline pain and global health retrospectively?	51
Chapter 5	The validity and reliability of the graphic rating scale and verbal rating scale for measuring pain across cultures: a study in Egyptian and Dutch women with rheumatoid arthritis	63
Chapter 6	A cross-cultural study of pain intensity in Egyptian and Dutch women with rheumatoid arthritis	71
Chapter 7	Confirmatory factor analysis of the Arthritis Impact Measurement Scales 2 Short Form in patients with rheumatoid arthritis	85
Chapter 8	A Rasch analysis of the Dutch Health Assessment Questionnaire (HAQ) Disability Index and HAQ-II in patients with rheumatoid arthritis	103
Chapter 9	Summary and discussion	121
	Samenvatting (Dutch summary)	127
	Dankwoord (Acknowledgements)	131
	Curriculum vitae	133
	Publications	135

1 General introduction

Rheumatology and rheumatic diseases

Rheumatology deals with the study and treatment of disorders affecting the musculoskeletal system. The musculoskeletal or rheumatic diseases comprise more than 100 different disorders with a wide variety of clinical presentations.[1] Common rheumatic diseases are rheumatoid arthritis and osteoarthritis, but rare systemic diseases such as scleroderma and general problems like chronic low back pain, bursitis and tendinitis, and fibromyalgia are also generally labeled as rheumatic conditions. Although rheumatic diseases can affect all people of all ages, their prevalence is especially high in females and elderly persons. The cause of most rheumatic diseases is still unknown and many are chronic, disabling, and progressive. Since the majority of the rheumatic diseases cannot be cured, treatment is usually aimed at controlling the symptoms and progression of the disease. Reported prevalence rates of musculoskeletal or rheumatic diseases vary widely, depending on the specific definitions used and the populations studied.[2,3] In general, however, it is estimated that approximately 15 to 20% of the general western population suffers from some form of rheumatic disease.[4–7] A recent Dutch general public survey confirmed that 1 in 5 persons aged 20 years and older reported having rheumatic complaints.[8] One in 8 persons was seeing a physician for their rheumatic complaints.

The burden of rheumatic diseases

Rheumatic diseases have a major impact on both the individuals with the disease and the society in terms of economic, social, and psychological burden. Most rheumatic diseases are associated with high levels of pain and reduced physical function. Compared with other major disease groups, rheumatic diseases are the most common cause of chronic health problems and pain, the leading cause of long-term disability, and accountable for a considerable part of the total health care costs in western countries.[1,6,9–13] Moreover, persons with rheumatic diseases have significantly lower health-related quality of life compared with persons without musculoskeletal problems[14–18] or compared with persons with other common chronic conditions,[19–23] particularly in terms of bodily pain and physical functioning. As the average age of the western population is increasing, the prevalence and impact of rheumatic diseases are also expected to increase substantially in the near future.[1,24–28]

Outcome measurement in rheumatology

Since there is no single "gold standard" to evaluate disease severity or treatment effectiveness in most rheumatic diseases, multiple measures are used in the assessment of rheumatic patients.[29] Traditionally, patient assessment consisted of objective clinical

measures such as sedimentation rate and radiographic damage or physician-based measures such as swollen and tender joint counts.[30,31] In the past 25 years, however, clinicians and health professionals have increasingly recognized the importance of assessing the patient's perspective on the impact of the disease and the effectiveness of treatment. This has resulted in the development of a large number of concepts and instruments to measure patient-reported outcomes (PROs). Within rheumatology, especially the domains of pain, health status, physical function (or disability), and disease activity have received much attention. Specific PROs such as the visual analog scale for pain[32,33], the Arthritis Impact Measurement Scales,[34,35] the Medical Outcomes Study 36-Item Short Form Health Survey,[36] and the Health Assessment Questionnaire[31] have been extensively tested in patients with diverse rheumatic diseases and have demonstrated excellent reliability and validity. Moreover, several studies in rheumatoid arthritis have shown that PROs are at least as sensitive to treatment effects as traditional clinical or physician-reported measures and may even be less susceptible to placebo effects.[37–43] Consequently, patient-reported pain, physical function, and global assessment of disease status are now widely accepted as major endpoints in rheumatology and are generally included in internationally agreed core sets of measures to be used in clinical trials.[44–50]

Objective of this thesis

Although PROs have proven their value in rheumatology, several issues have been raised about their current use and interpretation. Many of these issues stem from two contemporary paradigm shifts in health research in general and in the rheumatic diseases in particular. The first concerns the increasing interest in truly capturing and interpreting the individual patient's perspective on health and changes in health using PROs.[51,52] The second is the recent emergence of modern psychometric methods in health measurement which have the potential to strongly improve the precision and efficiency of PROs.[53,54] The studies in this thesis address some of these issues and were performed to further improve our understanding of PROs in rheumatology.

Data collection

The data presented in the studies were derived from four independent patient samples. The results in chapters 2 and 7 are based on data obtained from the ongoing Dutch Rheumatoid Arthritis Anti-TNF Monitoring (DREAM) study, a register for the prospective monitoring and evaluation of patients with active RA starting anti-tumor necrosis factor (TNF) treatment in 12 hospitals in the Netherlands. Chapters 3 and 4 are the result of a prospective study of 200 rheumatic patients with localized pain who

were treated with a local corticosteroid injection. The patients in this study were consecutively recruited from the outpatient rheumatology clinic at the Medisch Spectrum Twente (MST) hospital in Enschede, the Netherlands. The data for chapters 5 and 6 were collected in a cross-cultural study of 42 Egyptian and 30 Dutch young female patients with stable RA. The Egyptian patients were consecutively enrolled at the outpatient rheumatology clinic of the University Hospital of Cairo, Egypt. The age- and disease duration-matched Dutch patients were selected from the patient registry of the rheumatology clinic of the MST hospital. Finally, the data presented in chapter 8 were obtained during several waves of data collection at the outpatient rheumatology clinic of the MST hospital.

Outline

Patients' priorities for improvement

One concern about PROs in rheumatology is the current lack of understanding of the importance of different aspects of health to different patients. Whereas health outcomes are increasingly assessed by patients' self-reports, it is still usually the physician or investigator who judges the relative importance of these outcomes.[55] However, not all aspects of health are equally important to different patients.[56,57] Moreover, there is increasing evidence that patients' and physicians' perceptions of important outcomes are not congruent.[58–60] The importance of further research on patients' priorities for improvement was emphasized by both patients and professionals at the recent Outcome Measures in Rheumatoid Arthritis Clinical Trials (OMERACT) meetings.[51,61] Previous studies in cross-sectional RA populations showed that patients generally consider pain and physical disabilities as the most important targets for treatement.[35,59,62–68] More recently, however, it was suggested that these priorities for improvement can vary over the course of time and treatment. During a disease flare, for example, pain reduction may become more important than during stable disease.[61,69] To date, no studies have adequately examined the longitudinal course of patient priorities during a treatment with proven efficacy. **Chapter 2** of this thesis reports the results of a study of the 1-year course of patients' priorities for improvement in a cohort of 226 patients with active RA starting treatment with anti-TNF agents.

Meaningful improvements from the patient's perspective

Another contemporary issue in rheumatology concerns the interpretation of changes in continuous PROs such as pain. In evaluating the effectiveness of treatments it is important to determine the clinical significance of change scores, since even very small changes can reach statistical significance but may be clinically trivial.[70–73] As a result, much effort has been directed to the development and use of single definitions of

response, that is, to define cut-off points for clinically important improvement.[74] Two well-known (clinician-) opinion based or data-driven criteria for improvement are the American College of Rheumatology and the European League Against Rheumatism response criteria, which classify patients as improved or not based on a core set of outcome measures.[75,76] In recent years, more focus has been directed at determining minimal clinically important differences (MCIDs) for individual outcome measures. A MCID is the smallest amount of change on an outcome measure that represents a clinically meaningful change. MCIDs can be based on different perspectives of important improvement, such as the clinician's, society's, or the patient's perspective.[77,78] The most popular patient-centered approach uses patients' retrospective global ratings of change as an external anchor to determine mean values for important changes on the measure of interest.[79] Variations on this method have been used to determine the MCID for several PROs in rheumatic disease populations.[80–86] However, this method is designed to consider important differences at the group level[78] and mixes patients' and clinicians' perspectives since the latter still decide which ratings of change are important.[87] Moreover, the derived MCIDs do not meet the growing need for definitions of "major" clinically important improvements.[88,89] At a previous OMERACT meeting, it was concluded that methods should be developed that focus on major changes on rheumatology outcomes from the individual patient's perspective.[78,88,90] In **chapter 3** a new concept for measuring meaningful change from the individual patient's perspective is described. This concept called "patient-perceived satisfactory improvement" was used to determine the optimal cut-off point for meaningful change on the visual analog scale (VAS) for pain in a 2-week prospective study of 200 rheumatology outpatients treated with a local corticosteroid injection.

Patients' retrospective assessments of baseline states

A more general concern with PROs is the validity of retrospective reports of symptoms or health. In clinical research on treatment efficacy, changes in health states are usually measured prospectively as the difference between pre-treatment and post-treatment outcomes, obtained from serial measurements. In clinical practice, on the other hand, changes are commonly evaluated retrospectively by asking patients to compare their current state with their pre-treatment state.[91,92] Some study designs, such as cross-sectional or retrospective studies of treatment effectiveness, also rely on patients' recall of pre-treatment symptoms or overall health status.[93–96] Although retrospective measurement is obviously easier and more economic than prospective measurement, prospective designs are considered superior since recall of pre-treatment states may be flawed by memory problems and biases.[92,93,96,97] To date, no studies have examined the accuracy of retrospective reports of pre-treatment health states in rheumatic patients. **Chapter 4** is also based on the results from the local injection study. This chapter

examines the agreement between patients' actual assessments of pain and global health on the VAS collected just before the injection and retrospective assessments collected two weeks after the injection.

Cross-cultural assessment of pain

The increasingly multicultural society and the growing number of multinational clinical trials in rheumatology raises the issue of the validity and comparability of PROs like pain across patients from different cultural or ethnic backgrounds. Although pain is a universal experience for patients with a chronic rheumatic disease, little is known about ethno-cultural variations in the perception and reporting of rheumatic pain intensity. Several recent studies in other chronic pain populations have shown that patients' pain reports may indeed be affected by cultural or ethnic factors.[98–104] Some studies have also found ethno-cultural differences in the experience of pain in patients with a rheumatic disease.[105–108] However, most of these studies have compared more or less acculturated ethnic groups within the same nation or may have confounded ethnicity with other variables such as socioeconomic status. Only few studies have thoroughly examined differences in pain perception in patients living in culturally different countries or regions. Moreover, no studies have directly compared the cross-cultural validity and reliability of commonly used pain measures in rheumatology. In chapters 5 and 6 the results of a cross-cultural study in Egyptian and Dutch young female RA patients are presented. In **chapter 5** the validity and reliability of a verbal and a graphic rating scale for measuring pain intensity are compared in the Egyptian and Dutch samples. Ethnocultural differences in the perception and reporting of pain intensity and determinants of pain intensity are examined in **chapter 6**.

Application of modern psychometric methods to improve PROs

Finally, most current PROs in rheumatology have been developed and evaluated using classic test theory. In recent years, however, more powerful analysis techniques such as structural equation modeling (SEM) and item response theory (IRT) have entered the field of health status assessment.[109–112] These "modern" techniques, which have been in use for several decades in other research areas such as educational and psychological research, allow for the sophisticated analyses of existing measures and the development of more precise patient-reported health measures.[54,113,114] SEM can, for instance, be used to assess the adequacy of a hypothesized measurement structure underlying a PRO. This procedure is also known as confirmatory factor analysis (CFA) and has a major advantage over traditional exploratory factor analysis in that it allows for the statistical testing of the "goodness-of-fit" of an a-priori defined measurement model to a given data set and comparison of competing measurement models.[115] IRT consists of a family of mathematical models that relate the probability of a person's response to a

specific item to this person's location on some underlying latent construct (or trait) being measured by the questionnaire.[113] With IRT methods, items can be ordered in terms of difficulty on a continuous scale, so that overlapping items can be identified and those items can be selected that best measure the full range of a construct. IRT additionally enables the rigorous testing of whether items perform the same or differently across different subgroups, countries, or cultures trough analysis of differential item functioning. Finally, IRT provides the basis for computerized adaptive testing, where different respondents receive different sets of questions from a large item bank based on their level of health on the specific dimension being evaluated. **Chapter 7** presents the results of a confirmatory factor analysis of the short form Arthritis Impact Measurement Scales 2 (AIMS2-SF), an arthritis-specific measure of health status, using SEM techniques. Using baseline data from the cohort of patients starting anti-TNF treatment, three hypothesized measurement models were tested and compared. Finally, in **chapter 8** the Health Assessment Questionnaire Disability Index (HAQ-DI) and the revised Health Assessment Questionnaire (HAQ-II) were evaluated using Rasch analysis, a form of IRT methodology, in a cross-sectional sample of 472 patients with confirmed RA.

References

1. Sangha O. Epidemiology of rheumatic diseases. Rheumatology (Oxford) 2000;39: 3–12.
2. Symmons DPM. Population studies of musculoskeletal morbidity. In: Silman AJ, Hochberg MC, eds. Epidemiology of the rheumatic diseases. Oxford: Oxford University Press; 2001:5–28.
3. Picavet HS, Hazes JM. Prevalence of self reported musculoskeletal diseases is high. Ann Rheum Dis 2003;62:644–50.
4. Lee P, Helewa A, Smythe HA, Bombardier C, Goldsmith CH. Epidemiology of musculoskeletal disorders (complaints) and related disability in Canada. J Rheumatol 1985;12:1169–73.
5. Lawrence RC, Helmick CG, Arnett FC, et al. Estimates of the prevalence of arthritis and selected musculoskeletal disorders in the United States. Arthritis Rheum 1998;41:778–99.
6. Picavet HS, van den Bos GA. The contribution of six chronic conditions to the total burden of mobility disability in the Dutch population. Am J Public Health 1997;87:1680–2.
7. Helmick CG, Felson DT, Lawrence RC, et al. Estimates of the prevalence of arthritis and other rheumatic conditions in the United States: Part I. Arthritis Rheum 2007;58:15–25.
8. Chorus AMJ, van Overbeek K, Hopman-Rock M. Reumatische klachten in Nederland: resultaten Nationale Peiling van het Bewegingsapparaat 2006 Leiden: TNO Kwaliteit van Leven; 2007.
9. Badley EM, Rasooly I, Webster GK. Relative importance of musculoskeletal disorders as a cause of chronic health problems, disability, and health care utilization: findings from the 1990 Ontario Health Survey. J Rheumatol 1994;21:505–14.

10. Yelin E, Callahan LF. The economic cost and social and psychological impact of musculoskeletal conditions. National Arthritis Data Work Groups. Arthritis Rheum 1995;38:1351–62.
11. Meerding WJ, Bonneux L, Polder JJ, Koopmanschap MA, van der Maas PJ. Demographic and epidemiological determinants of healthcare costs in Netherlands: cost of illness study. BMJ 1998;317:111–5.
12. Elliott AM, Smith BH, Penny KI, Smith WC, Chambers WA. The epidemiology of chronic pain in the community. Lancet 1999;354:1248–52.
13. Lee P. The economic impact of musculoskeletal disorders. Qual Life Res 1994;3 S85–91.
14. Salaffi F, Carotti M, Stancati A, Grassi W. Health-related quality of life in older adults with symptomatic hip and knee osteoarthritis: a comparison with matched healthy controls. Aging Clin Exp Res 2005;17:255–63.
15. Picavet HS, Hoeymans N. Health related quality of life in multiple musculoskeletal diseases: SF-36 and EQ-5D in the DMC3 study. Ann Rheum Dis 2004;63:723–9.
16. Roux CH, Guillemin F, Boini S, et al. Impact of musculoskeletal disorders on quality of life: an inception cohort study. Ann Rheum Dis 2005;64:606–11.
17. Hill CL, Parsons J, Taylor A, Leach G. Health related quality of life in a population sample with arthritis. J Rheumatol 1999;26:2029–35.
18. West E, Jonsson SW. Health-related quality of life in rheumatoid arthritis in Northern Sweden: a comparison between patients with early RA, patients with medium-term disease and controls, using SF-36. Clin Rheumatol 2005;24:117–22.
19. Mason JH, Weener JL, Gertman PM, Meenan RF. Health status in chronic disease: a comparative study of rheumatoid arthritis. J Rheumatol 1983;10:763–8.
20. Kempen GI, Ormel J, Brilman EI, Relyveld J. Adaptive responses among Dutch elderly: the impact of eight chronic medical conditions on health-related quality of life. Am J Public Health 1997;87:38–44.
21. Sprangers MA, de Regt EB, Andries F, et al. Which chronic conditions are associated with better or poorer quality of life? J Clin Epidemiol 2000;53:895–907.
22. Wee HL, Cheung YB, Li SC, Fong KY, Thumboo J. The impact of diabetes mellitus and other chronic medical conditions on health-related Quality of Life: is the whole greater than the sum of its parts? Health Qual Life Outcomes 2005;3:2.
23. Stavem K, Lossius MI, Kvien TK, Guldvog B. The health-related quality of life of patients with epilepsy compared with angina pectoris, rheumatoid arthritis, asthma and chronic obstructive pulmonary disease. Qual Life Res 2000;9:865–71.
24. Lubeck DP. The costs of musculoskeletal disease: health needs assessment and health economics. Best Pract Res Clin Rheumatol 2003;17:529–39.
25. Elders MJ. The increasing impact of arthritis on public health. J Rheumatol 2000;27 6–8.
26. Helmick CG, Lawrence RC, Pollard RA, Lloyd E, Heyse SP. Arthritis and other rheumatic conditions: who is affected now, who will be affected later? National Arthritis Data Workgroup. Arthritis Care Res 1995;8:203–11.
27. Reginster JY. The prevalence and burden of arthritis. Rheumatology (Oxford) 2002;41 3–6.
28. Brooks PM. Impact of osteoarthritis on individuals and society: how much disability? Social consequences and health economic implications. Curr Opin Rheumatol 2002;14:573–7.

29. Pincus T, Sokka T. Complexities in the quantitative assessment of patients with rheumatic diseases in clinical trials and clinical care. Clin Exp Rheumatol 2005;23:S1–9.
30. Fries JF, Bruce B, Cella D. The promise of PROMIS: using item response theory to improve assessment of patient-reported outcomes. Clin Exp Rheumatol 2005;23:S53–7.
31. Fries JF, Spitz P, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. Arthritis Rheum 1980;23:137–45.
32. Huskisson EC. Measurement of pain. Lancet 1974;2:1127–31.
33. Scott J, Huskisson EC. Graphic representation of pain. Pain 1976;2:175–84.
34. Meenan RF, Gertman PM, Mason JH. Measuring health status in arthritis: the Arthritis Impact Measurement Scales. Arthritis Rheum 1980;23:146–52.
35. Meenan RF, Mason JH, Anderson JJ, Guccione AA, Kazis LE. AIMS2. The content and properties of a revised and expanded Arthritis Impact Measurement Scales health status questionnaire. Arthritis Rheum 1992;35:1–10.
36. Ware JE, Jr., Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care 1992;30:473–83.
37. Cohen SB, Strand V, Aguilar D, Ofman JJ. Patient- versus physician-reported outcomes in rheumatoid arthritis patients treated with recombinant interleukin-1 receptor antagonist (anakinra) therapy. Rheumatology (Oxford) 2004;43:704–11.
38. Gotzsche PC. Sensitivity of effect variables in rheumatoid arthritis: a meta-analysis of 130 placebo controlled NSAID trials. J Clin Epidemiol 1990;43:1313–8.
39. Strand V, Cohen S, Crawford B, Smolen JS, Scott DL. Patient-reported outcomes better discriminate active treatment from placebo in randomized controlled trials in rheumatoid arthritis. Rheumatology (Oxford) 2004;43:640–7.
40. Buchbinder R, Bombardier C, Yeung M, Tugwell P. Which outcome measures should be used in rheumatoid arthritis clinical trials? Clinical and quality-of-life measures' responsiveness to treatment in a randomized controlled trial. Arthritis Rheum 1995;38:1568–80.
41. Verhoeven AC, Boers M, van Der Linden S. Responsiveness of the core set, response criteria, and utilities in early rheumatoid arthritis. Ann Rheum Dis 2000;59:966–74.
42. Anderson JJ, Chernoff MC. Sensitivity to change of rheumatoid arthritis clinical trial outcome measures. J Rheumatol 1993;20:535–7.
43. Pincus T, Strand V, Koch G, et al. An index of the three core data set patient questionnaire measures distinguishes efficacy of active treatment from that of placebo as effectively as the American College of Rheumatology 20% response criteria (ACR20) or the Disease Activity Score (DAS) in a rheumatoid arthritis clinical trial. Arthritis Rheum 2003;48:625–30.
44. van der Heijde D, Bellamy N, Calin A, Dougados M, Khan MA, van der Linden S. Preliminary core sets for endpoints in ankylosing spondylitis. Assessments in Ankylosing Spondylitis Working Group. J Rheumatol 1997;24:2225–9.
45. Felson DT, Anderson JJ, Boers M, et al. The American College of Rheumatology preliminary core set of disease activity measures for rheumatoid arthritis clinical trials. Arthritis Rheum 1993;36:729–40.
46. Boers M, Tugwell P, Felson DT, et al. World Health Organization and International League of Associations for Rheumatology core endpoints for symptom modifying antirheumatic drugs in rheumatoid arthritis clinical trials. J Rheumatol 1994;21:86–9.

47. Bellamy N, Kirwan J, Boers M, et al. Recommendations for a core set of outcome measures for future phase III clinical trials in knee, hip, and hand osteoarthritis. Consensus development at OMERACT III. J Rheumatol 1997;24:799–802.
48. Leeb BF, Bird HA, Nesher G, et al. EULAR response criteria for polymyalgia rheumatica: results of an initiative of the European Collaborating Polymyalgia Rheumatica Group (subcommittee of ESCISIT). Ann Rheum Dis 2003;62:1189–94.
49. Taylor WJ. Preliminary identification of core domains for outcome studies in psoriatic arthritis using Delphi methods. Ann Rheum Dis 2005;64 ii110–2.
50. Merkel PA, Herlyn K, Martin RW, et al. Measuring disease activity and functional status in patients with scleroderma and Raynaud's phenomenon. Arthritis Rheum 2002;46:2410–20.
51. Kirwan J, Heiberg T, Hewlett S, et al. Outcomes from the Patient Perspective Workshop at OMERACT 6. J Rheumatol 2003;30:868–72.
52. Heller JE, Shadick NA. Outcomes in rheumatoid arthritis: incorporating the patient perspective. Curr Opin Rheumatol 2007;19:101–5.
53. Fries JF, Bruce B, Bjorner J, Rose M. More relevant, precise, and efficient items for assessment of physical function and disability: moving beyond the classic instruments. Ann Rheum Dis 2006;65 Suppl 3:iii16–iii21.
54. Chakravarty EF, Bjorner JB, Fries JF. Improving patient reported outcomes using item response theory and computerized adaptive testing. J Rheumatol 2007;34:1426–31.
55. Hewlett SA. Patients and clinicians have different perspectives on outcomes in arthritis. J Rheumatol 2003;30:877–9.
56. Gill TM, Feinstein AR. A critical appraisal of the quality of quality-of-life measurements. JAMA 1994;272:619–26.
57. O'Boyle CA, McGee H, Hickey A, O'Malley K, Joyce CR. Individual quality of life in patients undergoing hip replacement. Lancet 1992;339:1088–91.
58. Hewlett S, Smith AP, Kirwan JR. Values for function in rheumatoid arthritis: patients, professionals, and public. Ann Rheum Dis 2001;60:928–33.
59. Kwoh CK, Ibrahim SA. Rheumatology patient and physician concordance with respect to important health and symptom status outcomes. Arthritis Rheum 2001;45:372–7.
60. Rothwell PM, McDowell Z, Wong CK, Dorman PJ. Doctors and patients don't agree: cross sectional study of patients' and doctors' perceptions and assessments of disability in multiple sclerosis. BMJ 1997;314:1580–3.
61. Kirwan JR, Hewlett SE, Heiberg T, et al. Incorporating the patient perspective into outcome assessment in rheumatoid arthritis — progress at OMERACT 7. J Rheumatol 2005;32:2250–6.
62. Archenholtz B, Bjelle A. Reliability, validity, and sensitivity of a Swedish version of the revised and expanded Arthritis Impact Measurement Scales (AIMS2). J Rheumatol 1997;24:1370–7.
63. Gibson T, Clark B. Use of simple analgesics in rheumatoid arthritis. Ann Rheum Dis 1985;44:27–9.
64. Heiberg T, Kvien TK. Preferences for improved health examined in 1,024 patients with rheumatoid arthritis: pain has highest priority. Arthritis Rheum 2002;47:391–7.
65. McKenna F, Wright V. Pain and rheumatoid arthritis. Ann Rheum Dis 1985;44:805.

66. Minnock P, Fitzgerald O, Bresnihan B. Quality of life, social support, and knowledge of disease in women with rheumatoid arthritis. Arthritis Rheum 2003;49:221–7.
67. Riemsma RP, Taal E, Rasker JJ, Houtman PM, van Paassen HC, Wiegman O. Evaluation of a Dutch version of the AIMS2 for patients with rheumatoid arthritis. Br J Rheumatol 1996;35:755–60.
68. Taal E, Rasker JJ, Evers AW, Kraaimaat FW, Lanting PJH, Jacobs JW. Which priorities have rheumatoid arthritis (RA) patients for their health status improvement? [Abstract]. Arthritis Rheum 1997;40 S231.
69. Carr A, Hewlett S, Hughes R, et al. Rheumatology outcomes: the patient's perspective. J Rheumatol 2003;30:880–3.
70. Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. J Clin Epidemiol 2003;56:395–407.
71. Lydick E, Epstein RS. Interpretation of quality of life changes. Qual Life Res 1993;2:221–6.
72. Redelmeier DA, Lorig K. Assessing the clinical importance of symptomatic improvements. An illustration in rheumatology. Arch Intern Med 1993;153:1337–42.
73. Deyo RA, Patrick DL. The significance of treatment effects: the clinical perspective. Med Care 1995;33:AS286–91.
74. Kvien TK, Heiberg T, Hagen KB. Minimal clinically important improvement/difference (MCII/MCID) and patient acceptable symptom state (PASS): what do these concepts mean? Ann Rheum Dis 2007;66:iii40–iii1.
75. Felson DT, Anderson JJ, Boers M, et al. American College of Rheumatology. Preliminary definition of improvement in rheumatoid arthritis. Arthritis Rheum 1995;38:727–35.
76. van Gestel AM, Prevoo ML, van 't Hof MA, van Rijswijk MH, van de Putte LB, van Riel PL. Development and validation of the European League Against Rheumatism response criteria for rheumatoid arthritis. Comparison with the preliminary American College of Rheumatology and the World Health Organization/International League Against Rheumatism Criteria. Arthritis Rheum 1996;39:34–40.
77. Beaton DE, Boers M, Wells GA. Many faces of the minimal clinically important difference (MCID): a literature review and directions for future research. Curr Opin Rheumatol 2002;14:109–14.
78. Wells G, Beaton D, Shea B, et al. Minimal clinically important differences: review of methods. J Rheumatol 2001;28:406–12.
79. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. Control Clin Trials 1989;10:407–15.
80. Angst F, Aeschlimann A, Michel BA, Stucki G. Minimal clinically important rehabilitation effects in patients with osteoarthritis of the lower extremities. J Rheumatol 2002;29:131–8.
81. de Boer YA, Hazes JM, Winia PC, Brand R, Rozing PM. Comparative responsiveness of four elbow scoring instruments in patients with rheumatoid arthritis. J Rheumatol 2001;28:2616–23.
82. Dhanani S, Quenneville J, Perron M, Abdolell M, Feldman BM. Minimal difference in pain associated with change in quality of life in children with rheumatic disease. Arthritis Rheum 2002;47:501–5.

83. Pavy S, Brophy S, Calin A. Establishment of the minimum clinically important difference for the bath ankylosing spondylitis indices: a prospective study. J Rheumatol 2005;32:80–5.
84. Salaffi F, Stancati A, Silvestri CA, Ciapetti A, Grassi W. Minimal clinically important changes in chronic musculoskeletal pain intensity measured on a numerical rating scale. Eur J Pain 2004;8:283–91.
85. Stratford PW, Binkley JM, Riddle DL, Guyatt GH. Sensitivity to change of the Roland-Morris Back Pain Questionnaire: part 1. Phys Ther 1998;78:1186–96.
86. Tubach F, Ravaud P, Baron G, et al. Evaluation of clinically relevant changes in patient reported outcomes in knee and hip osteoarthritis: the minimal clinically important improvement. Ann Rheum Dis 2005;64:29–33.
87. Beaton DE, Bombardier C, Katz JN, Wright JG. A taxonomy for responsiveness. J Clin Epidemiol 2001;54:1204–17.
88. Tugwell P, Boers M, Brooks PM, Simon L, Strand CV. OMERACT 5: International consensus conference on outcome measures in rheumatology. J Rheumatol 2001;28:395–7.
89. Wolfe F, Michaud K, Strand V. Expanding the definition of clinical differences: from minimally clinically important differences to really important differences. Analyses in 8931 patients with rheumatoid arthritis. J Rheumatol 2005;32:583–9.
90. Wells G, Anderson J, Beaton D, et al. Minimal clinically important difference module: summary, recommendations, and research agenda. J Rheumatol 2001;28:452–4.
91. Fischer D, Stewart AL, Bloch DA, Lorig K, Laurent D, Holman H. Capturing the patient's view of change as a clinical outcome measure. JAMA 1999;282:1157–62.
92. Middel B, Goudriaan H, de Greef M, et al. Recall bias did not affect perceived magnitude of change in health-related functional status. J Clin Epidemiol 2006;59:503–11.
93. Mancuso CA, Charlson ME. Does recollection error threaten the validity of cross-sectional studies of effectiveness? Med Care 1995;33:AS77–88.
94. Lingard EA, Wright EA, Sledge CB. Pitfalls of using patient recall to derive preoperative status in outcome studies of total knee arthroplasty. J Bone Joint Surg Am 2001;83:1149–56.
95. Pellise F, Vidal X, Hernandez A, Cedraschi C, Bago J, Villanueva C. Reliability of retrospective clinical data to evaluate the effectiveness of lumbar fusion in chronic low back pain. Spine 2005;30:365–8.
96. Herrmann D. Reporting current, past, and changed health status. What we know about distortion. Med Care 1995;33:AS89–94.
97. Aseltine RH, Jr., Carlson KJ, Fowler FJ, Jr., Barry MJ. Comparing prospective and retrospective measures of treatment outcomes. Med Care 1995;33:AS67–76.
98. Bates MS, Edwards WT, Anderson KO. Ethnocultural influences on variation in chronic pain perception. Pain 1993;52:101–12.
99. Cano A, Mayo A, Ventimiglia M. Coping, pain severity, interference, and disability: the potential mediating and moderating roles of race and education. J Pain 2006;7:459–68.
100. Cohen MZ, Musgrave CF, Munsell MF, Mendoza TR, Gips M. The cancer pain experience of Israeli and American patients 65 years and older. J Pain Symptom Manage 2005;30:254–63.
101. Green CR, Baker TA, Sato Y, Washington TL, Smith EM. Race and chronic pain: a comparative study of young black and white Americans presenting for management. J Pain 2003;4:176–83.

102. McCracken LM, Matthews AK, Tang TS, Cuba SL. A comparison of blacks and whites seeking treatment for chronic pain. Clin J Pain 2001;17:249–55.
103. Portenoy RK, Ugarte C, Fuller I, Haas G. Population-based survey of pain in the United States: differences among white, African American, and Hispanic subjects. J Pain 2004;5:317–28.
104. Riley JL, 3rd, Wade JB, Myers CD, Sheffield D, Papas RK, Price DD. Racial/ethnic differences in the experience of chronic pain. Pain 2002;100:291–8.
105. Neumann L, Buskila D. Ethnocultural and educational differences in Israeli women correlate with pain perception in fibromyalgia. J Rheumatol 1998;25:1369–73.
106. Jordan MS, Lumley MA, Leisen JC. The relationships of cognitive coping and pain control beliefs to pain and adjustment among African-American and Caucasian women with rheumatoid arthritis. Arthritis Care Res 1998;11:80–8.
107. Thumboo J, Chew LH, Lewin-Koh SC. Socioeconomic and psychosocial factors influence pain or physical function in Asian patients with knee or hip osteoarthritis. Ann Rheum Dis 2002;61:1017–20.
108. Creamer P, Lethbridge-Cejku M, Hochberg MC. Determinants of pain severity in knee osteoarthritis: effect of demographic and psychosocial variables using 3 pain measures. J Rheumatol 1999;26:1785–92.
109. Revicki DA, Cella DF. Health status assessment for the twenty-first century: item response theory, item banking and computer adaptive testing. Qual Life Res 1997;6:595.
110. McHorney CA. Generic health measurement: past accomplishments and a measurement paradigm for the 21st century. Ann Intern Med 1997;127:743–50.
111. Cella D, Chang CH. A discussion of item response theory and its applications in health status assessment. Med Care 2000;38:II66–72.
112. Ware JE, Jr., Bjorner JB, Kosinski M. Practical implications of item response theory and computerized adaptive testing: a brief summary of ongoing studies of widely used headache impact scales. Med Care 2000;38:II73–82.
113. Reeve BB, Fayers P. Applying item response theory modelling for evaluating questionnaire item and scale properties. In: Fayers PM, Hays RD, eds. Assessing quality of life in clinical trials: Methods and practice. Oxford: Oxford University Press; 2005:55–73.
114. Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. Med Care 2000;38:II28–42.
115. Bryant FB, Yarnold PR, Michelson EA. Statistical methodology: VIII. Using confirmatory factor analysis (CFA) in emergency medicine research. Acad Emerg Med 1999;6:54–66.

2 Changes in priorities for improvement in patients with rheumatoid arthritis during 1 year of anti-tumour necrosis factor treatment

P.M. ten Klooster
M.M. Veehof
E. Taal
P.L.C.M. van Riel
M.A.F.J. van de Laar

Annals of the Rheumatic Diseases 2007; 66: 1485–1490.

Abstract

Objectives. To examine priorities for health status improvement in patients with active rheumatoid arthritis (RA) during anti-tumour necrosis factor (TNF) treatment.

Methods. Data were used from 173 patients with RA starting treatment with TNF-blocking agents. Outcome measures included assessment of health status with the Arthritis Impact Measurement Scales 2 (AIMS2) at baseline and after 3 and 12 months. The AIMS2 contains a priority list from which patients are asked to select from 12 areas of health the 3 in which they would most like to see improvement.

Results. After 1 year of treatment, 10 out of 12 areas of health on the AIMS2 were significantly improved. The most commonly selected priorities for improvement at baseline were pain (88%), hand and finger function (57%), walking and bending (42%), mobility (33%), and work (29%). At group level, this priority ranking remained largely unchanged during treatment. After adjustment for multiple comparisons, only pain was selected significantly less often at 3 and 12 months (71% at both assessments). Within individual patients, however, priorities often changed. Changes in the priority of pain were related to the achieved level of patient-perceived pain and disease activity.

Conclusions. This study shows that, at the group level, patients' priorities for improvement are fairly stable during 12 months of anti-TNF therapy, despite major improvements in health status. Although pain reduction becomes somewhat less important, it remains the most commonly selected priority. In contrast, individual patient priorities are not stable over the course of treatment and appear to be associated with differences in disease state.

Introduction

Rheumatoid arthritis (RA) is a common chronic inflammatory disease that greatly affects patients' physical, psychological and social wellbeing.[1–4] Over the years, various questionnaires have become available for measuring health status in patients with RA and multidimensional assessment of health status has now become common in clinical trials of RA. Clinicians or investigators usually determine the relative importance of these different dimensions of health. However, not all aspects of health are equally important to different patients[5,6] and patients' and doctors' perceptions of important health status outcomes may differ considerably.[7–10]

To accurately measure health status from the patient's perspective, it is essential to identify the aspects of health that patients would most like to see improved. Previous studies that have explored patient perceptions of the relative importance of improving different health aspects generally indicate that patients with RA consider pain and physical disabilities to be the most important targets for treatment.[7,11–18] However, it has been suggested that the relative importance of specific outcomes is not stable over the course of time or treatment.[6,19] During a flare, for example, pain reduction may be the most important priority, whereas other areas of health are more important during stable disease.[20,21]

To date, no studies have examined longitudinal changes in patients' priorities for improvement. One recent study examined 7-year changes in priorities for improvement in two cross-sectional RA cohorts.[22] Although all aspects of health had improved, patients' priorities for improvement remained mostly unchanged. This suggests that priorities for improvement are quite stable over time and not clearly associated with achieved improvements in health status. However, the authors performed a cross-sectional comparison on two partially overlapping populations, thus complicating the interpretation of the results.[23] Moreover, the observed improvements in health status were only minimal.[23]

The goals of the present study were to investigate the priorities for health status improvement in a cohort of patients with RA with high disease activity beginning tumour necrosis factor (TNF)-blocking treatment, and to examine changes in these priorities after 3 and 12 months.

Methods

Patients and study design

The data for this study were collected as part of the ongoing Dutch Rheumatoid Arthritis Anti-TNF Monitoring (DREAM) study, a register that started in April 2003 to prospectively monitor and evaluate the use of anti-TNF in patients with RA in 12 hospitals in the Netherlands. In this study, all patients with RA starting on anti-TNF

are evaluated every 3 months. Inclusion criteria for the DREAM study are a diagnosis of RA,[24] active disease (Disease Activity Score 28 (DAS28) >3.2),[25] previous treatment with at least two anti-rheumatic drugs including methotrexate at an optimum dose, or intolerance to methotrexate and no previous treatment with anti-TNF agents.

For this study, we used data from a subset of centres that included the following measures at baseline and at the 3-month and 12-month follow-up assessments: Health Assessment Questionnaire Disability Index (HAQ-DI),[26,27] Rheumatoid Arthritis Disease Activity Index (RADAI),[28,29] the 100 mm Visual Analogue Scale for General Health (VAS-GH) and the Arthritis Impact Measurement Scales 2 (AIMS2).[15,17]

Measures

The HAQ-DI contains 20 items measuring physical disabilities over the past week in eight categories of daily living: dressing, arising, eating, walking, hygiene, reach, grip, and common daily activities.[26,27] The HAQ was scored using the standard Disability Index, which takes into account the use of aids and devices. The HAQ-DI yields a score from 0 to 3, with higher scores indicating more disability.

The RADAI is a 5-item questionnaire for disease activity that asks patients to rate their global disease activity in the past 6 months, current disease activity in terms of swollen and tender joints, current arthritis pain, current duration of morning stiffness and number of tender joints in a joint list.[28,29] The first three items were rated on 11-point numerical rating scales. The combined RADAI score ranges from 0 to 10, with higher scores indicating more disease activity.

The VAS-GH is a 100 mm horizontal line ranging from 0 (best) to 100 (worst). Patients were asked to rate their current general health.

The AIMS2 is a disease-specific questionnaire designed to measure various components of health status in patients with arthritis.[15,17] The core part of the questionnaire contains 57 items that are categorised in 12 scales representing different areas of health. The scales can be combined into five summary component scores: physical (mobility level, walking and bending, hand and finger function, arm function, self-care, household tasks), affect (level of tension, mood), symptom (arthritis pain), social interaction (social activity, support from family and friends) and role (work). The scores on each scale or component range from 0 to 10, with higher scores representing poorer health status. Additionally, the AIMS2 contains sections on patient satisfaction with the 12 areas of health, effect of arthritis on each area of health, priorities for improvement, general perceptions of current and future health, and medical and demographic characteristics. The priority list (item 60) asks patients to select from 12 areas of health the 3 in which they would most like to see improvement. General satisfaction with current health (item 62) is assessed with a single-item Likert scale ranging from very satisfied (1) to very dissatisfied (5). For the analyses, responses to this item were dichotomised

into "satisfied" (very satisfied, somewhat satisfied) versus "not satisfied" (neither satisfied nor dissatisfied, somewhat dissatisfied, very dissatisfied).

Statistical analysis

Descriptive statistics were used to describe demographic and clinical characteristics and scores on outcome measures. Continuous data are presented as means with 95% confidence intervals (CI). Categorical data are presented as proportions with exact 95% CI for binomial distributions when appropriate.[30]

Paired two-tailed *t* tests with Bonferroni correction for multiple comparisons were used to compare differences in means of patient-reported outcomes between baseline and the 3-month and 12-month follow-ups. For each area of health listed in question 60 from the AIMS2, changes in the proportions of patients who listed this area as a priority for improvement at baseline and at the 3-month and 12-month assessments were analysed using McNemar tests with Yates continuity correction and Bonferroni adjustment for multiple comparisons.

Results

Between April 2003 and November 2004, 226 patients were enrolled in this part of the study. Of these patients, 173 (77%) completed the AIMS2 at baseline and at the 3-month and 12-month follow-ups. There were no significant differences in baseline age, gender, disease duration, DAS28 scores or Steinbrocker functional class[31] distribution between patients who did and those who did not complete all three AIMS2 questionnaires (data not shown). Data from patients who did not complete all three AIMS2 questionnaires were excluded from further analyses.

Of the included 173 patients, 70% were women. Mean (95% CI) age and disease duration at study entry were 53.2 (51.3 to 55.2) years and 9.9 (8.5 to 11.3) years, respectively. Assessment of disease severity at baseline generally indicated severe RA, with a DAS28 score of 5.5 (5.3 to 5.7). According to the Steinbrocker functional classification, 7% of the patients were classified as class I, 81% as class II and 12% as class III.

Three months after the start of treatment, patient-reported physical disabilities, disease activity and general health were significantly improved (Table 1), as were most aspects of health as measured with the AIMS2. Improvements were most pronounced for physical aspects of health. All improvements remained relatively stable at the 12-month follow-up, with the gradual improvements in self-care, work and level of tension on the AIMS2 becoming significantly different from baseline after 12 months. The proportion of patients who were satisfied with their current general health also significantly increased from 20.8% (15.0 to 27.6) at baseline to a relatively stable 48.6% (40.9 to

56.3, McNemar test, $P < 0.001$) at the 3-month follow-up and 51.4% (43.7 to 59.1) at the 12-month follow-up.

Table 1. Patient-reported outcomes at baseline and the 3-month and 12-month follow-ups

	Outcome, mean (95% CI)		
	Baseline	**3 months**	**12 months**
HAQ-DI (range 0 to 3)	1.4 (1.3 to 1.5)	1.1 (1.0 to 1.2)†	1.1 (1.0 to 1.2)†
RADAI (range 0 to 10)	5.5 (5.2 to 5.8)	3.7 (3.4 to 4.0)†	3.2 (2.9 to 3.5)†
VAS-GH (range 0 to 100)	58.0 (54.4 to 61.6)	42.5 (38.8 to 46.1)†	38.9 (35.1 to 42.6)†
AIMS2 (range 0 to 10)			
Mobility level	2.5 (2.2 to 2.8)	2.0 (1.7 to 2.3)†	2.0 (1.7 to 2.3)†
Walking and bending	5.8 (5.5 to 6.1)	4.8 (4.4 to 5.1)†	4.6 (4.3 to 5.0)†
Hand and finger	4.3 (3.9 to 4.6)	3.2 (2.9 to 3.5)†	3.1 (2.8 to 3.4)†
Arm function	2.8 (2.5 to 3.1)	1.8 (1.6 to 2.1)†	1.7 (1.4 to 2.0)†
Self-care	1.3 (1.1 to 1.6)	1.0 (0.7 to 1.2)	0.8 (0.6 to 1.1)†
Household tasks	2.6 (2.3 to 3.0)	2.0 (1.7 to 2.3)†	2.1 (1.7 to 2.4)†
Social activities	5.1 (4.9 to 5.3)	4.9 (4.8 to 5.1)	4.8 (4.7 to 5.0)
Support from family	2.7 (2.4 to 3.1)	2.6 (2.3 to 3.0)	2.5 (2.1 to 2.9)
Arthritis pain	6.7 (6.4 to 7.1)	4.6 (4.2 to 4.9)†	4.5 (4.1 to 4.9)†
Work*	4.4 (3.6 to 5.1)	3.6 (2.9 to 4.3)	3.1 (2.4 to 3.7)†
Level of tension	3.8 (3.5 to 4.1)	3.5 (3.2 to 3.7)	3.2 (3.0 to 3.5)†
Mood	3.2 (3.0 to 3.5)	2.6 (2.3 to 2.8)†	2.5 (2.3 to 2.7)†

AIMS2, Arthritis Impact Measurement Scales; HAQ-DI, Health Assessment Questionnaire Disability Index; RADAI, Rheumatoid Arthritis Disease Activity Index; VAS-GH, visual analogue scale for general health.

* n = 59.

† Significantly different from baseline after Bonferroni correction for multiple comparisons ($P < 0.05/45$). No significant differences between 3-month and 12-month follow-ups.

The proportions of patients who selected the different areas of health as a priority for improvement during the study period are shown in Table 2. At baseline, arthritis pain was the major priority for improvement, selected by about 90% of the patients. Other priorities were various aspects of physical function, including hand and finger function, walking and bending, and mobility. Almost one-third of the patients chose health status related to work as an important priority. Other aspects of health, including all psychosocial aspects, were selected by <20% of the patients.

At the group level, this priority ranking remained mostly unchanged during treatment (Figure 1). At both the 3-month and 12-month follow-ups, the top six priorities of improvement remained the same, with only minor shifts occurring within the less commonly selected areas of health, such as level of tension and arm function. Some changes were seen in the frequency in which individual areas of health were selected.

Table 2. Patients who listed various areas of health from the AIMS2 as a priority for improvement at baseline and the 3-month and 12-month follow-ups

	Outcome, n (%) (exact 95% binomial CI)		
	Baseline	**3 months**	**12 months**
Mobility level	57 (32.9; 26.0 to 40.5)	52 (30.1; 23.3 to 37.5)	55 (31.8; 24.9 to 39.3)
Walking and bending	73 (42.2; 34.7 to 49.9)	72 (41.6; 34.2 to 49.3)	74 (42.8; 35.3 to 50.5)
Hand and finger function	99 (57.2; 49.5 to 64.7)	83 (48.0; 40.3 to 55.7)	75 (43.4; 35.9 to 51.1)
Arm function	25 (14.5; 9.6 to 20.6)	15 (8.7; 4.9 to 13.9)	15 (8.7; 4.9 to 13.9)
Self-care	11 (6.4; 3.2 to 11.1)	10 (5.8; 2.8 to 10.4)	16 (9.2; 5.4 to 14.6)
Household tasks	28 (16.2; 11.0 to 22.5)	46 (26.6; 20.2 to 33.8)	42 (24.3; 18.1 to 31.4)
Social activities	16 (9.2; 5.4 to 14.6)	12 (6.9; 3.6 to 11.8)	13 (7.5; 4.1 to 12.5)
Support from family	5 (2.9; 0.9 to 6.7)	8 (4.6; 2.0 to 8.9)	10 (5.8; 2.8 to 10.4)
Arthritis pain	153 (88.4; 82.7 to 92.8)	122 (70.5; 63.1 to 77.2)	123 (71.1; 63.7 to 77.7)
Work	50 (28.9; 22.3 to 36.3)	50 (28.9; 22.3 to 36.3)	54 (31.2; 24.4 to 38.7)
Level of tension	20 (11.6; 7.2 to 17.3)	26 (15.0; 10.1 to 21.2)	29 (16.8; 11.5 to 23.2)
Mood	13 (7.5; 4.1 to 12.2)	12 (6.9; 3.6 to 11.8)	20 (11.6; 7.2 to 17.3)

AIMS2, Arthritis Impact Measurement Scales.

Table 3. Association between changes in the priority of pain and mean levels (95% CI) of patient-perceived pain and disease activity

	3 months				**12 months**			
	P/P (n = 112)	**NP/P (n = 10)**	**P/NP (n = 41)**	**NP/NP (n = 10)**	**P/P (n = 96)**	**NP/P (n = 27)**	**P/NP (n = 26)**	**NP/NP (n = 24)**
AIMS2 Pain	5.0 (4.6 to 5.4)	6.2 (4.3 to 8.0)*	3.4 (2.7 to 4.1)†	2.9 (1.5 to 4.3)	5.2 (4.7 to 5.6)	5.0 (4.2 to 5.7)*	3.7 (2.6 to 4.7)†	2.3 (1.7 to 3.0)
RADAI	3.9 (3.5 to 4.2)	5.3 (3.7 to 6.8)*	2.9 (2.3 to 3.5)†	3.3 (2.3 to 4.3)	3.6 (3.2 to 4.0)	3.5 (2.8 to 4.3)*	2.6 (1.8 to 3.3)†	2.0 (1.2 to 2.7)

AIMS2, Arthritis Impact Measurement Scales; NP/NP, no priority to no priority; NP/P, no priority to priority; P/NP, priority to no priority; P/P, priority to priority; RADAI, Rheumatoid Arthritis Disease Activity Index.

* Significantly different from NP/NP ($P < 0.05$).

† Significantly different from P/P ($P < 0.05$).

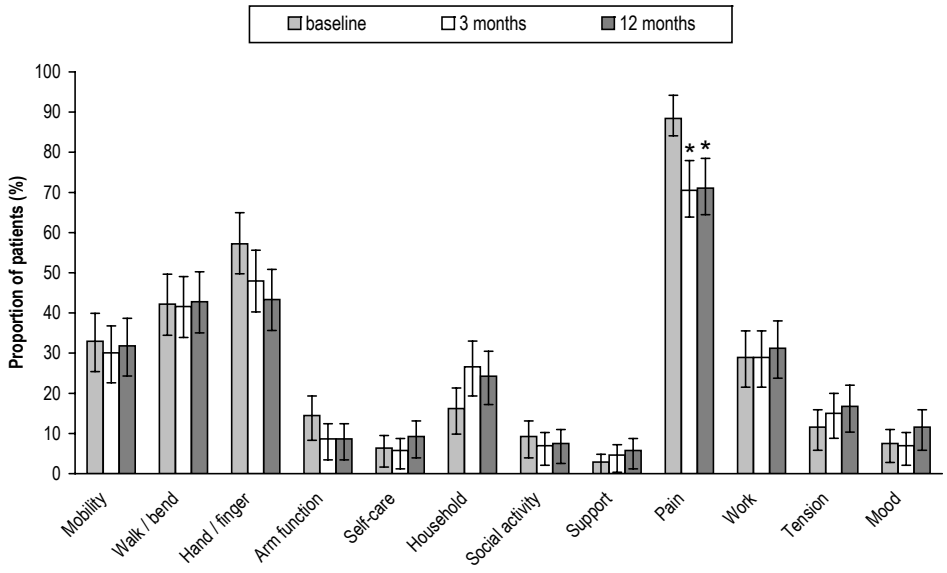

Figure 1. Proportion (exact 95% binomial CI) of patients who listed the different areas of health as a priority for improvement at baseline and the 3-month and 12-month follow-ups. *Significantly different from baseline after Bonferroni correction for multiple comparisons ($P < 0.05/36$); no significant differences between 3-month and 12-month follow-ups.

Most notable were the decreased priority of improvement in hand and finger function and pain and the increased priority of household tasks. However, after Bonferroni correction for multiple comparisons, only the decreased priority of pain reduction retained significance. Although arthritis pain remained the major priority for improvement during the study period, the proportion of patients who selected this health area significantly decreased at the 3-month follow-up and remained stable thereafter.

Although priorities for improvement during treatment were fairly stable at the group level, there was considerable intraindividual variation in priorities over time. The proportion of patients who changed the priority classification of an aspect of health (from either no priority to priority or from priority to no priority) at the two follow-ups ranged between 6.4% and 34.7% for the different aspects of health (see supplementary table W1, available at http://ard.bmjjournals.com/supplemental). From the patients who selected exactly three priorities both at baseline and after 3 months, only 19% selected the same list of priorities on both occasions (12% between 3 and 12 months), 56% made one change in their priority list (56% between 3 and 12 months), 23% selected two new priorities (29% between 3 and 12 months) and 2% selected none of the priorities from their previous list (3% between 3 and 12 months).

Individual changes in the priority status of pain at 3 and 12 months were related to concurrent levels of pain and disease activity (Table 3). Patients who dropped pain

from their priority list reported a significantly lower level of pain and disease activity than patients for whom pain remained a priority for improvement. Conversely, patients who changed pain from no priority to priority reported significantly more pain and disease activity than patients who continued to exclude it as a priority. Changes in the priority of pain improvement were not associated with different scores on the VAS-GH, HAQ-DI and the physical, affect, social and role components of the AIMS2.

Discussion

This is the first study to examine the longitudinal course of patients' priorities for improvement in a cohort of patients with RA during anti-TNF treatment. The results suggest that, at a group level, patients' priority rankings are fairly stable during 1 year of treatment, despite major improvements in health status. Although pain reduction becomes somewhat less important after 3 months of treatment, it remains the highest priority of improvement for patients with RA. At the individual patient level, however, priorities are not stable and appear to be associated with changes in disease state.

Our finding that improvements in pain and aspects of physical function are of primary importance to patients with RA is consistent with previous studies,[7,11–18] although some studies have suggested that physical disability or loss of mobility and dependency on others may be more important problems than pain itself.[32–34] In 1985, Gibson and Clark[12] found that 47% of 120 randomly selected patients with RA rated pain relief and 21% rated increased physical activity as the most desirable objectives of their treatment. A similar study of 250 patients with rheumatic disease showed that 66% of the included 120 patients with RA ranked pain and 22% ranked disability as the most important symptoms to be treated.[14] Both studies, however, focused only on physical aspects of RA and did not include any psychological or social dimensions of health.

Another study of 79 patients with various rheumatic diseases that did include psychosocial aspects of health reported that being free of pain was the symptom status outcome that the majority of patients (63%) identified as the most important outcome of treatment.[7] The feeling of being in control was the mental health outcome rated most important by the largest number of the patients (42%), activities involving the legs the most important physical health outcome (38%) and working at a job or around the house the most important social health outcome (62%). However, as the various dimensions of health were examined separately, this study did not assess the relative weight or importance attached to the physical, psychological and social aspects of health.

Several studies have used the more comprehensive priority list from the AIMS2 to describe patients' priorities for improvement in mostly cross-sectional RA populations. In the validation study of the original AIMS2, 62% of 299 patients with RA designated pain as a priority area.[15] Next were walking and bending (49%), hand and finger func-

tion (47%), and household tasks (30%). Another study reported comparable priorities in 92 patients with RA for pain (67%), walking and bending (41%), and hand and finger function (42%), but also a high priority for mobility (53%).[11] Two previous studies in Dutch patients with RA found similar high priorities for pain (74% and 75%, respectively), walking and bending (52% and 46%), and hand and finger function (41% and 38%), although in both studies household tasks was selected by <20% of the patients.[17,18] Minnock et al[16] found that 68% of 58 women with RA prioritised pain as an area of health needing improvement. Surprisingly, walking and bending (25%) and hand and finger function (25%) were less often selected as a priority in this study, whereas household tasks and mood were selected by 44% and 26%, respectively. Finally, in a recent study of 1024 patients with RA, pain was selected by 69%, walking and bending by 33%, and hand and finger function by 24% of the patients.[13]

The distributions of priorities in our study are reasonably consistent with these studies, with the exception of the relatively high priorities for hand and finger function and pain at baseline. During treatment, however, the proportion of patients that selected hand and finger function and pain as a priority decreased to comparable levels, as observed in the previous observations. Nonetheless, one other notable difference between priorities observed in this study and most other studies remained visible at all three assessments. The patients in this study more commonly selected aspects related to work as a priority for improvement. This may be related to cultural differences in the importance of being able to work, since a previous study in Dutch patients with RA also found a high priority for work-related aspects of health.[18]

This study also confirms the finding of Heiberg et al[22] that pain improvement remains the top priority for patients with RA over time, despite marked improvements in its intensity. Contrary to their findings, however, this study did show a significant decrease in the number of patients that selected pain as an area for improvement after 3 months of treatment. Moreover, within individual patients, priorities often changed and longitudinal changes in the priority for pain improvement were associated with the achieved level of pain and disease activity at the respective follow-up assessments. This gives some support to the idea that the importance of particular outcomes to patients may vary during different disease states and that existing measures may be enhanced by taking account of these variations in priorities.[20,21] However, as the current sample size was too small to permit extensive subgroup analyses, this association has yet to be confirmed for other areas of health.

The finding that pain remained the most selected priority for improvement during treatment may indicate that the improvements in pain (although significant at a group level) were still not large enough to lead to an "acceptable" level of pain for most individual patients. Although no established standards exist for a patient-acceptable symptom state on the AIMS2, the results showed that patients who dropped pain as a

priority for improvement had a mean pain score of about 3.5 on the AIMS2 pain scale. However, at both the 3-month and 12-month follow-ups, <40% of the patients actually achieved a pain score below 3.5.

Limitations

The study has some limitations. The first concerns the use of the priority list of the AIMS2 questionnaire to measure priorities for improvement. This priority list may not include all aspects of health that are important to patients with RA. For instance, the list does not include fatigue and general wellbeing, which have been identified by patients as important outcomes of treatment.[20,35–37] In addition, different dimensions of health are represented by different numbers of items on the priority list, which may have influenced the results. Finally, despite clear instructions to the contrary accompanying the priority list, it was possible for patients to select >3 priorities for improvement. However, as <5% of the patients selected >3 items at the different assessments, this is not likely to have significantly affected the results.

Another limitation concerns the generalisability of the current findings. Medical interventions such as anti-TNF treatment are primarily aimed at improving the pathophysiological processes of inflammation. Consequently, primary signs and symptoms such as pain, swollen and tender joints and impaired function are most likely to improve. Although theoretically, psychosocial aspects are induced by the disease process and should improve also in case of effective therapy, specific psychosocial interventions may very well result in different priority distributions.

Finally, as the duration of this study was limited to 1 year, no causal conclusions can be drawn about long-term changes in patients' priorities. A recent qualitative study in patients with RA suggested that the relative importance of different aspects of health changes as the disease progresses.[20] Patients reported that pain was most important in their early disease, and that mobility and independence were more important in later disease. However, to date there is no quantitative evidence that disease duration has long-term effects on patients' priorities for outcome improvement.[21]

Conclusion

This study suggests that patients' priorities for improvement are fairly stable over time, although individual priorities can change as a result of effective treatment. Pain reduction remains the most important priority for patients with RA, even after 1 year of anti-TNF treatment.

Acknowledgements

We thank T van Gaalen, W Kievit and P Welsing for their contribution to the organisation of the study and data management. We thank the following rheumatologists and research nurses for their assistance in patient recruitment and data collection: J Alberts, C Allaart, A ter Avest, P Barrera Rico, T Berends, H Bernelot Moens, K Bevers, C Bijkerk, A van der Bijl, J de Boer, A Boonen, E ter Borg, E Bos, Botha, A Branten, F Breedveld, H van den Brink, J Bürer, G Bruyn, H Cats, M Creemers, J Deenen, C De Gendt, K Drossaers-Bakker, A van Ede, A Eijsbouts, S Erasmus, M Franssen, I Geerdink, M Geurts, E Griep, E de Groot, C Haagsma, H Haanen, J Harbers, A Hartkamp, J Haverman, H van Heereveld, van de Helm-van Mil, I Henkes, S Herfkens, M Hoekstra, K van de Hoeven, DM Hofman, M Horbeek, F van den Hoogen, PM Houtman, T Huizinga, H Hulsmans, P Jacobs, T Jansen, M Janssen, M Jeurissen, A de Jong, M Kleine Schaar, G Kloppenburg, H Knaapen, P Koelmans, Kortekaas, B Kraft, A Krol, M Kruijssen, D Kuiper-Geertsma, I Kuper, R Laan, J van de Laan, J van Laar, P Lanting, H Lim, S van der Linden, A Mooij, J Moolenburgh, N Olsthoorn, P van Oijen, van Oosterhout, J Oostveen, P van 't Pad Bosch, K Rasing, K Ronday, D de Rooij, L Schalkwijk, P Seys, P de Sonnaville, A Spoorenberg, A Stenger, G Steup, W Swen, J Terwiel, M van der Veen, M Veerkamp, C Versteegden, H Visser, C Vogel, M Vonk, H Vonkeman, A Westgeest, H van Wijk, N Wouters.

Funding

This study was funded by an unrestricted educational grant by Schering-Plough and CVZ (the Dutch Health Care Insurance Board).

References

1. Anderson KO, Bradley LA, Young LD, McDaniel LK, Wise CM. Rheumatoid arthritis: review of psychological factors related to etiology, effects, and treatment. Psychol Bull 1985;98:358–87.
2. Lapsley HM, March LM, Tribe KL, Cross MJ, Courtenay BG, Brooks PM. Living with rheumatoid arthritis: expenditures, health status, and social impact on patients. Ann Rheum Dis 2002;61:818–21.
3. Meenan RF, Yelin EH, Nevitt M, Epstein WV. The impact of chronic disease: a sociomedical profile of rheumatoid arthritis. Arthritis Rheum 1981;24:544–9.
4. Yelin E, Lubeck D, Holman H, Epstein W. The impact of rheumatoid arthritis and osteoarthritis: the activities of patients with rheumatoid arthritis and osteoarthritis compared to controls. J Rheumatol 1987;14:710–7.
5. Gill TM, Feinstein AR. A critical appraisal of the quality of quality-of-life measurements. JAMA 1994;272:619–26.

6. O'Boyle CA, McGee H, Hickey A, O'Malley K, Joyce CR. Individual quality of life in patients undergoing hip replacement. Lancet 1992;339:1088–91.
7. Kwoh CK, Ibrahim SA. Rheumatology patient and physician concordance with respect to important health and symptom status outcomes. Arthritis Rheum 2001;45:372–7.
8. Rothwell PM, McDowell Z, Wong CK, Dorman PJ. Doctors and patients don't agree: cross sectional study of patients' and doctors' perceptions and assessments of disability in multiple sclerosis. BMJ 1997;314:1580–3.
9. Hewlett SA. Patients and clinicians have different perspectives on outcomes in arthritis. J Rheumatol 2003;30:877–9.
10. Hewlett S, Smith AP, Kirwan JR. Values for function in rheumatoid arthritis: patients, professionals, and public. Ann Rheum Dis 2001;60:928–33.
11. Archenholtz B, Bjelle A. Reliability, validity, and sensitivity of a Swedish version of the revised and expanded Arthritis Impact Measurement Scales (AIMS2). J Rheumatol 1997;24:1370–7.
12. Gibson T, Clark B. Use of simple analgesics in rheumatoid arthritis. Ann Rheum Dis 1985;44:27–9.
13. Heiberg T, Kvien TK. Preferences for improved health examined in 1,024 patients with rheumatoid arthritis: pain has highest priority. Arthritis Rheum 2002;47:391–7.
14. McKenna F, Wright V. Pain and rheumatoid arthritis. Ann Rheum Dis 1985;44:805.
15. Meenan RF, Mason JH, Anderson JJ, Guccione AA, Kazis LE. AIMS2. The content and properties of a revised and expanded Arthritis Impact Measurement Scales health status questionnaire. Arthritis Rheum 1992;35:1–10.
16. Minnock P, Fitzgerald O, Bresnihan B. Quality of life, social support, and knowledge of disease in women with rheumatoid arthritis. Arthritis Rheum 2003;49:221–7.
17. Riemsma RP, Taal E, Rasker JJ, Houtman PM, van Paassen HC, Wiegman O. Evaluation of a Dutch version of the AIMS2 for patients with rheumatoid arthritis. Br J Rheumatol 1996;35:755–60.
18. Taal E, Rasker JJ, Evers AW, Kraaimaat FW, Lanting PJH, Jacobs JW. Which priorities have rheumatoid arthritis (RA) patients for their health status improvement? [Abstract]. Arthritis Rheum 1997;40 S231.
19. Carr AJ, Higginson IJ. Are quality of life measures patient centred? BMJ 2001;322:1357–60.
20. Carr A, Hewlett S, Hughes R, et al. Rheumatology outcomes: the patient's perspective. J Rheumatol 2003;30:880–3.
21. Kirwan JR, Hewlett SE, Heiberg T, et al. Incorporating the patient perspective into outcome assessment in rheumatoid arthritis — progress at OMERACT 7. J Rheumatol 2005;32:2250–6.
22. Heiberg T, Finset A, Uhlig T, Kvien TK. Seven year changes in health status and priorities for improvement of health in patients with rheumatoid arthritis. Ann Rheum Dis 2005;64:191–5.
23. Boonen A, Landewe R. Health status in rheumatoid arthritis over 7 years. Ann Rheum Dis 2005;64:173–5.
24. Arnett FC, Edworthy SM, Bloch DA, et al. The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. Arthritis Rheum 1988;31:315–24.
25. Prevoo ML, van 't Hof MA, Kuper HH, van Leeuwen MA, van de Putte LB, van Riel PL. Modified Disease Activity Scores that include twenty-eight–joint counts: development and

validation in a prospective longitudinal study of patients with rheumatoid arthritis. Arthritis Rheum 1995;38:44–8.
26. Fries JF, Spitz P, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. Arthritis Rheum 1980;23:137–45.
27. Siegert CE, Vleming LJ, Vandenbroucke JP, Cats A. Measurement of disability in Dutch rheumatoid arthritis patients. Clin Rheumatol 1984;3:305–9.
28. Stucki G, Liang MH, Stucki S, Bruhlmann P, Michel BA. A self-administered rheumatoid arthritis disease activity index (RADAI) for epidemiologic research. Psychometric properties and correlation with parameters of disease activity. Arthritis Rheum 1995;38:795–8.
29. Fransen J, van Halm VP, Nurmohamed MT, van Riel PL, de Ryck Y, Dijkmans BA. Validity of the rheumatoid arthritis disease activity index (RADAI) in a two year open label study with Leflunomide. Ann Rheum Dis 2005;64 (suppl 3):194
30. Bland M. An introduction to medical statistics. 3rd ed. Oxford: Oxford University Press; 2000.
31. Steinbrocker O, Traeger CH, Battman RG. Therapeutic criteria in rheumatoid arthritis. JAMA 1949;140:659–62.
32. Chamberlain MA, Buchanan JM, Hanks H. The arthritic in an urban environment. Ann Rheum Dis 1979;38:51–6.
33. Cornelissen PG, Rasker JJ, Valkenburg HA. The arthritis sufferer and the community: a comparison of arthritis sufferers in rural and urban areas. Ann Rheum Dis 1988;47:150–6.
34. Taal E, Rasker JJ, Seydel ER, Wiegman O. Health status, adherence with health recommendations, self-efficacy and social support in patients with rheumatoid arthritis. Patient Educ Couns 1993;20:63–76.
35. Kirwan J, Heiberg T, Hewlett S, et al. Outcomes from the Patient Perspective Workshop at OMERACT 6. J Rheumatol 2003;30:868–72.
36. Ahlmen M, Nordenskiold U, Archenholtz B, et al. Rheumatology outcomes: the patient's perspective. A multicentre focus group interview study of Swedish rheumatoid arthritis patients. Rheumatology (Oxford) 2005;44:105–10.
37. Hewlett S, Carr M, Ryan S, et al. Outcomes generated by patients with rheumatoid arthritis: how important are they? Musculoskeletal Care 2005;3:131–42.

Supplementary table W1. Number of patients (%) who did or did not change the priority classification the different aspects of health between baseline and the 3-month follow-up and between the 3-month and 12-month follow-up.

		3 months			12 months		
	Baseline priority for improvement	**No change**	**Change from priority to no priority**	**Change from no priority to priority**	**No change**	**Change from priority to no priority**	**Change from no priority to priority**
Mobility level	57 (32.9)	120 (69.4)	29 (16.8)	24 (13.6)	120 (69.4)	25 (14.5)	28 (16.2)
Walking and bending	73 (42.2)	122 (70.5)	26 (15.0)	25 (14.5)	113 (65.3)	29 (16.8)	31 (17.9)
Hand and finger function	99 (57.2)	113 (65.3)	38 (22.0)	22 (12.7)	117 (67.6)	32 (18.5)	24 (13.9)
Arm function	25 (14.5)	149 (86.1)	17 (9.8)	7 (4.0)	155 (89.6)	9 (5.2)	9 (5.2)
Self-care	11 (6.4)	158 (91.3)	8 (4.6)	7 (4.0)	153 (88.4)	7 (4.0)	13 (7.5)
Household tasks	28 (16.2)	139 (80.3)	8 (4.6)	26 (15.0)	127 (73.4)	25 (14.5)	21 (12.1)
Social activities	16 (9.2)	155 (89.6)	11 (6.4)	7 (4.0)	154 (89.0)	9 (5.2)	10 (5.8)
Support from family	5 (2.9)	162 (93.6)	4 (2.3)	7 (4.0)	159 (91.9)	6 (3.5)	8 (4.6)
Arthritis pain	153 (88.4)	122 (70.5)	41 (23.7)	10 (5.8)	120 (69.4)	26 (15.0)	27 (15.6)
Work	50 (28.9)	127 (73.4)	23 (13.3)	23 (13.3)	119 (68.8)	25 (14.5)	29 (16.8)
Level of tension	20 (11.6)	149 (86.1)	9 (5.2)	15 (8.7)	146 (84.4)	12 (6.9)	15 (8.7)
Mood	13 (7.5)	154 (89.0)	10 (5.8)	9 (5.2)	153 (88.4)	6 (3.5)	14 (8.1)

3 Patient-perceived satisfactory improvement (PPSI): interpreting meaningful change in pain from the patient's perspective

P.M. ten Klooster
K.W. Drossaers-Bakker
E. Taal
M.A.F.J. van de Laar

Pain 2006; 121: 151–157.

Abstract

The assessment of clinically meaningful changes in patient-reported pain has become increasingly important when interpreting results of clinical studies. However, proposed response criteria, such as the minimal clinically important difference, do not correspond with the growing need for information on truly meaningful, individual improvements. The aim of the present study was to investigate satisfactory improvements in pain from the patient's perspective. Data were collected in a 2-week prospective study of 181 arthritis patients treated with a local corticosteroid injection. Baseline and follow-up pain were assessed on 100 mm visual analogue scales for pain intensity (VAS-PI). At baseline, patients also marked a hypothetical level on a VAS-PI representing a satisfactory improvement in pain. Patient-perceived satisfactory improvement (PPSI) was constructed using a 5-point categorical rating of change scale at follow-up as the anchor. PPSI was associated with a minimal reduction of 30 mm or 55% on the VAS-PI. Since absolute change in pain associated with satisfactory improvement proved highly dependent on baseline pain, percent change scores performed better in classifying improved patients. The 55% threshold for satisfactory improvement was consistent over the course of treatment and reasonably consistent across groups of patients. Our data suggest that PPSI is a clinically relevant and stable concept for interpreting truly meaningful improvements in pain from the individual perspective.

Introduction

In recent years, both clinicians and investigators have become increasingly interested in the patient's perspective on the meaning of changes on core outcome measures.[1] A commonly used method to determine thresholds for patient-perceived meaningful change is to compare changes in pain scores with patients' global ratings of the magnitude of change.[2,3] Variations on this approach have been used to define the minimal clinically important difference (MCID) in pain in various clinical settings.[4–9]

A frequently overlooked concern with this approach is that it actually mixes perspectives.[10] Whereas the patient rates the magnitude of change, the investigator determines which rating serves as the cut-off for important or satisfactory improvement. Another concern is that patients are often unable to accurately recall their initial pain and that retrospective self-reports of pain relief do not always reflect true changes in pain.[11–13] Instead of comparing pre-treatment and current pain, these patients seem to focus mainly on the acceptability of their current status when judging the magnitude of change.[14–17]

To address these problems Tubach et al[18] recently suggested to complement the MCID with the patient acceptable symptom state (PASS), an absolute value on the follow-up measure beyond which patients consider themselves well. The PASS does not deal with changes, but concentrates on the concept of achieving a satisfactory state. In this sense, the concept of the patient-derived PASS is very similar to arbitrarily defined or data driven concepts as adequate analgesia[19] and low disease activity state.[20]

Since achieving adequate pain relief is the ultimate goal of pain treatment, the PASS is a clinically relevant concept. Moreover, the patient driven PASS meets the growing need for measures of major improvement from the patients' perspective as opposed to measures of minimal important difference.[1,21] A drawback of its use, however, is that it entails separate analyses for patients achieving a relevant change and patients achieving an acceptable state.

This study presents an investigation into meaningful changes in pain from the patient's perspective that combines the strengths of both the MCID and the PASS. The first objective of the study was to assess the magnitude of change on the VAS-PI that most closely represents patient-perceived satisfactory improvement (PPSI) in arthritis patients with localized musculoskeletal pain. The second objective was to investigate the stability of PPSI across groups of patients. Since patients' perceptions may also change over the course of treatment, the third objective was to examine whether PPSI corresponds with the change in pain that patients before treatment consider necessary for satisfactory improvement.

Methods

Patient selection and study design

Participants were recruited at the outpatient rheumatology clinic. All consecutive patients with localized musculoskeletal pain and an indication for a local corticosteroid injection were asked to participate. Patients were excluded if they were aged <16 years or unable to mark a visual analogue scale for pain intensity (VAS-PI). The study did not interfere with usual treatment.

Prior to the injection, patients indicated the average level of localized pain in the past week on a 100 mm, unmarked VAS-PI with endpoints "no pain" and "unbearable pain". Subsequently, patients marked the level of pain that would represent a satisfactory improvement on a separate VAS-PI. After 2 weeks, a follow-up questionnaire was mailed to the patients. After marking the VAS-PI for pain in the past week, patients judged the change in pain by answering the following question: "Compared to 2 weeks ago (before the local injection) the pain in the injected area is…" The response categories were "worse", "unchanged", "unsatisfactory improved", "satisfactory improved" and "good to very good improved".

Analyses

Statistics. Statistical analyses were performed using SPSS 11.0 for Windows. The valid use of parametric statistics was verified by testing for normal distribution of the variables (Kolmogorov–Smirnov test, normal distribution assumed when $P > 0.05$). When the assumption of normality was not met, non-parametric statistics were used. P values <0.05 were considered to indicate statistical significance. The mean and standard deviation (SD) were used for descriptive statistics unless otherwise specified.

Patients' judgments of change. To study meaningful changes in pain, an anchoring method based on the patient's judgment of change at follow-up was used. This categorical rating scale, however, has not been previously validated. Supporting evidence for its valid use as an external anchor would be an appreciable relationship between patients' ratings of change and actual changes on the VAS-PI.[16] To explore this relationship, the categorical ratings were compared with absolute change scores (VAS-PI follow-up – VAS-PI baseline) and percent change scores ((absolute change / VAS-PI baseline) x 100)) by means of one-way analyses of variance (ANOVAs) followed by post hoc multiple comparisons (Bonferroni adjustment). Secondly, Spearman rank correlation coefficients of the categorical rating scale with absolute and percent change in pain on the VAS-PI were calculated. Correlations ≥0.5 were considered indicative for the valid use of the rating scale.[15,22]

Patient-perceived satisfactory improvement on the VAS-PI. Patient-perceived satisfactory improvement (PPSI) was defined as the change in pain on the VAS-PI associated with a minimal rating of satisfactory improvement at follow-up. Ratings of "satisfactory improved" and "good to very good improved" were pooled to define satisfactory improved patients. Patients were considered unimproved when they rated themselves as worsened, unchanged or unsatisfactory improved. To evaluate the change in pain that was most closely associated with PPSI, receiver operating characteristic (ROC) curves were computed for both absolute and percent change scores.[23,24] As opposed to the analyses of group means, as suggested by Jaeschke et al,[25] ROC analysis offers the opportunity to study patient-perceived improvement at the individual level. An area under the ROC curve (AUC) ≥0.7 was considered adequately accurate in classifying satisfactory improved patients.[26] The change score with the highest combination of sensitivity and specificity was selected as the optimal cut-off point for PPSI. The comparative accuracy of absolute and percent change scores was determined by comparing the areas under the curve.[27]

Consistency of PPSI over groups of patients. The consistency of PPSI across baseline demographic and clinical variables was investigated using the data of patients who rated their pain as satisfactory or good to very good improved. Dependency of absolute change on baseline VAS-PI scores was determined by linear regression analysis. The consistency of absolute change over age and disease duration was assessed using Pearson correlation coefficients and the differences between men and women and between the five primary diagnoses were investigated using an independent *t* test and a one-way ANOVA. Since percent change in these patients was not normally distributed, the non-parametric Spearman rank correlation coefficient, Mann–Whitney *U* test and Kruskal–Wallis *H* tests, were used to assess the stability of percent change scores.

Consistency of PPSI over the course of treatment. To assess whether patient's perceptions of satisfactory improvement had changed over the course of treatment, the mean actual change scores of improved patients were compared with the mean change scores patients initially judged necessary to be satisfied. The agreement between actual and initially defined change scores of satisfactory improved patients was calculated using the intraclass correlation coefficient (ICC). ICCs were considered excellent when >0.75, fair to good when ≥0.40 ≤0.75 and poor when <0.40.[28] Since the ICC does not provide information on the magnitude of within-person differences, a Bland–Altman plot of the difference against the average of the actual and initially defined change scores was constructed.[29]

Results

Patient characteristics

Between May and December 2004, 200 patients agreed to participate in the study and completed the baseline questionnaire. Despite sending reminders, 6 follow-up questionnaires were not returned. Thirteen follow-up questionnaires were not interpretable. Descriptive baseline characteristics of the included patients are listed in Table 1.

Table 1. Baseline demographic and clinical characteristics (N = 181)

Age in years (mean ± SD)	59.5 ± 14.7
Gender (% female/male)	70.7/29.3
Primary diagnosis	
Rheumatoid arthritis (%)	37.0
Osteoarthritis (%)	17.7
Psoriatic arthritis (%)	8.8
Tendinitis/bursitis (%)	8.3
Other (%)	28.2
Disease duration (median, range)	5, 0–52
VAS-PI (mean ± SD)	58.6 ± 24.0

There were no significant differences in baseline VAS-PI scores between patients based on primary diagnosis (ANOVA) and baseline pain was not related to age (Pearson *r*). Women tended to report more pain then men, although this difference was not significant (60.9 ± 24.1 vs. 53.2 ± 23.1 mm, $t(179) = -1.96$, $P = 0.052$). Patients with longer disease duration reported more pain (Spearman $r = 0.14$, $P < 0.05$).

Patients' judgments of change

ANOVAs showed that absolute and percent change scores on the VAS-PI were significantly different – in the expected direction – between groups based on the patients' ratings of change (Table 2). Both absolute and percent change scores were significantly different between satisfactory improved patients and worsened, unchanged, or unsatisfactory improved patients. The association between patient-perceived ratings of change and actual change scores was supported by moderate (Spearman $r = -0.51$, $P < 0.001$) to good (Spearman $r = -0.70$, $P < 0.001$) correlation for absolute and percent change, respectively.

Patient-perceived satisfactory improvement on the VAS-PI

Figure 1 presents the ROC curves for absolute and percent change on the VAS-PI at 2-week follow-up, associated with patients' ratings of satisfactory or good to very good improvement. Both absolute and percent change scores had good diagnostic power in identifying satisfactory improved patients, with AUCs of 0.80 (95% CI: 0.73–0.85, $P < 0.0001$) and 0.86 (95% CI: 0.80–0.91, $P < 0.0001$), respectively. The better overall accuracy of PPSI expressed as a percent change score was represented by a significantly higher

Table 2. Group level analysis of mean change in pain on the VAS-PI (in mm) at follow-up associated with categories of patient-perceived rating of change (N = 181)

	Absolute change (mean ± SD)[a]	Percent change (mean ± SD)[b]
Worsened (n = 3)	16.3 ± 21.0$_a$	35.5 ± 42.8$_a$
Unchanged (n = 17)	–2.8 ± 18.9$_a$	–6.7 ± 38.2$_{a,b}$
Unsatisfactory improved (n = 49)	–16.3 ± 19.8$_a$	–22.7 ± 31.3$_b$
Satisfactory improved (n = 76)	–37.2 ± 25.4$_b$	–56.1 ± 34.6$_c$
Good to very good improved (n = 35)	–43.5 ± 23.5$_b$	–85.6 ± 15.5$_d$

[a] One-way ANOVA: $F(4,175)=18.0$, $P < 0.001$.

[b] One-way ANOVA: $F(4,175)=34.5$, $P < 0.001$. Means in the same column that do not have the same subscript differ at $P < 0.05$ (Bonferroni adjustment).

AUC for percent change scores ($P < 0.05$). The optimal cut-off point for an absolute change in pain was –30 mm, corresponding to a sensitivity of 0.68 (95% CI: 0.58–0.76) and specificity of 0.84 (95% CI: 0.73–0.92). The best cut-off for a percent change from baseline was –54.6%, with a sensitivity of 0.74 (95% CI: 0.65–0.82) and a specificity of 0.91 (95% CI: 0.82–0.97).

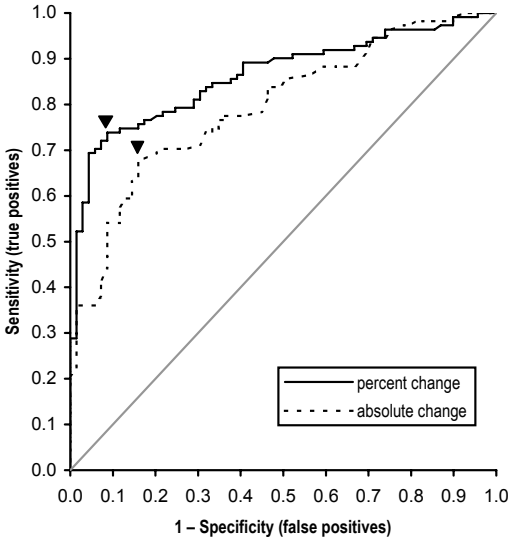

Figure 1. Receiver operating characteristic curves for absolute and percent change in pain on the VAS-PI at 2-week follow-up associated with PPSI (N = 181). ▼ = Optimal cut-off point: absolute change = –30 mm (sensitivity 0.68, specificity 0.84); percent change = –54.6% (sensitivity 0.74, specificity 0.91). AUC absolute change = 0.80 (95% CI: 0.73–0.85); AUC percent change = 0.86 (95% CI: 0.80–0.91); $P < 0.001$ for difference between AUCs.

Consistency of PPSI over groups of patients

The results from the ROC analyses indicated that percent change scores performed better in identifying satisfactory improved patients than absolute change scores. This dependency of PPSI on baseline pain was confirmed by analysis of the change scores of satisfactory and good to very good improved patients (n = 111). The relation between absolute change in pain and baseline pain is illustrated in Figure 2. Patients with high baseline pain required greater absolute reductions in pain to reach a satisfactory improvement ($r^2 = 0.58$, $P < 0.001$). The magnitude of both absolute and percent change in pain was not related to age or disease duration and did not vary between groups based on primary diagnosis. However, absolute change scores in female patients were significantly larger than those in male patients (–42.0 ± 25.4 vs. –31.2 ± 22.1 mm, $t(109) = 2.03$, $P < 0.05$). Percent change scores did not significantly differ between men and women.

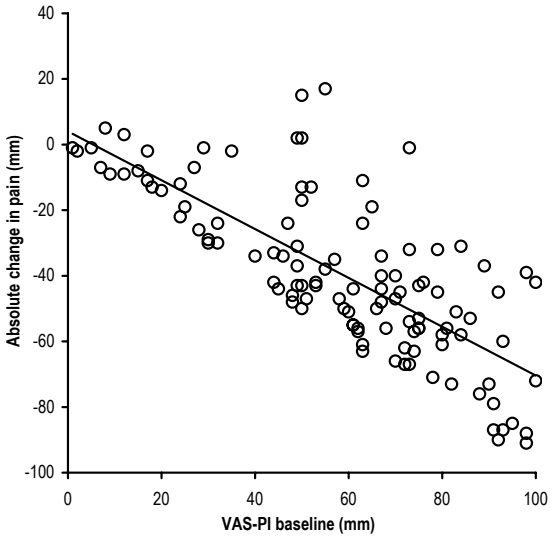

Figure 2. Scatter plot of absolute change in pain in satisfactory or good to very good improved patients related to baseline pain intensity (n = 111). The straight line represents the linear regression line through the data points ($r^2 = 0.58$, $P < 0.001$), demonstrating the dependency of PPSI on baseline pain.

Consistency of PPSI over the course of treatment

The absolute change in pain that patients (N = 181) initially considered necessary to achieve a satisfactory improvement was –32.0 ± 19.7 mm, corresponding to a percent change of –54.7 ± 27.8%. The actual change scores of satisfactory improved patients (n = 76) were adequately correlated with the initially defined satisfactory change scores (ICC = 0.61). However, Bland–Altman analysis (Figure 3) showed that actual change scores were systematically larger than initially defined change scores (mean difference:

6.1 ± 19.9 mm; paired *t* test, $P < 0.05$). As expected, this difference was weakly related to the magnitude of actual change in pain (Pearson $r = 0.23$, $P < 0.05$), indicating that patients with relatively high actual change scores had improved more than they initially considered necessary for satisfactory improvement. The difference between initially defined and actual change in pain was highly variable, as represented by the relatively wide limits of agreement (±39.0 mm) of the Bland–Altman plot. Since systematic bias was only moderate, the predominant source of error, however, was due to random variation instead of a systematic difference between actual and initially defined satisfactory improvement.

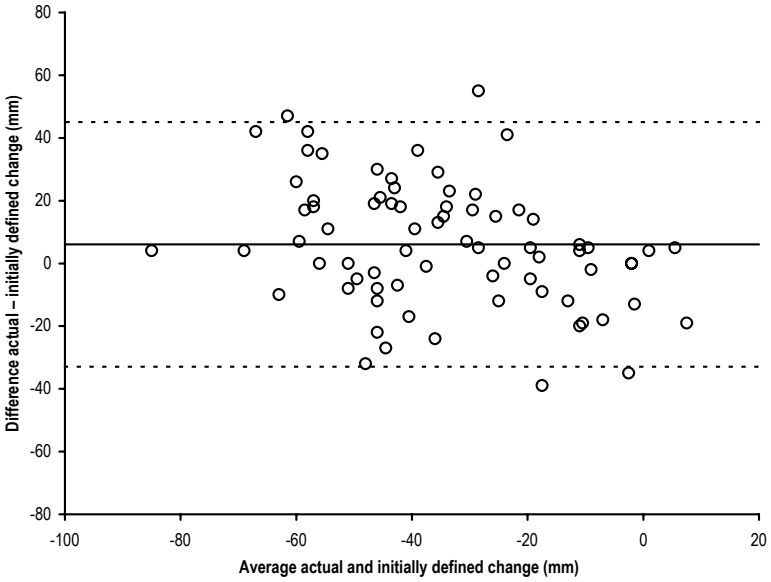

Figure 3. Agreement between actual and initially defined change for satisfactory improvement within satisfactory improved patients (n = 76). The dashed lines represent the 95% limits of agreement (–32.9 and 45.1 mm). The horizontal solid line represents the mean difference between both change scores (6.1 mm).

Discussion

This study presents the PPSI as a new outcome for individual, within-person improvement in pain intensity. Defining meaningful improvement is becoming increasingly important in interpreting the effectiveness of the treatment of pain. However, currently used data driven constructs for identifying clinical improvements such as the MCID do not satisfy the need for information on relevant changes from the patient's perspective. PPSI is assessed using patients' judgments of satisfactory change as the only criterion and prevents arbitrarily chosen cut-off points on the external anchor.

Moreover, it gives a better representation of relevant change since patients tend to judge changes based on the acceptability of their present state. As such, it allows for a patient-centred approach in determining thresholds for true meaningful change, which combines the strengths of both the MCID and the PASS.

In the present study, the threshold for PPSI in musculoskeletal pain was best represented by a decrease on the VAS-PI of at least 55% or 30 mm. This threshold is characterized by a high sensitivity and specificity, supporting the responsiveness of the VAS-PI in measuring musculoskeletal pain. The magnitude of change required to achieve PPSI is considerably larger than most current definitions of meaningful change, such as the proposed 30% improvement criterion.[30] In fact, it corresponds more closely with the formally most often used 50% pain relief threshold.[31,32] Although this higher threshold is neither supported by empirical research[33] nor has its importance to patients been established,[34] a 50% reduction in pain does have clinical intuitive appeal as a threshold for satisfactory improvement.[31] The difference in magnitude between PPSI and the previously established MCIDs on the VAS-PI could have several possible explanations. The difference could be related to the specific clinical setting of this study, the patients' demographic characteristics or diverging patient expectations. A more likely explanation, however, is that patients are not as easily satisfied with an improvement in pain as investigators are. Changes in pain may need to exceed the cut-offs defined by investigators to be considered satisfactory by patients. Evidence supporting this assumption is that the 55% or 30 mm cut-off for satisfactory improvement is in close accordance with recent studies examining patient-perceived, relevant improvements on the VAS-PI. Concepts defined as "adequate pain treatment",[35] "important improvement or recovery"[36] or "considerable improvement"[37] show similar cut-off points. As such, the cut-off for satisfactory improvement seems to answer the growing need for definite, relevant response criteria as opposed to minimal detectable responses.[21,38–40]

The present study confirms that patient-perceived improvement is not uniformly distributed over the range of the VAS. Whereas initially important improvements were considered to be absolute values,[41,42] more recently it was shown that the magnitude of a MCID increases as baseline pain intensity increases.[6,7,43,44] This dependency on baseline pain status also applies to satisfactory improvements. Patients with high baseline pain need larger reductions in pain to consider themselves satisfactory improved. The ROC analyses also indicate that the diagnostic accuracy of the VAS-PI in discriminating between satisfactory and not-satisfactory improved patients increases when change scores are expressed as a percent change from baseline. Like MCID, PPSI is thus best represented as a percent change from baseline.

The magnitude of a satisfactory improvement proves to be consistent across groups of patients, except for gender. The lower absolute value of PPSI in men can be partly explained by their lower baseline pain scores, since percent change scores were more

consistent between men and women. PPSI is also consistent over the course of treatment. Retrospective judgements of satisfactory improvement are adequately correlated with the change in pain patients at baseline consider satisfactory. The relatively wide limits of agreement between actual and initially defined satisfactory change scores may be due to the inherent problem of high measurement error associated with the use of visual analogue scales.[45–47]

The results of the study support the valid use of the 5-point categorical rating scale as an anchor to assess PPSI. The rating scale allows for a clear distinction between satisfactory and unsatisfactory improved patients. Moreover, the categorical rating scale correlates adequately with the absolute change on the VAS-PI and good with percent change from baseline. However, the assumption that the categorical rating scale is also a reliable standard for measuring change could be a concern. The design of the study did not allow for an assessment of the reliability of this scale. This is a common problem for global rating scales, since internal consistency (Cronbach's α) cannot be computed for a single-item scale and test-retest studies are often complicated or impractical.[17] Future studies are required using this scale on successive occasions in patients with a stable VAS-PI after the first follow-up, in order to assess the test-retest reliability of the categorical rating scale.

Another concern is the exclusive focus on improvements in pain. Since only three patients indicated an increase in pain, clinically meaningful deteriorations on the VAS-PI could not be calculated. The magnitude of the change that patients perceive as meaningful may differ between improvements and deteriorations.[16,22,48] The goal of this investigation and most clinical studies, however, was to study important improvement since this is the result that clinicians and researchers are usually most interested in.

Moreover, the correlation between actual changes in satisfactory improved patients and initially defined changes may have been influenced by a testing effect, i.e. patients may have recalled the position they originally marked on the VAS-PI that would constitute satisfactory improvement.

A final issue concerns the generalizability of the findings. In the current sample, only patients who were treated with a corticosteroid injection were included. These injections are usually administered to patients who experience an exacerbation of pain. The relatively acute nature of their pain may have influenced patients' ratings of their pain and improvement. To determine the generalizability of the study, the findings should be confirmed in different clinical settings. Moreover, the magnitude of PPSI may very well differ for other outcome domains, such as physical functioning, global health status or quality of life. Since the procedures for assessing PPSI can be applied to all patient-reported outcomes, meaningful improvements from the patient's perspective can also be determined for these outcome domains.

In conclusion, PPSI is a clinically relevant and stable concept and can be used to assess true meaningful change in pain from the patient's perspective. Its straightforward character and analyses allows for the unambiguous assessment of satisfactory improved patients. The application of this measure in future clinical studies could lead to new standards for defining clinically meaningful improvement in pain and other outcome domains.

Acknowledgements

The authors thank the respondents who participated in this study and the rheumatologists of Medisch Spectrum Twente for their help in recruiting patients.

References

1. Wells G, Anderson J, Beaton D, et al. Minimal clinically important difference module: summary, recommendations, and research agenda. J Rheumatol 2001;28:452–4.
2. Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. J Clin Epidemiol 2003;56:395–407.
3. Deyo RA, Patrick DL. The significance of treatment effects: the clinical perspective. Med Care 1995;33:AS286–91.
4. Cepeda MS, Africano JM, Polo R, Alcala R, Carr DB. What decline in pain intensity is meaningful to patients with acute pain? Pain 2003;105:151–7.
5. Dhanani S, Quenneville J, Perron M, Abdolell M, Feldman BM. Minimal difference in pain associated with change in quality of life in children with rheumatic disease. Arthritis Rheum 2002;47:501–5.
6. Farrar JT, Young JP, LaMoreaux L, Werth JL, Poole RM. Clinical importance of changes in chronic pain intensity measured on an 11-point numerical pain rating scale. Pain 2001;94:149–58.
7. Jensen MP, Chen C, Brugger AM. Interpretation of visual analog scale ratings and change scores: a reanalysis of two clinical trials of postoperative pain. J Pain 2003;4:407–14.
8. Salaffi F, Stancati A, Silvestri CA, Ciapetti A, Grassi W. Minimal clinically important changes in chronic musculoskeletal pain intensity measured on a numerical rating scale. Eur J Pain 2004;8:283–91.
9. Stratford PW, Binkley JM, Riddle DL, Guyatt GH. Sensitivity to change of the Roland-Morris Back Pain Questionnaire: part 1. Phys Ther 1998;78:1186–96.
10. Beaton DE, Bombardier C, Katz JN, Wright JG. A taxonomy for responsiveness. J Clin Epidemiol 2001;54:1204–17.
11. Feine JS, Lavigne GJ, Dao TT, Morin C, Lund JP. Memories of chronic pain and perceptions of relief. Pain 1998;77:137–41.
12. Fischer D, Stewart AL, Bloch DA, Lorig K, Laurent D, Holman H. Capturing the patient's view of change as a clinical outcome measure. JAMA 1999;282:1157–62.

13. Haas M, Nyiendo J, Aickin M. One-year trend in pain and disability relief recall in acute and chronic ambulatory low back pain patients. Pain 2002;95:83–91.
14. Aseltine RH, Jr., Carlson KJ, Fowler FJ, Jr., Barry MJ. Comparing prospective and retrospective measures of treatment outcomes. Med Care 1995;33:AS67–76.
15. Guyatt GH, Norman GR, Juniper EF, Griffith LE. A critical look at transition ratings. J Clin Epidemiol 2002;55:900–8.
16. Guyatt GH, Osoba D, Wu AW, Wyrwich KW, Norman GR. Methods to explain the clinical significance of health status measures. Mayo Clin Proc 2002;77:371–83.
17. Norman GR, Stratford P, Regehr G. Methodological problems in the retrospective computation of responsiveness to change: the lesson of Cronbach. J Clin Epidemiol 1997;50:869–79.
18. Tubach F, Ravaud P, Baron G, et al. Evaluation of clinically relevant states in patient reported outcomes in knee and hip osteoarthritis: the patient acceptable symptom state. Ann Rheum Dis 2005;64:34–7.
19. Benedetti F, Vighetti S, Amanzio M, et al. Dose-response relationship of opioids in nociceptive and neuropathic postoperative pain. Pain 1998;74:205–11.
20. van der Heijde DM, van 't Hof MA, van Riel PL, et al. Judging disease activity in clinical practice in rheumatoid arthritis: first step in the development of a disease activity score. Ann Rheum Dis 1990;49:916–20.
21. Wolfe F, Michaud K, Strand V. Expanding the definition of clinical differences: from minimally clinically important differences to really important differences. Analyses in 8931 patients with rheumatoid arthritis. J Rheumatol 2005;32:583–9.
22. Cella D, Hahn EA, Dineen K. Meaningful change in cancer-specific quality of life scores: differences between improvement and worsening. Qual Life Res 2002;11:207–21.
23. Deyo RA, Centor RM. Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance. J Chronic Dis 1986;39:897–906.
24. Ward MM, Marx AS, Barry NN. Identification of clinically important changes in health status using receiver operating characteristic curves. J Clin Epidemiol 2000;53:279–84.
25. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. Control Clin Trials 1989;10:407–15.
26. Grzybowski M, Younger JG. Statistical methodology: III. Receiver operating characteristic (ROC) curves. Acad Emerg Med 1997;4:818–26.
27. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983;148:839–43.
28. Fleiss JL. The design and analysis of clinical experiments. New York: John Wiley & Sons; 1986.
29. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;i:307–10.
30. Dworkin RH, Turk DC, Farrar JT, et al. Core outcome measures for chronic pain clinical trials: IMMPACT recommendations. Pain 2005;113:9–19.
31. Moore A, McQuay H, Gavaghan D. Deriving dichotomous outcome measures from continuous data in randomised controlled trials of analgesics. Pain 1996;66:229–37.
32. Scott DL, Dacre JE, Greenwood A, Treasure L, Huskisson EC. Can we develop simple response criteria for slow acting antirheumatic drugs? Ann Rheum Dis 1990;49:196–8.

33. Seres JL. The fallacy of using 50% pain relief as the standard for satisfactory pain treatment outcome. Pain Forum 1999;8:183–88.
34. Farrar JT. What is clinically meaningful: outcome measures in pain clinical trials. Clin J Pain 2000;16:S106–12.
35. Lee JS, Hobden E, Stiell IG, Wells GA. Clinically important change in the visual analog scale after adequate pain control. Acad Emerg Med 2003;10:1128–30.
36. Giraudeau B, Rozenberg S, Valat JP. Assessment of the clinically relevant change in pain for patients with sciatica. Ann Rheum Dis 2004;63:1180–1.
37. Kvien TK, Dougados M, Mowinckel P, Skomsvoll JF, Mikkelsen K. Values for Minimal Clinically Important Improvement (MCII) in patients with Ankylosing Spondylitis (Poster). In: Annual Scientific Meeting of the American College of Rheumatology; 2004; San Antonio, Texas; 2004.
38. Felson DT, Anderson JJ. A review of evidence on the discriminant validity of outcome measures in rheumatoid arthritis. J Rheumatol 2001;28:422–6.
39. Kelly AM. Setting the benchmark for research in the management of acute pain in emergency departments. Emerg Med (Fremantle) 2001;13:57–60.
40. Tugwell P, Boers M, Brooks PM, Simon L, Strand CV. OMERACT 5: International consensus conference on outcome measures in rheumatology. J Rheumatol 2001;28:395–7.
41. Kelly AM. The minimum clinically significant difference in visual analogue scale pain score does not differ with severity of pain. Emerg Med J 2001;18:205–7.
42. Todd KH. Patient-oriented outcome measures: the promise of definition. Ann Emerg Med 2001;38:672–4.
43. Bird SB, Dickson EW. Clinically significant changes in pain along the visual analog scale. Ann Emerg Med 2001;38:639–43.
44. Tubach F, Ravaud P, Baron G, et al. Evaluation of clinically relevant changes in patient reported outcomes in knee and hip osteoarthritis: the minimal clinically important improvement. Ann Rheum Dis 2005;64:29–33.
45. DeLoach LJ, Higgins MS, Caplan AB, Stiff JL. The visual analog scale in the immediate postoperative period: intrasubject variability and correlation with a numeric scale. Anesth Analg 1998;86:102–6.
46. Kropmans TJ, Dijkstra PU, Stegenga B, Stewart R, de Bont LG. Repeated assessment of temporomandibular joint pain: reasoned decision-making with use of unidimensional and multidimensional pain scales. Clin J Pain 2002;18:107–15.
47. Lassere MN, van der Heijde D, Johnson KR, Boers M, Edmonds J. Reliability of measures of disease activity and disease damage in rheumatoid arthritis: implications for smallest detectable difference, minimal clinically important difference, and analysis of treatment effects in randomized controlled trials. J Rheumatol 2001;28:892–903.
48. Hays RD, Woolley JM. The concept of clinically meaningful difference in health-related quality-of-life research. How meaningful is it? Pharmacoeconomics 2000;18:419–23.

4 Can we assess baseline pain and global health retrospectively?

P.M. ten Klooster
K.W. Drossaers-Bakker
E. Taal
M.A.F.J. van de Laar

Clinical and Experimental Rheumatology 2007;25:176–181.

Abstract

Objective. To study the agreement between patients' actual baseline assessments of pain and global health before treatment and retrospective assessments collected 2 weeks after treatment.

Methods. Data were collected in a prospective study of 200 rheumatology outpatients treated with a local corticosteroid injection. At baseline and 2-week follow-up, localized pain and global health were assessed on 100 mm visual analogue scales. The follow-up questionnaire was extended with a retrospective assessment of pain and global health before treatment.

Results. At follow-up patients slightly overestimated the severity of pain and global health before treatment. Actual and retrospective assessments were adequately correlated (pain: $r_s = 0.73$; global health: $r_s = 0.67$). Bland–Altman analysis showed that both pain and global health were characterized by high intra-individual variation between actual and retrospective assessments, with the 95% limits of agreement (–37.3 to 32.3 mm for pain and –49.7 to 37.8 mm for global health) far exceeding proposed values for minimal clinically important differences.

Conclusion. Over a 2-week interval, patients' retrospective assessments of baseline pain and global health are fairly accurate and adequately correlated with actual baseline scores. At the group level, retrospective assessments can provide acceptable data on baseline pain and global health. The wide variability between actual and retrospective assessments, however, indicates that even over short time intervals there is poor individual agreement between the two methods.

Introduction

The assessment of changes in patient-perceived pain and global health plays a key role in both clinical trials and routine practice. In clinical practice, physicians often rely on patients' retrospective accounts of previous states or perceived changes in state to evaluate the effectiveness of treatment. In clinical trials, on the other hand, retrospective measurement is usually discouraged[1] and patients' retrospective perceptions of change or baseline states are rarely measured. However, prospective research designs are usually expensive and time consuming, and sometimes impractical or even impossible.[2,3] In these situations, retrospective assessments of baseline health states collected at follow-up could provide an attractive alternative, provided that these assessments yield reasonably accurate data.

The main concern with retrospective research designs is the extent to which patients are able to accurately recall their symptoms or overall health before treatment.[1,4–6] In patients with arthritis, pain is the most prominent symptom and is best measured with a visual analogue scale (VAS).[7] Several studies have investigated the accuracy of pain recall, but their findings vary considerably. Whereas some found that patients are quite able to recall previous pain states,[8–12] others concluded that recall is inaccurate or systematically biased.[13–19]

Besides patient-perceived pain, the VAS for patient global health status has become a central outcome measure in rheumatology. In contrast to pain recall, however, very little is known about patients' ability to remember previous global health states. Two studies that have examined similar constructs, indicate that recall of global health may be susceptible to error and bias.[20,21] Moreover, it would seem plausible that patients generally will have more difficulties in accurately recalling general health states than concrete symptoms such as pain.[6,21]

One important factor in recalling pre-treatment pain or global health is the time between the actual and the retrospective assessment. Most studies on pain recall in chronic pain patients have used long time intervals between both assessments, ranging from several months to years. Since errors in pain recall generally get worse with the passage of time,[22–24] retrospective assessments after a relatively short time interval may yield sufficiently reliable data.

Finally, an additional drawback of studies comparing actual and retrospective assessments is their reliance on comparison of means and correlation analysis, which are likely to overestimate the actual agreement. A more informative measure of agreement was developed by Bland and Altman,[25] who suggested to plot the absolute individual differences between both methods against their mean and comparing their 95% limits of agreement with a clinically acceptable difference between the two methods.

The aim of the present study was therefore to examine the agreement between patients' actual assessments of baseline pain and global health and retrospective assessments collected after a relatively short period of 2 weeks, using additional Bland–Altman analyses.

Materials and methods

Patients

The data for this study were collected at the outpatient rheumatology clinic. Arthritis patients older than 16 years who experienced localized musculoskeletal pain and who were treated with a local corticosteroid injection were eligible for inclusion. Informed consent was obtained from all patients.

Measurements

The study consisted of two serial assessments. The baseline assessment was completed during the patient's visit at the outpatient clinic, just before the injection procedure. The 2-week follow-up questionnaire was mailed to the patients. At baseline and follow-up, average localized pain and global health in the past week were measured on 100 mm, unmarked VASs, anchored by "no pain – unbearable pain" and "very well – very poor." At the end of the follow-up questionnaire, patients were asked to recall their average level of pain and global health in the week before the injection on identical VASs (e.g., In general, how much pain did you experience in the affected joint in the week before the local injection?).

Statistical analysis

Normal distribution of age, disease duration, VAS scores and differences between actual and retrospective VAS scores was examined by the Kolmogorov–Smirnov (K–S) test and inspection of normality plots. Since several VAS scores were not normally distributed (K–S, $P < 0.05$), all comparisons were conducted using non-parametric tests. Differences between actual and retrospective assessments were tested using paired Wilcoxon signed ranks tests, with Hodges–Lehmann estimates for median differences and 95% confidence intervals (CI). Correlations between actual and retrospective assessments were expressed using Spearman's rank correlation coefficient (r_s). Individual agreement between the two methods of baseline assessment was assessed by plotting the difference between both assessments against their mean.[25]

Results

Patient characteristics

In the period between May and December 2004, 200 consecutive patients were recruited. Six patients (3%) did not return the follow-up questionnaire and 13 patients (6.5%) did not complete the retrospective assessments. Data from these patients were excluded from further analyses. Baseline characteristics of the excluded patients did not differ from the included patients. The descriptive characteristics of the 181 included patients are shown in Table 1.

Table 1. Patient baseline characteristics and actual baseline, follow-up and retrospective baseline VAS scores from the included patients

Age (years), median (IQR)	60 (51–71)
Female, n (%)	128 (71)
Primary diagnosis	
Rheumatoid arthritis, n (%)	67 (37.0)
Osteoarthritis, n (%)	33 (18.2)
Psoriatic arthritis, n (%)	16 (8.8)
Tendinitis / bursitis, n (%)	15 (8.3)
Other, n (%)*	50 (27.6)
Disease duration (years), median (IQR)	4 (0–11)
Baseline pain (VAS, 0–100 mm), median (IQR)	61.0 (46.0–78.0)
Follow-up pain (VAS, 0–100 mm), median (IQR)	25.5 (10.0–47.0)
Retrospective baseline pain (VAS, 0–100 mm), median (IQR)	67.0 (45.5–79.0)†
Baseline global health (VAS, 0–100 mm), median (IQR)	38.0 (10.5–59.0)
Follow-up global health (VAS, 0–100 mm), median (IQR)	31.0 (9.0–48.0)
Retrospective baseline global health (VAS, 0–100 mm), median (IQR)	46.0 (20.5–63.5)‡

IQR: interquartile range, VAS: visual analogue scale.

* Includes several diagnoses such as polymyalgia rheumatica, shoulder complaints, and gout.

† Significantly different from actual baseline pain, Wilcoxon (2-tailed), $Z = -2.02$, Hodges–Lehmann estimated median difference –2.5 mm, 95% CI: –4.5 to 0, $P = 0.044$.

‡ Significantly different from actual baseline global health, Wilcoxon (2-tailed), $Z = -3.60$, Hodges–Lehmann estimated median difference –5.0 mm, 95% CI: –8.0 to –2.5, $P < 0.001$.

Difference between actual and retrospective baseline assessments

Two weeks after treatment patients slightly overestimated the severity of their baseline pain (estimated median difference –2.5, 95% CI: –4.5 to 0) and global health (estimated median difference –5.0, 95% CI: –8.0 to –2.5). The difference between actual and retrospective assessments was correlated with the respective actual level of pain or health before treatment (pain $r_s = 0.28$, 95% CI: 0.14 to 0.41; global health $r_s = 0.34$, 95% CI: 0.21 to 0.47) and the prospective change in pain or health between baseline and follow-up (pain $r_s = 0.27$, 95% CI: 0.13 to 0.40; global health $r_s = 0.50$, 95% CI: 0.38 to 0.60). Patients with low baseline pain or global health tended to exaggerate its severity afterwards,

while patients with high baseline scores tended to underestimate baseline states. Moreover, prospectively improved patients tended to underestimate baseline severity, whereas patients whose condition deteriorated tended to overestimate baseline severity. Differences between both methods of baseline assessment were not significantly correlated with patients' baseline characteristics and present level of pain or health status at the moment of recall.

Correlation between actual and retrospective baseline assessments

The retrospective assessments of baseline pain and global health correlated adequately with the actual baseline assessments (pain r_s = 0.73, 95% CI: 0.65 to 0.79; global health r_s = 0.66, 95% CI: 0.57 to 0.74).

Agreement between actual and retrospective baseline assessments

Bland–Altman analysis of the difference between actual and retrospective baseline assessments against the mean of both methods (Figure 1 and Figure 2) confirmed that the systematic bias between actual and retrospective assessments was small. Both pain and global health were, however, characterized by high intra-individual variation, with the 95% limits of agreement ranging from –37.3 to 32.3 mm for pain and 49.7 to 37.8 mm for global health.

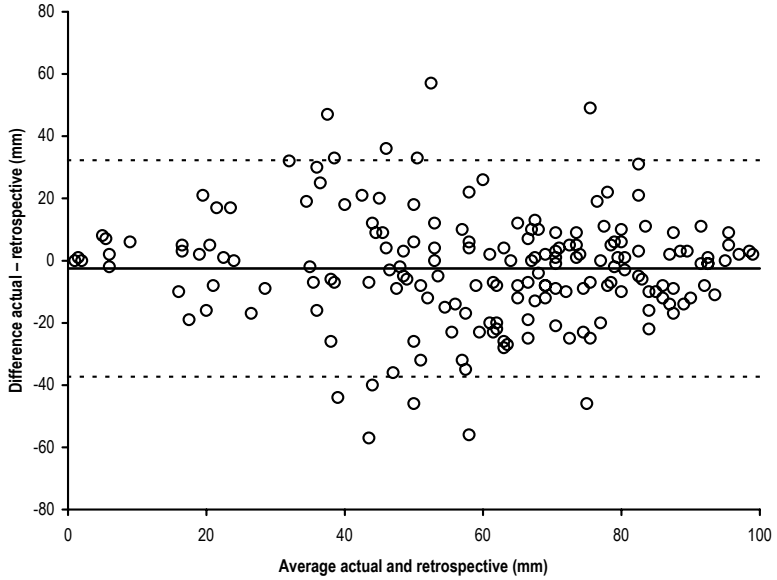

Figure 1. Individual agreement between actual and 2-week retrospective assessments of baseline pain on the VAS. The horizontal solid line represents the mean difference (bias) between both change scores (–2.5 mm). The dashed lines represent the 95% limits of agreement (mean difference ± 1.96 SD of the difference), ranging from –37.3 to 32.3 mm.

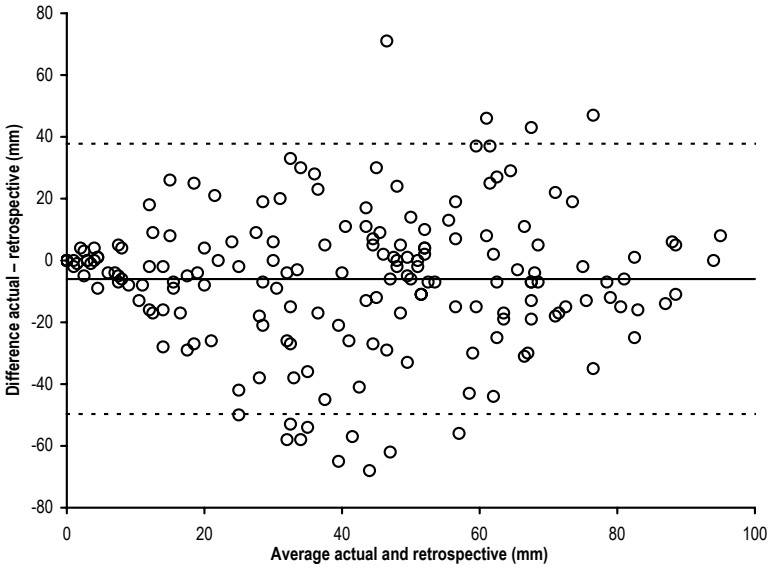

Figure 2. Individual agreement between actual and 2-week retrospective assessments of baseline global health on the VAS. The horizontal solid line represents the mean difference (bias) between both change scores (–6.0 mm). The dashed lines represent the 95% limits of agreement (mean difference ± 1.96 SD of the difference), ranging from –49.7 to 37.8 mm.

Discussion

Prospective measurement of changes in patient-reported outcomes such as pain and global health is the gold standard for clinical research. In this study we investigated whether patients' baseline pain and global health states can be reliably assessed retrospectively. The results of the study indicate that although retrospective assessments of baseline pain and global health are fairly accurate at the group level and adequately correlated with actual baseline scores, there is poor agreement within individual patients.

The results showed that, as a group, patients tended to overestimate both the severity of baseline pain and global health retrospectively. This tendency of patients to overestimate the severity of their pre-treatment situation has been reported in previous studies.[15–17,19–22] Two possible theoretical explanations have been proposed for this systematic bias in recall. The first explanation is motivational bias (e.g., cognitive dissonance or social desirability), where patients who have undergone a treatment will be motivated to exaggerate the benefits of that treatment.[4] The second explanation is response shift bias, which refers to a change in the meaning of one's self-evaluation of their health status as a result of a change in their internal standards, values or conceptualization of the measured construct.[26] However, since the patients in this study were

asked to recall their baseline status, as opposed to give a renewed judgment with the insights they have now (a so-called then-test), true response shift could not be assessed.

In accordance with other pain studies,[13,22] the differences between actual and retrospective assessments in this study were related to the actual baseline level of pain or global health and its prospective change. However, the accuracy of recall was not influenced by the present level of pain or global health at the moment of recall, as previously suggested.[11,22,27,28]

Although the group differences between actual and retrospective assessments in this study were statistically significant, their small magnitude suggests that they are not likely to be of clinical significance. Several studies have demonstrated that patient-perceived pain and global health on the VAS have poor test-retest reliability and high random measurement error compared to multi-item measures.[29,30] The observed differences on the VAS can therefore not be reliably distinguished from random error.

The small average differences between retrospective and actual assessments and the adequate correlation between them, would suggest that retrospective assessments after a 2-week period can capture quite reliable data on baseline pain and global health at the group level. However, within individual patients, the difference between actual and retrospective assessments proved to be highly variable and subject to error. Although there are no established rules for clinically acceptable differences between the two methods, using retrospective assessments should at the least not lead to different conclusions about the efficacy of treatment. In this study, however, the 95% limits of agreement of the Bland–Altman plots far exceeded proposed values of approximately 15–20 mm for minimal clinically important improvements in pain and global health.[31,32] Using patients' retrospective instead of actual baseline assessments to measure change over treatment, could thus result in a high number of patients being incorrectly classified as having significantly improved or deteriorated.

Although several previous studies have examined patients' recall of pain, this study is one of the first to examine patients' ability to recall previous global health states. The findings support the assumption that patients' memory of global health status is even more problematic than their recall of pain. Recall bias was larger in global health assessments, and patients' actual and retrospective assessments of global health were less strongly correlated. Moreover, Bland–Altman analyses indicated that actual and retrospective assessments of global health were more susceptible to intra-individual variability. This suggests that patients have more trouble remembering previous global health states than previous pain states.

Some reservations should be made regarding the generalisability of the present findings. Firstly, the study population included patients with heterogeneous diagnoses. Since pain and global health are known to vary across different rheumatic diseases, the findings may not be applicable to specific rheumatic conditions. Moreover, since most

patients experienced a major improvement in pain at the 2-week follow-up, the findings may not apply to stable pain recall. A further limitation of this study is that it is not clear whether patients at follow-up truly recalled their baseline pain and global health status, or tried to recall the physical position of their mark on the baseline VAS. Moreover, patients completed the baseline questionnaire in the clinic and in the presence of an investigator, whereas the follow-up questionnaire was mailed the patient's home. The contexts in which the data were collected may have affected patients' reporting.[33] Finally, the study design did not incorporate the influence of personality characteristics or psychosocial factors, which can contribute to the variability in the memory of previous pain or health states.[15,34–38]

In conclusion, retrospective assessments can provide fairly reliable data on aggregate baseline pain and global health and can be used for descriptive and exploratory purposes. However, at the individual level there is poor agreement between actual and retrospective assessments of baseline health states. The unacceptably high variability in the magnitude and direction of the differences confirms that even over relatively short time intervals, retrospective assessments should not be used as substitutes for individual baseline status or to measure individual changes over treatment in clinical trials.

Acknowledgements

The authors thank all the respondents for their participation in the study and the rheumatologists of Medisch Spectrum Twente for their help in the inclusion of the patients.

References

1. Felson DT, Anderson JJ, Boers M, et al. The American College of Rheumatology preliminary core set of disease activity measures for rheumatoid arthritis clinical trials. Arthritis Rheum 1993;36:729–40.
2. Emberton M, Challands A, Styles RA, Wightman JA, Black N. Recollected versus contemporary patient reports of pre-operative symptoms in men undergoing transurethral prostatic resection for benign disease. J Clin Epidemiol 1995;48:749–56.
3. Kreulen GJ, Stommel M, Gutek BA, Burns LR, Braden CJ. Utility of retrospective pretest ratings of patient satisfaction with health status. Res Nurs Health 2002;25:233–41.
4. Aseltine RH, Jr., Carlson KJ, Fowler FJ, Jr., Barry MJ. Comparing prospective and retrospective measures of treatment outcomes. Med Care 1995;33:AS67–76.
5. Fischer D, Stewart AL, Bloch DA, Lorig K, Laurent D, Holman H. Capturing the patient's view of change as a clinical outcome measure. JAMA 1999;282:1157–62.
6. Herrmann D. Reporting current, past, and changed health status. What we know about distortion. Med Care 1995;33:AS89–94.
7. Sokka T. Assessment of pain in rheumatic diseases. Clin Exp Rheumatol 2005;23:S77–84.

8. Babul N, Darke AC, Johnson DH, Charron-Vincent K. Using memory for pain in analgesic research. Ann Pharmacother 1993;27:9–12.
9. Beese A, Morley S. Memory for acute pain experience is specifically inaccurate but generally reliable. Pain 1993;53:183–9.
10. Brauer C, Thomsen JF, Loft IP, Mikkelsen S. Can we rely on retrospective pain assessments? Am J Epidemiol 2003;157:552–7.
11. Salovey P, Smith AF, Turk DC, Jobe JB, Willis GB. The accuracy of memory for pain: not so bad most of the time. Am Pain Soc J 1993;2:184–91.
12. Singer AJ, Kowalska A, Thode HC, Jr. Ability of patients to accurately recall the severity of acute painful events. Acad Emerg Med 2001;8:292–5.
13. Bryant RA. Memory for pain and affect in chronic pain patients. Pain 1993;54:347–51.
14. Dawson EG, Kanim LE, Sra P, et al. Low back pain recollection versus concurrent accounts: outcomes analysis. Spine 2002;27:984–93.
15. Jamison RN, Sbrocco T, Parris WC. The influence of physical and psychosocial factors on accuracy of memory for pain in chronic pain patients. Pain 1989;37:289–94.
16. Lingard EA, Wright EA, Sledge CB. Pitfalls of using patient recall to derive preoperative status in outcome studies of total knee arthroplasty. J Bone Joint Surg Am 2001;83:1149–56.
17. Linton SJ, Melin L. The accuracy of remembering chronic pain. Pain 1982;13:281–5.
18. Liu WH, Aitkenhead AR. Comparison of contemporaneous and retrospective assessment of postoperative pain using the visual analogue scale. Br J Anaesth 1991;67:768–71.
19. Pellise F, Vidal X, Hernandez A, Cedraschi C, Bago J, Villanueva C. Reliability of retrospective clinical data to evaluate the effectiveness of lumbar fusion in chronic low back pain. Spine 2005;30:365–8.
20. Bernhard J, Lowy A, Maibach R, Hurny C. Response shift in the perception of health for utility evaluation: an explorative investigation. Eur J Cancer 2001;37:1729–35.
21. Mancuso CA, Charlson ME. Does recollection error threaten the validity of cross-sectional studies of effectiveness? Med Care 1995;33:AS77–88.
22. Feine JS, Lavigne GJ, Dao TT, Morin C, Lund JP. Memories of chronic pain and perceptions of relief. Pain 1998;77:137–41.
23. Jensen MP, Chen C, Brugger AM. Postsurgical pain outcome assessment. Pain 2002;99:101–9.
24. McGorry RW, Webster BS, Snook SH, Hsiang SM. Accuracy of pain recall in chronic and recurrent low back pain. J Occup Rehabil 1999;9:169–78.
25. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;i:307–10.
26. Sprangers MA, Schwartz CE. Integrating response shift into health-related quality of life research: a theoretical model. Soc Sci Med 1999;48:1507–15.
27. Eich E, Reeves JL, Jaeger B, Graff-Radford SB. Memory for pain: relation between past and present pain intensity. Pain 1985;23:375–80.
28. Smith WB, Safer MA. Effects of present pain level on recall of chronic pain and medication use. Pain 1993;55:355–61.
29. Lassere MN, van der Heijde D, Johnson KR, Boers M, Edmonds J. Reliability of measures of disease activity and disease damage in rheumatoid arthritis: implications for smallest detect-

able difference, minimal clinically important difference, and analysis of treatment effects in randomized controlled trials. J Rheumatol 2001;28:892–903.
30. Russell AS, Conner-Spady B, Mintz A, Maksymowych WP. The responsiveness of generic health status measures as assessed in patients with rheumatoid arthritis receiving infliximab. J Rheumatol 2003;30:941–7.
31. Farrar JT, Young JP, LaMoreaux L, Werth JL, Poole RM. Clinical importance of changes in chronic pain intensity measured on an 11-point numerical pain rating scale. Pain 2001;94:149–58.
32. Tubach F, Ravaud P, Baron G, et al. Evaluation of clinically relevant changes in patient reported outcomes in knee and hip osteoarthritis: the minimal clinically important improvement. Ann Rheum Dis 2005;64:29–33.
33. Campbell R, Quilty B, Dieppe P. Discrepancies between patients' assessments of outcome: qualitative study nested within a randomised controlled trial. BMJ 2003;326:252–3.
34. Gedney JJ, Logan H. Memory for stress-associated acute pain. J Pain 2004;5:83–91.
35. Lefebvre JC, Keefe FJ. Memory for pain: the relationship of pain catastrophizing to the recall of daily rheumatoid arthritis pain. Clin J Pain 2002;18:56–63.
36. Linton SJ. Memory for chronic pain intensity: correlates of accuracy. Percept Mot Skills 1991;72:1091–5.
37. Porzelius J. Memory for pain after nerve-block injections. Clin J Pain 1995;11:112–20.
38. Tasmuth T, Estlanderb AM, Kalso E. Effect of present pain and mood on the memory of past postoperative pain in women treated surgically for breast cancer. Pain 1996;68:343–7.

5 The validity and reliability of the graphic rating scale and verbal rating scale for measuring pain across cultures: a study in Egyptian and Dutch women with rheumatoid arthritis

P.M. ten Klooster
A.P.J. Vlaar
E. Taal
R.E. Gheith
J.J. Rasker
A.K. El-Garf
M.A.F.J. van de Laar

The Clinical Journal of Pain 2006; 22: 827–830.

Abstract

Objective. To compare the validity and reliability of a graphic rating scale (GRS) and a verbal rating scale (VRS) for measuring pain intensity in young female Egyptian and Dutch patients with rheumatoid arthritis (RA).

Methods. Data were obtained in a cross-cultural study of 42 Egyptian and 30 Dutch female outpatients with stable RA. Construct validity was assessed by correlating the scales with other core measures of disease activity in RA. Test-retest reliability was assessed over a 1-week interval.

Results. The GRS and the VRS were strongly intercorrelated in the total study cohort and in the Egyptian and Dutch subgroups. In the individual subgroups, only the GRS demonstrated the expected pattern of correlations with other disease activity measures. Test-retest reliability of the GRS was adequate in both Egyptian and Dutch patients (intraclass correlation coefficient 0.78 vs. 0.83, respectively), whereas reliability of the VRS was unsatisfactory in the Egyptian subgroup (weighted κ 0.60 vs. 0.82 in the Netherlands).

Discussion. The study confirmed that the GRS and VRS were reliable and valid in the total study cohort. Within the individual countries, the GRS seemed to perform better than the VRS.

Introduction

Single-item continuous rating scales such as the visual analog scale (VAS) or the similar graphic rating scale (GRS) and categorical scales like the verbal rating scale (VRS) are among the most commonly used measures of pain intensity.[1,2] Both types of pain scales have shown good psychometric properties, although the VAS and the GRS generally tend to be more sensitive to change.[3–6]

The evidence supporting the use of these scales, however, is largely based on research conducted in Western settings. To our knowledge, no studies have directly compared the psychometric qualities of pain scales between patients from Arabic and Western cultures. The aim of this study was to examine the comparative validity and reliability of the GRS and the VRS in Egyptian and Dutch young female RA patients.

Materials and methods

Study design and patients

The Egyptian participants in this study were consecutively recruited at the outpatient rheumatology clinic of the University Hospital of Cairo, Egypt. Forty-two female RA patients aged >18 years and fulfilling the American College of Rheumatology revised criteria for RA7 were included. During the first visit an Egyptian rheumatologist (R.E.G.) administered a set of questionnaires including the GRS and the VRS. Additionally, a clinical examination was performed. The questionnaires were readministered 1 week after the first visit.

For comparison, a Dutch sample of 30 female RA patients matched for age and disease duration was selected from the patient registry of the rheumatology clinic of Medisch Spectrum Twente in Enschede, the Netherlands. The Dutch patients followed the same assessment procedure as the Egyptian patients. Clinical examinations in both samples were performed by the same investigator (A.P.J.V.).

Measures

The GRS is very similar to the more often used VAS, the primary difference being that the GRS has specific markers along the line. In this study, the GRS consisted of a 10-cm horizontal line divided by vertical marks into 10 equal segments, anchored at each end with 0 (no pain) and 100 (severe pain). Patients were asked to mark the line at a point that best represented the severity of their pain. The VRS consisted of 5 words describing different levels of pain: 1 = none, 2 = very mild, 3 = mild, 4 = moderate, and 5 = severe. Patients were asked to select the word that best described their usual pain. To assess physical functioning, patients additionally completed a language-specific Health

Assessment Questionnaire Disability Index (HAQ-DI, range 0 to 3 with higher scores indicating more disability).[8,9]

The baseline clinical examination included a tender joint count (TJC, 68 joints evaluated), swollen joint count (66 joints evaluated), examination of recent radiographs of the hands and wrists using the Sharp/van der Heijde scoring method (SHS, range 0 to 280),[10] and laboratory evaluation of erythrocyte sedimentation rate (ESR) and rheumatoid factor.

Statistical analysis

Construct validity was assessed by intercorrelating the GRS and the VRS and correlating the scales with other variables from the American College of Rheumatology core set of disease activity measures.[11] Spearman correlation coefficients were used for all correlations within the individual countries, while Spearman partial correlation coefficients, controlling for country effects, were used in the total study cohort. For convergent and divergent validity, it was hypothesized that a valid measure of pain would be strongly (>0.60) associated with physical functioning, moderately (0.30 to 0.60) with the number of tender and swollen joints, and weakly or not at all (<0.30) with ESR and radiographic damage.[12] One-week test-retest reliability of the scales was assessed using the intraclass correlation coefficient (ICC) for the GRS and the quadratic weighted kappa (κ_w) for the VRS.[13] Reliability was considered adequate for group comparisons when >0.70.[14]

Results

Table 1 shows some baseline characteristics of the Egyptian and Dutch patients. The subgroups were comparable regarding age, disease duration, radiographic damage, and presence of rheumatoid factor. There were significant differences in pain scores, physical functioning, joint counts, and ESR between Egyptian and Dutch patients. The educational level of the Egyptian subgroup was significantly lower with 86% of the patients having completed no formal education at all or primary school only, compared with 3% of the Dutch patients ($P < 0.001$).

The GRS and the VRS were strongly intercorrelated in the total study cohort (partial $r = 0.70$, $P < 0.001$) and in the Egyptian ($r = 0.68$, $P < 0.001$) and Dutch ($r = 0.78$, $P < 0.001$) subgroups. In the total cohort, the GRS and the VRS showed similar patterns of correlations with other clinical measures (Table 2). As expected, pain scores were strongly correlated with the HAQ-DI, moderately with the TJC, and weakly or not at all with ESR and SHS. Contrary to expectations, pain was only weakly associated with the swollen joint count. In the Egyptian and Dutch subgroups, the GRS demonstrated the

expected correlations with other core measures. VRS scores were less clearly associated with the TJC, but more strongly with the ESR and SHS than the GRS scores.

Table 1. Baseline demographic and clinical characteristics

	Egyptian patients (n = 42)	Dutch patients (n = 30)	*P*
Age (y)	31.9 (6.0)	34.5 (5.6)	0.064
Disease duration (y)	4.0 (3.0–7.0)	5.0 (3.0–7.5)	0.113
GRS, 0–100	59.8 (21.8)	26.0 (24.2)	<0.001
VRS, 1–5	4.0 (3.0–5.0)	3.0 (2.0–4.0)	<0.001
HAQ-DI, 0–3	1.4 (0.8)	0.7 (0.7)	<0.001
TJC, 0–68	30.5 (14.5–55.3)	7.5 (1.0–17.5)	<0.001
SJC, 0–66	1.0 (0.0–1.0)	4.0 (2.0–8.3)	<0.001
ESR (mm/h)	49.4 (23.5)	12.9 (9.0)	<0.001
RF positive (%)	67	70	0.485
SHS, 0–280	48.7 (32.1)	48.9 (42.6)	0.975

Values are expressed as mean (SD) or median (interquartile range) unless otherwise noted.
RF positive indicates rheumatoid factor positive; SJC, swollen join count.

Table 2. Spearman correlation coefficients between the GRS and VRS for pain and other measures of disease activity

	Total study cohort (N = 72)		Egyptian patients (n = 42)		Dutch patients (n = 30)	
	GRS†	VRS†	GRS	VRS	GRS	VRS
HAQ-DI	0.65***	0.59***	0.61***	0.62***	0.75***	0.57**
TJC	0.41***	0.24*	0.44**	0.28	0.51**	0.28
SJC	0.28*	0.25*	0.38*	0.31*	0.30	0.34
ESR	0.15	0.28*	0.10	0.25	0.05	0.32
SHS	–0.07	0.09	0.21	0.52***	–0.22	–0.31

* $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$.
† Spearman partial correlation coefficients, controlling for country.

Test-retest reliability of the GRS exceeded the commonly accepted standard of 0.70 in the total study cohort (ICC = 0.85) and in both the Egyptian (ICC = 0.78) and Dutch (ICC = 0.83) patients. The VRS demonstrated adequate reliability in the total cohort (κ_w = 0.76) and the Dutch subgroup (κ_w = 0.82), but performed unsatisfactory in Egyptian patients (κ_w = 0.60).

Discussion

The assessment of patient-perceived pain is central to both clinical trials and clinical practice in RA. Single-item ratings of pain intensity are usually recommended for measuring pain in patients with RA.[11] Although the psychometric qualities of the GRS and the VRS have been established, no studies have directly compared the validity and reliability of both scales in patients from different cultures. In this study, we examined the comparative validity and reliability of the GRS and the VRS in Egyptian and Dutch young female RA patients. The results of the study showed that the GRS and VRS were both reliable and valid in the total study cohort, but that the GRS performed better within the Egyptian and Dutch subgroups.

In the total study cohort and in both subgroups, the GRS and the VRS were strongly intercorrelated. This is consistent with other studies with chronic pain patients where correlations ranging from 0.60 to 0.80 between the unmarked VAS and the VRS have been reported.[1–3,6,15–17] Although the high correlation between the GRS and VRS does provide support for the concurrent validity of the scales, correlation coefficients generally tend to overestimate actual agreement.[18] Moreover, the GRS and the VRS demonstrated a somewhat different pattern of correlations with other measures of disease activity both within and between the two countries. This suggests that the GRS and the VRS should not be used interchangeably to measure pain in RA patients.

An interesting finding of this study was that the test-retest reliability of the GRS was nearly equivalent in the Egyptian and Dutch patients. Several studies have emphasized that an important limitation of the unmarked VAS is that some populations have difficulty understanding the abstract nature of the scale.[15,19] Besides the known comprehension problems in elderly or cognitively impaired persons, Clark et al[6] recently suggested that patients with low education may also experience difficulties completing the VAS. Moreover, in a study in literate and illiterate RA patients, Ferraz et al[20] found that the test-retest reliability of the VAS was significantly lower in illiterate RA patients and that the VRS performed better in this group. A possible explanation for the relatively high reliability of the GRS in the predominantly illiterate Egyptian patients in this study is that the marks along the GRS assisted the patients in choosing the appropriate position on the line. Because both illiterate and literate people can usually count to 10, a GRS may be more suitable for patients with less formal education.

Some caution is needed in interpreting the results of this study. Firstly, it should be noted that the cohort of patients studied is not representative of the general Western population of RA patients. Because the Egyptian subgroup was selected first, the Dutch patients were matched to this subgroup. However, the epidemiology of RA in Egypt and the clinical characteristics of the Egyptian hospital-based population differ considerably from those in the Netherlands. In general, the average age, age at onset, and

disease duration of Egyptian RA patients is much lower than those of Western patients and the female to male ratio higher.[21]

Moreover, disease activity was significantly higher in the Egyptian patients, which may have been caused by different treatment strategies and by ethnic or cultural differences in the perception and reporting of pain and disability. Consequently, there was a considerable ceiling effect on the VRS in the Egyptian subgroup where 15 patients (35%) reported their pain as severe, as opposed to only 2 Dutch patients (7%). Responses on the GRS, however, were normally distributed. This may have contributed to decreased performance of the VRS and would suggest that the 5-point scale used in this study is not sensitive enough in patients with higher levels of pain.

Finally, although the cultural difference between Egypt and the Netherlands was one of the primary reasons for this study, the educational differences may have also limited the comparability of the results. Because most Egyptian patients were illiterate, all questionnaires were administered in a face-to-face interview with an investigator, whereas the Dutch questionnaires were mostly self-completed. The presence of an investigator may have affected the patients' reporting of pain or physical functioning.

In summary, this study confirmed that the GRS and the VRS are valid and reliable measures of pain. The psychometric properties of the specific scales may, however, differ between countries or cultures.

References

1. Ohnhaus EE, Adler R. Methodological problems in the measurement of pain: a comparison between the verbal rating scale and the visual analogue scale. Pain 1975;1:379–84.
2. Linton SJ, Gotestam KG. A clinical comparison of two pain scales: correlation, remembering chronic pain, and a measure of compliance. Pain 1983;17:57–65.
3. Downie WW, Leatham PA, Rhind VM, et al. Studies with pain rating scales. Ann Rheum Dis 1978;37:378–81.
4. Langley GB, Sheppeard H. Problems associated with pain measurement in arthritis: comparison of the visual analogue and verbal rating scales. Clin Exp Rheumatol 1984;2:231–4.
5. Scott J, Huskisson EC. Graphic representation of pain. Pain 1976;2:175–84.
6. Clark P, Lavielle P, Martinez H. Learning from pain scales: patient perspective. J Rheumatol 2003;30:1584–8.
7. Arnett FC, Edworthy SM, Bloch DA, et al. The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. Arthritis Rheum 1988;31:315–24.
8. El Meidany YM, El Gaafary MM, Ahmed I. Cross-cultural adaptation and validation of an Arabic Health Assessment Questionnaire for use in rheumatoid arthritis patients. Joint Bone Spine 2003;70:195–202.
9. Siegert CE, Vleming LJ, Vandenbroucke JP, et al. Measurement of disability in Dutch rheumatoid arthritis patients. Clin Rheumatol 1984;3:305–9.

10. van der Heijde D. How to read radiographs according to the Sharp/van der Heijde method. J Rheumatol 2000;27:261–3.
11. Felson DT, Anderson JJ, Boers M, et al. The American College of Rheumatology preliminary core set of disease activity measures for rheumatoid arthritis clinical trials. Arthritis Rheum 1993;36:729–40.
12. Sokka T. Assessment of pain in rheumatic diseases. Clin Exp Rheumatol 2005;23:S77–84.
13. Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educ Psychol Meas 1973;33:613–19.
14. Lohr KN. Assessing health status and quality-of-life instruments: attributes and review criteria. Quality of Life Research 2002;11:193.
15. Kremer E, Atkinson JH, Ignelzi RJ. Measurement of pain: patient preference does not confound pain measurement. Pain 1981;10:241–8.
16. Breivik EK, Bjornsson GA, Skovlund E. A comparison of pain rating scales by sampling from clinical trial data. Clin J Pain 2000;16:22–8.
17. Bolognese JA, Schnitzer TJ, Ehrich EW. Response relationship of VAS and Likert scales in osteoarthritis efficacy measurement. Osteoarthritis Cartilage 2003;11:499–507.
18. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1:307–10.
19. Jensen MP, Karoly P, Braver S. The measurement of clinical pain intensity: a comparison of six methods. Pain 1986;27:117–26.
20. Ferraz MB, Quaresma MR, Aquino LR, et al. Reliability of pain scales in the assessment of literate and illiterate patients with rheumatoid arthritis. J Rheumatol 1990;17:1022–4.
21. Abdel-Nasser AM, Rasker JJ, Valkenburg HA. Epidemiological and clinical aspects relating to the variability of rheumatoid arthritis. Semin Arthritis Rheum 1997;27:123–40.

6 A cross-cultural study of pain intensity in Egyptian and Dutch women with rheumatoid arthritis

A.P.J. Vlaar
P.M. ten Klooster
E. Taal
R.E. Gheith
A.K. El-Garf
J.J. Rasker
M.A.F.J. van de Laar

The Journal of Pain 2007; 8: 730–736.

Abstract

It has been suggested that patients from Mediterranean cultures tend to report more intense pain than their Northern or Western European counterparts in comparable medical conditions. However, empirical data to support this hypothesis are limited. The goals of the present study were to examine differences in pain intensity reports between Dutch and Egyptian women with rheumatoid arthritis (RA) and to examine the influence of possible confounding variables using multivariate analyses. We performed a cross-sectional study in 30 Dutch and 42 Egyptian women with comparable RA, matched for age and disease duration. Pain intensity was measured on a 100-mm graphic rating scale. Additionally, we assessed physical function, radiographic joint damage, progression of RA, disease activity, number of swollen and tender joints, medication, rheumatoid factor, and socioeconomic variables. The progression of RA and radiographic damage were not significantly different between Egyptian and Dutch patients. However, the Egyptian population reported significantly worse pain and physical function and demonstrated higher disease activity. Multiple linear regression analysis showed that the country of residence and the number of tender and swollen joints were significant independent determinants of pain reports. The results provide some support for the idea that there are ethnocultural differences in pain reports between Egyptian and Dutch women with RA, although the mechanisms underlying these differences remain unclear.

Perspective. This article shows that after controlling for differences in demographic, socioeconomic, and clinical variables, Egyptian women with RA reported more pain than Dutch women with RA. Clinicians and investigators should recognize that cultural or ethnic factors may play an important role in patients' pain reports.

Introduction

Since pain is a highly subjective experience, the assessment of pain depends primarily upon patients' self-reports. Over the years, there has been increasing interest in the relationship between cultural background or ethnicity and the experience of laboratory and clinical pain.[1] Although many studies have focused on more or less acculturated ethnic groups within the same nation, such as African Americans and Caucasians in the United States, a growing body of evidence now suggests that the perception and reporting of pain intensity can vary across ethnic or cultural groups.[2–15] However, few studies have examined possible confounders that may explain the observed ethnocultural differences in pain, such as socioeconomic factors, and the definitions used for ethnic categories are often problematic.[16–19]

One specific cross-cultural issue that has not received much empirical investigation is the widespread belief that patients from Mediterranean cultures tend to express more intense pain than Northern or Western European patients in comparable medical conditions. This idea is hardly new. Already in 1944, Chapman and Jones[20] reported that healthy subjects from Mediterranean races were more sensitive to pain than subjects from Northern European stock. In another early and influential study, Zborowski[21] suggested that Italian-Americans and Jewish-Americans were more openly expressive about their pain and tended to complain more than Anglo-Americans. Finally, a more recent study of Israeli women with fibromyalgia demonstrated that patients of Mediterranean origin reported higher levels of pain and more severe symptoms than patients of European-American origin.[22] However, these differences disappeared when the results were adjusted for age and education.

Despite the pervasiveness of the idea that Mediterranean patients tend to report more pain, empirical evidence for existing differences in pain intensity reports between persons from Mediterranean and Northern or Western European cultures between is very limited.[18] In a previous study we compared the severity and impact of rheumatoid arthritis in Egyptian and Dutch patients.[23] The Egyptian patients reported more pain than the Dutch (5.9 vs. 3.0 on a 0–10 numerical rating scale, $P < 0.001$), the disease was more active in the Egyptian patients, and they were more disabled despite a comparable disease otherwise. The goals of the present study were to further examine differences in pain intensity scores in a new cohort of Dutch and Egyptian women with RA and to investigate the influence of possible confounding variables using multivariate analyses.

Materials and methods

Patients and study design

Participants in this study were Egyptian and Dutch female RA patients fulfilling the American College of Rheumatology revised criteria for RA.[24] An a priori power analysis indicated that 26 patients in each group would be required to achieve 80% power (α = 0.05, expected standard deviation = 25 mm) for a two-sided test to detect a clinically meaningful difference of 20 mm in pain on a 100-mm graphic rating scale.

The Egyptian patients were consecutively recruited in October 2004 at the outpatient rheumatology clinic of the University Hospital of Cairo, Egypt. The Dutch sample was recruited between December 2004 and April 2005 at the outpatient rheumatology clinic of Medisch Spectrum Twente in Enschede, The Netherlands. To reduce the number of possible confounding variables, the Dutch patients were group-matched for age, gender, and disease duration. Eligible Dutch participants were identified by regular review of the clinical records of RA patients who were scheduled for a visit to the clinic.

Both groups followed the same assessment protocol. During the patients' visit to the clinic, the study was explained in the patients' own language by the local physician (R.E.G. in Egypt and A.P.J.V. in The Netherlands). Patients who agreed to participate were administered a questionnaire that included items on demographic, socioeconomic, and clinical variables and standard self-report measures of pain and physical function. Since most Egyptian patients were illiterate, the questions were read aloud by the physician in a face-to-face interview, whereas the Dutch questionnaires were mostly self-completed.

Additionally, a standardized physical examination, radiographic examination, and laboratory tests were performed. The physical and radiographic examinations in both groups were performed by the same trained physician (A.P.J.V.). Additional data on currently prescribed disease-modifying antirheumatic drugs (DMARDs), such as methotrexate (MTX) and sulfasalazine, and simple analgesics, including paracetamol, non-steroidal anti-inflammatory drugs and opioids, was obtained from the patients' medical records.

In accordance with local regulations, the institutional ethics committees of both hospitals decided that the study did not require formal approval. Verbal informed consent, however, was obtained from all patients before enrollment.

Measures

Pain intensity was measured using a horizontal 100-mm graphic rating scale (GRS) ranging from "no pain" (0) to "severe pain" (100). The GRS is similar to the more commonly known visual analog scale (VAS), the primary difference being that the GRS

adds specific markers along the line.[25,26] The GRS we used in this study had 11 vertical marks dividing the scale into 10 equal segments. Patients were asked to place a mark at any point on the line that best represented the severity of their pain over the past week. The scale was scored by measuring the distance in mm from the "no pain" end to the patient's mark. The GRS was selected over the VAS since several studies have shown that certain patients experience difficulties in understanding and completing the VAS.[27,28] Especially in populations with a high level of illiteracy, as is the case in Egypt, the VAS may not be the best choice for assessing pain.[29,30] Previous analyses demonstrated that the GRS had adequate test-retest reliability and an expected pattern of correlations with established disease activity measures in both the Egyptian and Dutch patients.[31]

Validated translations of the Health Assessment Questionnaire Disability Index (HAQ-DI)[32,33] were used to assess physical function. The HAQ-DI consists of 20 items and measures physical disabilities over the past week in 8 categories of daily living. Scores on the standard disability index, which corrects for the use of devices or help from others, range from 0 to 3, with higher scores indicating more disability. The culturally adapted Arabic HAQ-DI was recently developed and tested in a sample of 184 rheumatoid arthritis patients from different Arabic countries, including 78 Egyptian patients.[34] The questionnaire showed high test-retest reliability and internal consistency and its validity was supported by correlations with other disease activity parameters. In The Netherlands, slightly different translations of the HAQ-DI have been extensively used and validated in RA patients.[35–37] We used the most recent updated and validated version.[38]

The physical examination consisted of a tender joint count (68 joints assessed) and swollen joint count (66 joints assessed).[39] Radiographs were taken of the hands and wrists and scored using the Sharp/van der Heijde scoring method (SHS, range 0–280).[40] Moreover, the progression of RA was determined using the classification of Steinbrocker, which ranges from stage I (early) to IV (terminal).[41] Laboratory tests included measurement of the erythrocyte sedimentation rate (ESR) after one hour and rheumatoid factor (RF). Finally, a pooled disease activity index was calculated by the Disease Activity Score (DAS28-3), which combines a 28 tender and swollen joint count and the ESR into a single, continuous index ranging from 0 (no disease activity) to 10 (severe disease activity).[42]

Data analysis

Data analysis was performed using SPSS 12.0 for Windows (SPSS, Chicago, IL). The valid use of parametric statistics was verified by testing for normal distribution of the variables (Kolmogorov–Smirnov test, normal distribution assumed when $P > 0.05$). Differences between women from Egypt and The Netherlands were tested with inde-

pendent samples *t* tests (normally distributed variables), Mann–Whitney *U* tests (not normally distributed variables) and Pearson chi-square tests (categorical variables). The level of significance for these analyses was set at $P < 0.05$ (two-tailed).

Differences in pain intensity between subgroups of patients (e.g., married vs. not married, employed vs. unemployed, RF positive vs. RF negative, use of MTX vs. no use of MTX) were tested with independent samples *t* tests. Univariate correlations between pain and socioeconomic and clinical variables were analyzed separately for the total study sample and the Egyptian and Dutch subgroups using Spearman's correlation coefficients. Since differences in the perception of pain intensity between Egyptian and Dutch women might be confounded by differences in sociodemographic or clinical variables, we applied hierarchical multiple linear regression analyses. In the first block demographic, socioeconomic and clinical variables that were univariately related with pain intensity ($P < 0.10$) were entered. In the second block the country of residence was entered.

Results

The Egyptian sample consisted of 42 patients and the Dutch sample of 30 patients. There were no significant differences in age, disease duration, and marital status between the Egyptian and Dutch women (Table 1). Egyptian women had significantly more children than Dutch women ($P < 0.001$). The educational level of the Dutch women was higher ($P < 0.001$) than of the Egyptian women, of whom almost half had received no formal education at all. The percentage of employed women was significantly higher ($P < 0.001$) in The Netherlands compared to Egypt.

Neither disease characteristics nor progression (RF, Steinbrocker classification, SHS) were different between women from the 2 countries (Table 2). Disease activity, including ESR and tender joint counts, but with the exception of swollen joint counts, was significantly higher in Egyptian women than in Dutch women. Egyptian women were prescribed more DMARDs and more often MTX than Dutch women, but there was no significant difference in the total number of pain medications prescribed. Physical function and perception of pain intensity were significantly worse in Egyptian women compared to Dutch women. GRS pain intensity scores were 57% lower in Dutch women.

Pain intensity scores were significantly ($P < 0.001$) lower in employed women (26.8 ± 23.4 mm) than in women without employment (54.6 ± 25.9 mm) and significantly ($P < 0.01$) higher in women who were prescribed MTX (50.8 ± 27.1 mm) than in women who were not prescribed MTX (29.1 ± 26.1). No significant differences in pain intensity were found for marital status and RF. Table 3 shows the sociodemographic and clinical

Table 1. Demographic characteristics of the Egyptian (n = 42) and Dutch (n = 30) patients

	Egypt	Netherlands	*P*
Age, years	31.9 (6.0)	34.5 (5.6)	0.06*
Disease duration, years	4.0 (3.0–7.0)	5.0 (3.0–7.5)	0.11†
Marital status (n [%])			
Married	34 (81)	22 (73)	
Unmarried	6 (14)	7 (23)	
Divorced	2 (5)	1 (3)	0.60‡
No. of children	3 (1–5)	1 (0–2)	< 0.001†
Education (n [%])			
None	20 (48)	0 (0)	
Primary school	16 (38)	1 (3)	
Secondary school	4 (10)	21 (70)	
Higher profession school	1 (2)	7 (23)	
University	1 (2)	1 (3)	< 0.001†
Work status (n [%])			
Employed	3 (7)	20 (67)	
Housewife	39 (93)	6 (20)	
Unemployed	0 (0)	4 (13)	< 0.001

NOTE. Values are mean (standard deviation) or median (interquartile range) unless otherwise noted.
* Independent *t* test.
† Mann–Whitney *U* test.
‡ Pearson χ^2.

variables that were univariately correlated with pain intensity in the total sample or in the Egyptian and Dutch subgroups. In the total sample, pain intensity was significantly associated with number of children, educational level, ESR, TJC, DAS28-3, HAQ-DI, and number of DMARDs prescribed. Some correlations in the total sample appeared to be spuriously affected by combining disparate subgroups. Pain intensity was significantly correlated with number of children, educational level and ESR in the total sample, but this relation was not apparent within the separate subgroups. Number of DMARDS was significantly associated with pain intensity in the total sample and in the Egyptian subgroup, but not in the Dutch subgroup. Conversely, pain intensity and SJC were significantly associated in the Egyptian subgroup but not in the total sample and in the Dutch subgroup.

To avoid multicollinearity, employment, DAS28-3, HAQ-DI, and present MTX prescription were omitted from the regression analysis since these were highly correlated (*r*'s >0.60) with and well represented by respectively education, the separate disease activity parameters (TJC, SJC, ESR), and number of DMARDs. Table 4 shows the results of the linear regression analysis for pain intensity. Pain intensity was independently associated with the number of tender and swollen joints. Number of children and

Table 2. Clinical characteristics and currently prescribed medication of the Egyptian (n = 42) and Dutch (n = 30) Patients

	Egypt	Netherlands	*P*
Classification of progression (n [%])			
Early	8 (19)	8 (27)	
Moderate	20 (48)	16 (53)	
Severe	11 (26)	5 (17)	
Terminal	3 (7)	1 (3)	0.63*
RF positive (n [%])	28 (67)	21 (70)	0.97*
ESR, mm/h	49.4 (23.5)	12.9 (9.0)	<0.001†
SHS, 0–280	48.7 (32.1)	48.9 (42.6)	0.98†
TJC, 0–68	30.5 (14.5–55.3)	7.5 (1.0–17.5)	<0.001†
SJC, 0–66	1.0 (0.0–2.0)	4.0 (2.0–8.3)	<0.001†
DAS28-3, 0–10	5.3 (1.2)	3.4 (1.1)	<0.001†
No. of DMARDs	2 (1–2)	1 (1–2)	<0.01‡
MTX at present, yes (n [%])	38 (90)	14 (47)	<0.001*
No. of analgesics	1 (1–1)	1 (1–1)	0.40‡
HAQ-DI, 0–3	1.4 (0.8)	0.7 (0.7)	<0.001†
GRS for pain, 0–100	59.8 (21.8)	26.0 (24.2)	<0.001†

NOTE. Values are mean (standard deviation) or median (interquartile range) unless otherwise noted.
Abbreviations: RF, rheumatoid factor; ESR, erythrocyte sedimentation rate; SHS, Modified Sharp/van der Heijde method for scoring radiographs; TJC, tender joint count; SJC, swollen joint count; DAS, Disease Activity Score; DMARDs, disease-modifying antirheumatic drugs; MTX, methotrexate; HAQ-DI, Health Assessment Questionnaire Disability Index; GRS, graphic rating scale.
* Pearson χ^2.
† Independent *t* test.
‡ Mann–Whitney *U* test.

Table 3. Spearman correlations between pain intensity and sociodemographic and clinical variables

	Total sample (N = 72)		Egypt (n = 42)		Netherlands (n = 30)	
	r	*P*	*r*	*P*	*r*	*P*
No. of children	0.33	<0.01	0.28	0.07	–0.06	0.76
Educational level	–0.47	<0.001	–0.08	0.61	–0.19	0.31
ESR	0.51	<0.001	0.10	0.54	0.05	0.78
TJC	0.59	<0.001	0.44	<0.01	0.50	<0.01
SJC	–0.06	0.60	0.38	0.01	0.30	0.11
DAS28-3	0.71	<0.001	0.55	<0.001	0.48	<0.01
HAQ-DI	0.72	<0.001	0.61	<0.001	0.75	<0.001
No. of DMARDs	0.35	<0.01	0.34	0.03	0.04	0.82

Abbreviations: ESR, erythrocyte sedimentation rate; TJC, tender joint count; SJC, swollen joint count; DAS, Disease Activity Score; HAQ-DI, Health Assessment Questionnaire Disability Index; DMARDs, disease-modifying antirheumatic drugs.

educational level and ESR and number of DMARDs did not remain significant in the multivariate model. After controlling for sociodemographic and clinical variables, country of residence remained a significant independent predictor of pain intensity.

Table 4. Hierarchical multiple linear regression analysis predicting pain intensity (N = 72)

	B	*SEB*	*Beta standardized*	*P*	R^2	*F Change*	*P*
Step 1					0.51	10.99	<0.001
No. of children	1.36	1.48	0.09	0.36			
Educational level	–0.69	3.37	–0.03	0.84			
ESR	0.12	0.13	0.11	0.38			
TJC	0.38	0.14	0.30	<0.01			
SJC	1.02	0.48	0.21	0.04			
No. of DMARDs	6.14	3.16	0.18	0.06			
Step 2					0.55	5.44	0.02
Country*	–20.40	8.74	–0.36	0.02			

Abbreviations: SEB, standard error of B; ESR, erythrocyte sedimentation rate; TJC, tender joint count; SJC, swollen joint count; DMARDs, disease-modifying antirheumatic drugs.
* 1 = Egypt, 2 = Netherlands.

Discussion

The results of this study showed that, after controlling for differences in disease activity and socioeconomic and clinical variables, Egyptian women with RA reported more pain than Dutch women with RA.

This finding is consistent with other studies that have described ethnic or cultural differences in the pain experience in both experimental and clinical settings.[2–15] Several biological, social, psychological, and medical mechanisms that may differ between cultural or ethnic groups have been suggested to influence differences in reported pain between groups.[16,43] In this study, we focused on differences in demographic variables, disease characteristics, disease activity, medication, and socioeconomic status (marital status, education, and employment) between Egyptian and Dutch patients as possible confounders for differences in pain intensity scores. Despite controlling for these variables, patients' country of residence remained a significant determinant of pain.

There are several possible explanations for this finding. First, the measures used in this study may not cover all relevant demographic, socioeconomic, and clinical variables. For instance, socioeconomic status is a broad concept that, besides marital status, education, and employment, may also include other variables such as income and medical insurance. Moreover, ethnicity and socioeconomic status are inseparably related to access to care, which may be a distinct predictor of health outcomes.[44–46]

Also, other factors that were not included in this study could have contributed to the observed differences in reported pain between the Egyptian and Dutch patients. Several studies have shown that psychosocial variables that influence the perception of pain, such as depression, helplessness, and coping strategies, may differ between cultures.[3,7,8,13,43,47–49] Other unmeasured variables that could have affected the results include genetic or biological differences and differences in comorbidities between the 2 ethnic populations. Inclusion of such variables in future studies could help to identify the underlying mechanisms that explain ethnic or cultural differences in pain experience.

Another interesting issue concerns the possible influence of semantic differences in pain descriptions across cultures.[50,51] Although Wolff[52] has suggested the use of nonverbal, graphic methods of pain measurement in cross-cultural comparisons, even the anchor points of such a scale are grounded in language.[53] However, since other indicators of pain or disability in this study were also higher in the Egyptian sample, this is not likely to have significantly affected the results.

Finally, only the intensity of pain was measured in this study. In a study on cancer pain, Greenwald found significant differences in affective pain ratings between ethnicities, while pain intensity ratings were not different.[54] This finding was confirmed in a study on ethnic differences in the experience of chronic pain[9] and in a study on the perception of experimental pain.[55] This would suggest that ethnocultural factors are more strongly related with the emotional experience of pain than with the sensory experience. To examine this distinction in RA pain, future studies examining ethnocultural differences in the experience of pain could use multiple pain measures that cover both sensory and affective dimensions of pain.

An important finding of this study was that although disease severity as measured by radiological and serological parameters was similar in the Egyptian and Dutch patients, pain, disease activity, and physical function were significantly worse in the Egyptian cohort. The high disease activity and pain scores of the Egyptian patients could indicate that pain and disease activity are generally under-treated in this population. In this light, it was somewhat surprising that the patients in Egypt were prescribed more DMARDs than the Dutch patients. However, it should be noted that we could not assess whether prescribed medications were actually filled and taken by the patients. Especially in the Egyptian sample the number of prescribed medications may not correspond to number of medications taken as prescribed. The Egyptian patients attended the outpatient clinic of the University Hospital, which is the main source of free and specialized medical care to patients. Since the preferred DMARD is not always available in this hospital, the physicians sometimes prescribe additional alternative DMARDs. Consequently, the medical records of some Egyptian patients are likely to have included more DMARDs than they actually received.

This study has some specific methodological limitations that suggest caution in interpreting the results. Since the results are based on a relatively small and selective sample size, the conclusions should be considered tentative and exploratory and have to be confirmed in further studies with larger numbers of participants. Also, the patient selection method used in this study limits the generalizability of the findings. The Egyptian patient group consisted of a convenience sample of consecutive female RA patients visiting the outpatient clinic. Since pain has been found to be independently associated with variables such as age and gender,[2,56] the Dutch patients were group-matched to these Egyptian patients. However, the epidemiology of RA in Egypt differs considerably from that in The Netherlands. The age, age at onset, and disease duration of Egyptian RA patients is generally much lower than those of Western patients and the female to male ratio higher.[57] Therefore, the findings of this study should not be generalized to the general Dutch and Egyptian RA populations. Moreover, physical and radiographic examinations were performed by a single physician. Although this physician was trained in reading and scoring radiographic films and performing joint examinations, the reliability of these examinations could not be assessed. Finally, the difference in administration of the questionnaires to the Egyptian and Dutch patients could have introduced bias into the study. The presence of the physician in the Egyptian group may have consciously or unconsciously motivated patients to under- or over-report their pain and disability in order to please the doctor or receive better treatment.

Despite these limitations, this study offers some interesting insights into the relationship between culture and pain in RA and offers some support to the idea that there are ethnocultural differences in pain reporting between patients from Mediterranean and Western European countries. Given our increasingly multicultural society and the growing number of multinational clinical trials, it is important that we recognize that the experience of pain is affected by more than clinical variables alone. However, the study also points to the need for further research to better understand the present findings. Future studies should attempt to further explore the mechanisms that may underlie ethnocultural differences in pain reports.

References

1. Moore R, Brodsgaard I. Cross-cultural investigations of pain. In: Crombie IK, Croft PR, Linton SJ, LeResche L, Von Korff M, eds. Epidemiology of pain. Seattle, WA: IASP Press; 1999:53–80.
2. Green CR, Ndao-Brumblay SK, Nagrant AM, Baker TA, Rothman E. Race, age, and gender influences among clusters of African American and white patients with chronic pain. J Pain 2004;5:171–82.

3. Bates MS, Edwards WT, Anderson KO. Ethnocultural influences on variation in chronic pain perception. Pain 1993;52:101–12.
4. Chibnall JT, Tait RC, Andresen EM, Hadler NM. Race and socioeconomic differences in post-settlement outcomes for African American and Caucasian Workers' Compensation claimants with low back injuries. Pain 2005;114:462–72.
5. Edwards RR, Doleys DM, Fillingim RB, Lowery D. Ethnic differences in pain tolerance: clinical implications in a chronic pain population. Psychosom Med 2001;63:316–23.
6. Faucett J, Gordon N, Levine J. Differences in postoperative pain severity among four ethnic groups. J Pain Symptom Manage 1994;9:383–9.
7. Green CR, Baker TA, Sato Y, Washington TL, Smith EM. Race and chronic pain: a comparative study of young black and white Americans presenting for management. J Pain 2003;4:176–83.
8. McCracken LM, Matthews AK, Tang TS, Cuba SL. A comparison of blacks and whites seeking treatment for chronic pain. Clin J Pain 2001;17:249–55.
9. Riley JL, 3rd, Wade JB, Myers CD, Sheffield D, Papas RK, Price DD. Racial/ethnic differences in the experience of chronic pain. Pain 2002;100:291–8.
10. Thomas VJ, Rose FD. Ethnic differences in the experience of pain. Soc Sci Med 1991;32:1063–6.
11. Weisenberg M, Caspi Z. Cultural and educational influences on pain of childbirth. J Pain Symptom Manage 1989;4:13–9.
12. Cohen MZ, Musgrave CF, Munsell MF, Mendoza TR, Gips M. The cancer pain experience of Israeli and American patients 65 years and older. J Pain Symptom Manage 2005;30:254–63.
13. Cano A, Mayo A, Ventimiglia M. Coping, pain severity, interference, and disability: the potential mediating and moderating roles of race and education. J Pain 2006;7:459–68.
14. Watson PJ, Latif RK, Rowbotham DJ. Ethnic differences in thermal pain responses: a comparison of South Asian and White British healthy males. Pain 2005;118:194–200.
15. Portenoy RK, Ugarte C, Fuller I, Haas G. Population-based survey of pain in the United States: differences among white, African American, and Hispanic subjects. J Pain 2004;5:317–28.
16. Edwards CL, Fillingim RB, Keefe F. Race, ethnicity and pain. Pain 2001;94:133–7.
17. Chaturvedi N. Ethnicity as an epidemiological determinant — crudely racist or crucially important? Int J Epidemiol 2001;30:925–7.
18. Ernst G. The myth of the 'Mediterranean syndrome': do immigrants feel different pain? Ethn Health 2000;5:121–6.
19. Todd KH. Pain assessment and ethnicity. Ann Emerg Med 1996;27:421–3.
20. Chapman WP, Jones CM. Variations in cutaneous and visceral pain sensitivity in normal subjects. J Clin Invest 1944;23:81–91.
21. Zborowski M. Cultural components in response to pain. J Soc Issues 1952;8:16–30.
22. Neumann L, Buskila D. Ethnocultural and educational differences in Israeli women correlate with pain perception in fibromyalgia. J Rheumatol 1998;25:1369–73.
23. Abdel-Nasser AM. Egyptian and Dutch rheumatoid arthritis patients: a biopsychosocial analysis [thesis]. Enschede: University of Twente; 1996.

24. Arnett FC, Edworthy SM, Bloch DA, et al. The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. Arthritis Rheum 1988;31:315–24.
25. Jensen MP, Karoly P. Self-report scales and procedures for assessing pain in adults. In: Turk DC, Melzack R, eds. Handbook of Pain Assessment. New York: Guilford Press; 2001:15–34.
26. Jensen MP. The validity and reliability of pain measures in adults with cancer. J Pain 2003;4:2–21.
27. Jensen MP, Karoly P, Braver S. The measurement of clinical pain intensity: a comparison of six methods. Pain 1986;27:117–26.
28. Kremer E, Atkinson JH, Ignelzi RJ. Measurement of pain: patient preference does not confound pain measurement. Pain 1981;10:241–8.
29. Clark P, Lavielle P, Martinez H. Learning from pain scales: patient perspective. J Rheumatol 2003;30:1584–8.
30. Ferraz MB, Quaresma MR, Aquino LR, Atra E, Tugwell P, Goldsmith CH. Reliability of pain scales in the assessment of literate and illiterate patients with rheumatoid arthritis. J Rheumatol 1990;17:1022–4.
31. ten Klooster PM, Vlaar AP, Taal E, et al. The validity and reliability of the graphic rating scale and verbal rating scale for measuring pain across cultures: a study in Egyptian and Dutch women with rheumatoid arthritis. Clin J Pain 2006;22:827–30.
32. Bruce B, Fries JF. The Stanford Health Assessment Questionnaire: a review of its history, issues, progress, and documentation. J Rheumatol 2003;30:167–78.
33. Fries JF, Spitz P, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. Arthritis Rheum 1980;23:137–45.
34. El Meidany YM, El Gaafary MM, Ahmed I. Cross-cultural adaptation and validation of an Arabic Health Assessment Questionnaire for use in rheumatoid arthritis patients. Joint Bone Spine 2003;70:195–202.
35. Bijlsma JW, Oude Heuvel CH, Zaalberg A. Development and validation of the Dutch questionnaire capacities of daily life (VDF) for patients with rheumatoid arthritis. J Rehabil Sci 1990;3:71–4.
36. Siegert CE, Vleming LJ, Vandenbroucke JP, Cats A. Measurement of disability in Dutch rheumatoid arthritis patients. Clin Rheumatol 1984;3:305–9.
37. van der Heijde DM, van Riel PL, van de Putte LB. Sensitivity of a Dutch Health Assessment Questionnaire in a trial comparing hydroxychloroquine vs. sulphasalazine. Scand J Rheumatol 1990;19:407–12.
38. Zandbelt MM, Welsing PM, van Gestel AM, van Riel PL. Health Assessment Questionnaire modifications: is standardisation needed? Ann Rheum Dis 2001;60:841–5.
39. Deandrade JR, Casagrande PA. A seven-day variability study of 499 patients with peripheral rheumatoid arthritis. Arthritis Rheum 1965;8:302–34.
40. van der Heijde D. How to read radiographs according to the Sharp/van der Heijde method. J Rheumatol 2000;27:261–3.
41. Steinbrocker O, Traeger CH, Battman RG. Therapeutic criteria in rheumatoid arthritis. JAMA 1949;140:659–62.
42. Prevoo ML, van 't Hof MA, Kuper HH, van Leeuwen MA, van de Putte LB, van Riel PL. Modified Disease Activity Scores that include twenty-eight–joint counts: development and

validation in a prospective longitudinal study of patients with rheumatoid arthritis. Arthritis Rheum 1995;38:44–8.
43. Lipton JA, Marbach JJ. Ethnicity and the pain experience. Soc Sci Med 1984;19:1279–98.
44. Andrulis DP. Access to care is the centerpiece in the elimination of socioeconomic disparities in health. Ann Intern Med 1998;129:412–6.
45. Nguyen M, Ugarte C, Fuller I, Haas G, Portenoy RK. Access to care for chronic pain: racial and ethnic differences. J Pain 2005;6:301–14.
46. Pincus T, Esther R, DeWalt DA, Callahan LF. Social conditions and self-management are more powerful determinants of health than access to care. Ann Intern Med 1998;129:406–11.
47. Jordan MS, Lumley MA, Leisen JC. The relationships of cognitive coping and pain control beliefs to pain and adjustment among African-American and Caucasian women with rheumatoid arthritis. Arthritis Care Res 1998;11:80–8.
48. Campbell CM, Edwards RR, Fillingim RB. Ethnic differences in responses to multiple experimental pain stimuli. Pain 2005;113:20–6.
49. Hastie BA, Riley JL, 3rd, Fillingim RB. Ethnic differences in pain coping: factor structure of the coping strategies questionnaire and coping strategies questionnaire-revised. J Pain 2004;5:304–16.
50. Diller A. Cross-cultural pain semantics. Pain 1980;9:9–26.
51. Fabrega H, Jr., Tyma S. Language and cultural influences in the description of pain. Br J Med Psychol 1976;49:349–71.
52. Wolff BB. Ethnocultural factors influencing pain and illness behavior. Clin J Pain 1985;1:23–30.
53. Zatzick DF, Dimsdale JE. Cultural variations in response to painful stimuli. Psychosom Med 1990;52:544–57.
54. Greenwald HP. Interethnic differences in pain perception. Pain 1991;44:157–63.
55. Edwards RR, Fillingim RB. Ethnic differences in thermal pain responses. Psychosom Med 1999;61:346–54.
56. Affleck G, Tennen H, Keefe FJ, et al. Everyday life with osteoarthritis or rheumatoid arthritis: independent effects of disease and gender on daily pain, mood, and coping. Pain 1999;83:601–9.
57. Abdel-Nasser AM, Rasker JJ, Valkenburg HA. Epidemiological and clinical aspects relating to the variability of rheumatoid arthritis. Semin Arthritis Rheum 1997;27:123–40.

7 Confirmatory factor analysis of the Arthritis Impact Measurement Scales 2 Short Form in patients with rheumatoid arthritis

P.M. ten Klooster
M.M. Veehof
E. Taal
P.L.C.M. van Riel
M.A.F.J. van de Laar

Arthritis & Rheumatism (Arthritis Care & Research), in press.

Abstract

Objective. To examine the factorial validity of the short form Arthritis Impact Measurement Scales 2 (AIMS2-SF) in patients with rheumatoid arthritis (RA).

Methods. Data were used from a sample of 279 patients with active RA who completed the long form AIMS2 before starting treatment with tumor necrosis factor-α blocking agents. Confirmatory factor analyses were conducted to test and compare the fit of the currently used theoretical measurement model of the AIMS2-SF, originally suggested for the long form AIMS2, and 2 alternative models based on previous exploratory research.

Results. A model with the physical dimension divided into upper and lower body limitations was superior to the current model, and both models provided a clearly better fit than a model without a separate symptom dimension. Under the restrictive assumption of uncorrelated error terms, none of the models achieved a consistent and acceptable fit as judged by several goodness-of-fit indices. Allowing error covariances between 6 pairs of items within the same dimension resulted in an improved and acceptable fit of both the current model and the model with a separate upper and lower body component.

Conclusion. This study generally supports the factorial validity of the AIMS2-SF and suggests the use of separate scores for upper and lower body limitations. Further research is needed to resolve the issue of high error correlations associated with particular items.

Introduction

Since its introduction by Meenan et al in 1980,[1] the Arthritis Impact Measurement Scales (AIMS) and the revised AIMS2[2] have been widely used for measuring health status in patients with rheumatic diseases. Because the length of the AIMS2 limited its use in clinical research and routine practice, Guillemin et al[3] developed a 26-item short form of the questionnaire (AIMS2-SF) for patients with rheumatoid arthritis (RA). The investigators attempted to preserve the content validity of the questionnaire and the final short form version showed similar psychometric properties as the original long form AIMS2.

Despite its increasing use, however, empirical support for the factorial validity of the AIMS2-SF is still somewhat limited. For multidimensional measures such as the AIMS2-SF, factorial validation is important for understanding how to score and interpret the different dimensions. According to the study by Guillemin et al,[3] the AIMS2-SF is usually scored using the 5 dimensions originally suggested as second-order components for the long form AIMS2. Although a number of studies have examined the underlying structure of the AIMS2-SF, all of the studies used exploratory factor analysis. Moreover, somewhat different factor solutions were found across different study samples (Table 1).

Analysis of the original AIMS2-SF in a French cohort study of patients with RA starting treatment with methotrexate identified 5 factors, representing upper body function, lower body function, affect, symptom, and social interaction.[3] This factor structure was indeed close to the original dimensions of the long form, with the physical dimension, however, split into 2 parts. In a cross-sectional study of patients with osteoarthritis (OA) in the US, Ren et al[4] reported a very similar 5-factor solution. Contrary to these findings, a Dutch study using cross-sectional data from 3 studies of outpatients with RA found a 3-factor solution representing a physical, psychological, and social dimension.[5] All lower and upper body items loaded on 1 factor and the 3 symptom items loaded on both the physical and psychological dimension, but more strongly on the psychological dimension. This 3-factor solution was closely replicated in another cross-sectional study of German patients with OA in primary care.[6]

Given the inconclusiveness of previous efforts to identify the factor structure of the AIMS2-SF, it is unclear whether its current scoring procedure appropriate. Therefore, the objective of the present study was to further examine the factor structure of the AIMS2-SF using confirmatory factor analysis (CFA) in a new sample of patients with active RA. CFA provides a more powerful test of factorial validity than exploratory approaches by examining whether a hypothesized measurement model adequately fits the data of a given sample. Moreover, it allows for the comparison of competing

Table 1. Sample characteristics and factor structures from previous exploratory factor analyses of the Arthritis Impact Measurement Scales 2 Short Form (AIMS2-SF)*

Author, year (ref.)	Study sample	No. of patients	Study design	Disease severity†	Mean age	Mean disease duration	Female, %	Factor solution‡
Guillemin et al, 1997[3]	RA	127	Prospective cohort of patients starting methotrexate	Severe	51	8	76	5 factors: upper body, lower body, affect, symptom, social interaction
Ren et al, 1999[4]	OA	147	Cross-sectional performance test sample	Moderate	66	14	81	5 factors: upper body, lower body, affect, symptom, social interaction
Taal et al, 2003[5]	RA	587	Cross-sectional out-patients samples	Moderate	61	16	63	3 factors: physical, psychological, social
Rosemann et al, 2005[6]	OA	220	Cross-sectional primary care patients sample	Moderate	47	10	44	3 factors: physical, psychological, social

* RA = rheumatoid arthritis; OA = osteoarthritis.
† Arbitrarily defined using the reported mean AIMS2-SF symptom scale score (<3 = mild, 3–6 = moderate, >6 = severe).
‡ Items from the role component were excluded from all factor analyses

models and selection of the best fitting model. Finally, CFA can be used to refine existing models to increase their parsimony and statistical power. Therefore, CFAs were conducted to test the current measurement model of the AIMS2-SF and compare it with 2 alternative models based on previous exploratory analyses.

Patients and methods

Participants

Participants in this study were patients with RA who were included in the Dutch Rheumatoid Arthritis Anti-Tumor Necrosis Factor Monitoring (DREAM) register. The DREAM register is an ongoing observational cohort study of patients with RA starting anti-tumor necrosis factor treatment in 11 hospitals in The Netherlands. Details on the inclusion criteria, methodology, and cohort characteristics of the DREAM study are reported elsewhere.[7] For this part of the study, we used data from a subset of patients who additionally completed the AIMS2 at baseline. The DREAM cohort study was approved by the appropriate institutional ethics committees and all patients provided written informed consent.

Measures

The AIMS2 is a self-administered questionnaire designed to measure various dimensions of health status in patients with arthritis.[2] It has been used in different rheumatic conditions and an increasing number of validated translations are available.[8–15] The Dutch version that was used in this study has shown good psychometric properties in patients with RA.[16–18] The core part of the AIMS2 contains 57 items that are categorized in 12 scales representing different areas of health. Each scale contains 4 or 5 items measured on 5-point Likert-type rating scales. The scales can be combined into 5 second-order summary component scores: physical (mobility level, walking and bending, hand and finger function, arm function, self-care, household tasks), affect (level of tension, mood), symptom (arthritis pain), social interaction (social activity, support from family), and role (work).

The AIMS2-SF[3] comprises 26 items from the long form version. Item reduction for the AIMS2-SF was mainly based on a Delphi exercise of patients' and experts' judgments of relevance and was aimed at preserving the 5 original dimensions of the AIMS2. The AIMS2-SF showed similar convergent validity, reliability, and sensitivity to change as the original AIMS2. Several studies have since confirmed the psychometric qualities of the AIMS2-SF.[4–6,18,19] Although in the Dutch version[5] items 31 and 38 from the long form version are recommended instead of items 33 and 42, the items from the original French version were used in the current study. The AIMS2-SF is usually scored by combining the items from the 5 dimensions of the long form AIMS2. First, responses

to 16 items are recoded so that higher scores indicate worse health. After recoding the raw responses, the scores on each item within a dimension are summed and transformed to a score ranging from 0 to 10, with higher scores representing poorer health status.

Analyses

CFAs were conducted using LISREL 8.70 (Scientific Software International, Lincolnwood, IL). Because the AIMS2-SF consists of ordinal Likert-type items, robust maximum likelihood (RML) estimation with Satorra–Bentler (SB)-scaled statistics was used.[20] This estimation procedure corrects for non-normality using the asymptotic covariance matrix and provides more accurate fit indices. Given the relatively small sample size in this study, this procedure was considered to be preferable to alternative asymptotic distribution-free methods.[21–24]

The conventional overall measure of fit in CFA is the chi-square statistic, where small, non-significant values indicate a good fit with the data. Because of several problems associated with this statistic, such as its sensitivity to large sample sizes, the chi-square statistic is likely to overstate the lack of fit of a model. Therefore, this measure was primarily used to statistically compare the relative fit of the nested models by means of chi-square difference tests. Because the simple difference between 2 SB-scaled chi-squares does not yield a correct test statistic, we used the SB-scaled difference test statistic ($\Delta SB\chi^2$) procedure.[25]

Additionally, a variety of indices have been developed that account for the problems associated with the chi-square statistic. As suggested by Hu and Bentler,[26,27] multiple fit indices were used to examine the fit of the models: the non-normed fit index (NNFI; also known as the Tucker–Lewis index), the comparative fit index (CFI), the standardized root mean square residual (SRMR), and the root mean square error of approximation (RMSEA). Values ≥0.95 for the NNFI and CFI, ≤0.08 for the SRMR, and ≤0.06 for the RMSEA indicate a good fit between the hypothesized model and the data.[26,27]

As in other studies exploring the factor structure of the AIMS2-SF, the 2 role items were excluded from the factor analyses because these are not completed by patients who are unemployed, disabled, or retired at the time of study. At baseline, 48% of the patients fell into one of these categories. Moreover, constructing a factor with only 2 items may lead to possible identification and convergence problems.[28–30] The percentage of missing values for the 24 remaining items ranged from 0.0 to 3.6. Because no missing data patterns were identified, missing values were imputed using the expectation-maximization algorithm procedure in LISREL.

We tested and compared 3 different factor models. Model 1 is the currently used measurement model of the AIMS2-SF that combines the remaining items of the AIMS2-

SF in the same dimensions as the long form AIMS2: physical, social interaction, symptom, and affect. Model 2 is based on the findings of 2 previous exploratory factor analyses indicating that the items in the physical dimension should be split into 2 factors, 1 reflecting upper body limitations and 1 reflecting lower body limitations.[3,4] Finally, Model 3 reflects the 3-factor solution found in 2 independent studies, in which the symptom items did not form a separate factor but, instead, loaded highly on the psychological (affect) dimension.[5,6]

In all models, the items were constrained to load on a single factor, the variance of the hypothesized latent factors was fixed at 1.0, and the factors were allowed to correlate freely. Initial comparison of the models was based on the restrictive assumption that the error terms of the items were uncorrelated. Correlated error terms between items generally indicate that these items share a common variance that is not accounted for by the hypothesized factor structure, such as the presence of 1 or more meaningful unspecified factors. However, findings of correlated error terms are not unusual in the validation of assessment instruments in general, and of psychological measures in particular.[31] The presence of correlated error terms can also reflect certain method effects, especially perceived redundancy or overlap in item content.[32,33] A model can be further improved by allowing such error terms to correlate, but only when this can be justified and interpreted substantively.[34] In general, it is considered justifiable to allow error correlations between items within the same factor. When none of the restrictive models achieved acceptable fit, the constraints of the models were relaxed one at a time by allowing the error terms with the largest modification index within each factor to correlate, provided that this made substantive sense.

Results

Between March 2003 and November 2004, 302 patients were enrolled in this part of the study. Of these patients, 279 (92.4%) completed the long form AIMS2 at study entry. There were no significant differences in demographic and clinical characteristics between patients who did and those who did not complete the AIMS2 (data not shown). Of the participants who completed the AIMS2, 70% were female and the mean ± SD age and median (interquartile range) disease duration were 54.6 ± 12.8 years and 8.0 (3.0–15.0) years, respectively. Assessment of disease severity at baseline generally indicated severe RA, with a mean ± SD 28-joint Disease Activity Score[35] of 5.43 ± 1.22, AIMS2-SF symptom scale score of 6.9 ± 2.0, and Health Assessment Questionnaire Disability Index[36,37] score of 1.49 ± 0.59.

The means, SDs, skewness, and kurtosis of the recoded scores on the items are listed in Table 2. The full range of the response options was used for all items, except items 32 and 53, for which the highest response option (“never” or “no days”) was not used.

Skewness and kurtosis values of the items suggested low to moderate non-normality of the items, further supporting the use of RML estimation with SB-scaled statistics.

Table 2. Distribution of the recoded Arthritis Impact Measurement Scales 2 Short Form (AIMS2-SF) scores*

	Mean ± SD	Skewness	Kurtosis
1. Use a car or public transportation	1.95 ± 1.01	0.49	–0.65
5. In a bed or a chair for most or all of the day	1.82 ± 0.96	0.65	–0.60
6. Trouble doing vigorous activities	4.59 ± 0.77	–1.35	0.59
7. Trouble walking several blocks or climbing a few flights of stairs	3.76 ± 1.19	–0.36	–0.82
10. Unable to walk unless assisted	2.00 ± 1.36	0.76	–0.81
11. Write with a pen or pencil	2.34 ± 1.07	0.24	–0.60
12. Button a shirt or blouse	2.56 ± 1.14	0.17	–0.63
13. Turn a key in a lock	2.46 ± 1.06	0.17	–0.55
18. Comb or brush your hair	2.10 ± 1.03	0.39	–0.69
20. Reach a shelve that was above your head	2.76 ± 1.25	0.10	–0.75
22. Need help to get dressed	1.90 ± 1.10	0.69	–0.66
24. Need help to get in or out of bed	1.45 ± 0.79	1.31	0.46
29. Get together with friends or relatives	2.95 ± 0.62	–0.03	1.29
32. On the telephone with close friends or relatives	2.48 ± 0.74	–0.09	–0.16
33. Go to a meeting of a church, club, team, or other group	3.66 ± 0.95	–0.14	–0.49
35. Family and friends sensitive to your personal needs	1.91 ± 0.98	0.54	–0.62
39. Severe pain from your arthritis	3.72 ± 1.07	–0.30	–0.65
41. Morning stiffness lasts more than 1 hour	3.49 ± 1.31	–0.26	–0.91
42. Pain makes it difficult for you to sleep	3.14 ± 1.15	–0.05	–0.57
48. Felt tense or high strung	2.67 ± 0.91	0.03	–0.17
49. Bothered by nervousness or your nerves	2.32 ± 0.92	0.16	–0.42
53. Enjoyed the things you do	2.19 ± 0.73	0.16	0.08
54. In low or very low spirits	3.11 ± 0.89	–0.04	–0.04
56. Others better off if you were dead	1.38 ± 0.76	1.55	1.09

* The item numbers refer to the original numbers in the long form AIMS2. Higher scores indicate poorer status for all items.

The results of the CFAs are presented in Table 3. None of the restricted models satisfied the recommended criteria of acceptable fit on any of the fit indices. From the 3 models tested, the alternative measurement model with the physical dimension split into upper and lower body limitations (model 2) provided the best fit to the data, with an NNFI and CFI close to 0.95 and an SRMR and RMSEA just above the cutoff values of 0.08 and 0.06, respectively. Although the fit indices of the current model (model 1) were only marginally worse, the difference between these nested models was significant ($\Delta SB\chi^2$ (4) = 87.44, $P < 0.001$). Both models performed substantially better than the model without a separate factor for symptoms (model 3).

Subsequent examination of the different models showed that the assumption of no correlation between the error terms did not hold. The modification indices indicated

Table 3. Summary of fit indices for the different models of the Arthritis Impact Measurement Scales 2 Short Form (AIMS2-SF)*

	χ^2	df	NNFI	CFI	SRMR	RMSEA (90% CI)
Model 1	805.03	246	0.92	0.93	0.10	0.09 (0.08–0.10)
Refined: 6 correlated errors	483.90	240	0.96	0.97	0.08	0.06 (0.05–0.07)
Model 2	727.17	242	0.93	0.94	0.10	0.09 (0.08–0.09)
Refined: 6 correlated errors	413.72	236	0.97	0.98	0.08	0.05 (0.04–0.06)
Model 3	1038.23	249	0.88	0.90	0.11	0.11 (0.10–0.11)
Refined: 7 correlated errors	549.51	242	0.95	0.96	0.10	0.07 (0.06–0.08)

* χ^2 = Satorra–Bentler–scaled chi-square; NNFI = non-normed fit index; CFI = comparative fit index; SRMR = standardized root mean square residual; RMSEA = root mean square error of approximation; 90% CI = 90% confidence interval; model 1 = current model (physical, social interaction, symptom, and affect); model 2 = alternative model with the physical dimension split into upper and lower body limitations; model 3 = alternative model with the symptom items loading on the affect dimension.

the presence of several high error covariances between pairs of items within the same dimension. The 6 largest modification indices (all >20) were consistently present in all models. These error covariances involved items 48 (felt tense or high strung) and 49 (bothered by nervousness or your nerves), items 6 (trouble doing vigorous activities) and 7 (trouble walking several blocks or climbing a few flights of stairs), items 22 (need help to get dressed) and 24 (need help to get in or out of bed), items 11 (write with a pen or pencil) and 12 (button a shirt or blouse), items 12 and 13 (turn a key in a lock), and items 11 and 13. In model 3, an additional high correlation was identified between the error terms of items 53 (enjoyed the things you do) and 54 (in low or very low spirits). All of these correlated error terms appeared to involve pairs of items with a high degree of similarity in feeling state or items reflecting similar degrees of severity on the same functional limitation.

Because allowing these error terms to correlate did not seriously compromise the structure of the original models, the models were respecified to include these correlations. The final refined models of the AIMS2-SF are shown in Figure 1, including the standardized factor loadings, correlations between factors, and correlations between error terms. The fit indices of the refined models with correlated error terms showed marked improvements in model fit for all 3 models (see Table 3). Moreover, the refined versions of both the current measurement model (model 1) and the alternative model with the physical dimension split into upper and lower body limitations (model 2) now satisfied the criteria for acceptable model fit for the NNFI, CFI, SRMR, and RMSEA. As with the restricted versions of the models, the latter model (model 2) performed significantly better than model 1 ($\Delta SB\chi^2$ (4) = 144.86, $P < 0.001$) and both models provided a clearly better fit than the model without a separate symptom factor (model 3).

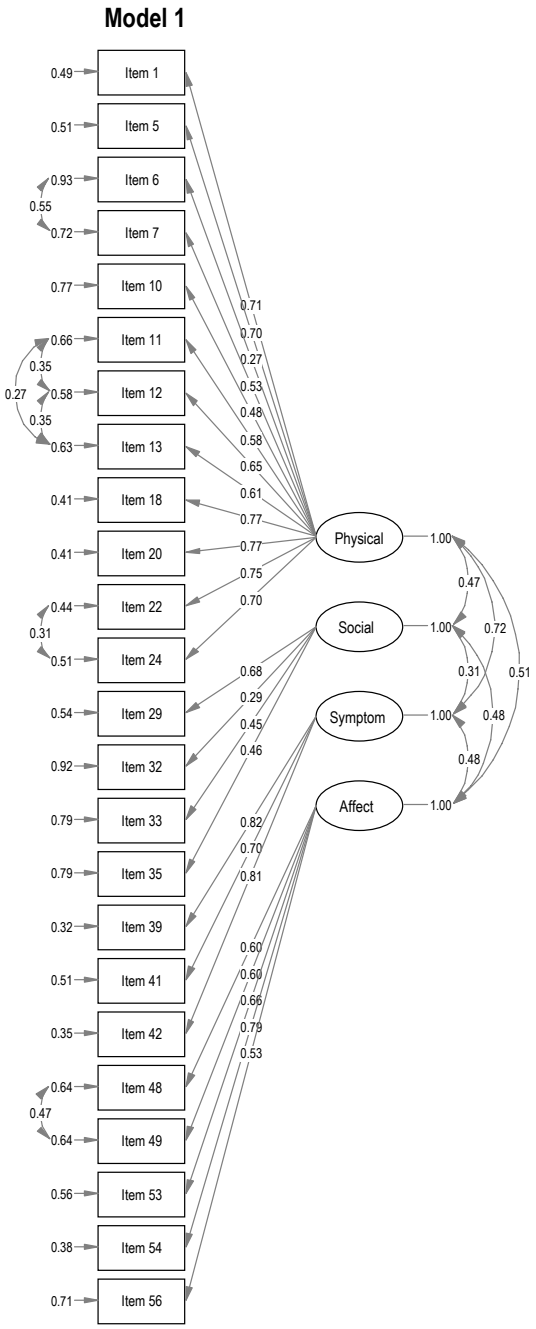

Figure 1. Standardized parameter estimates for the 3 refined models of the Arthritis Impact Measurement Scales 2 Short Form. Rectangles represent the observed variables (items) and ellipses represent the hypothesized latent constructs (factors). Values on the single-headed arrows leading from the factors to the items are standardized factor loadings. Values to the left of the items represent error variances. Values on the curved double-headed arrows are correlations between factors or error terms.

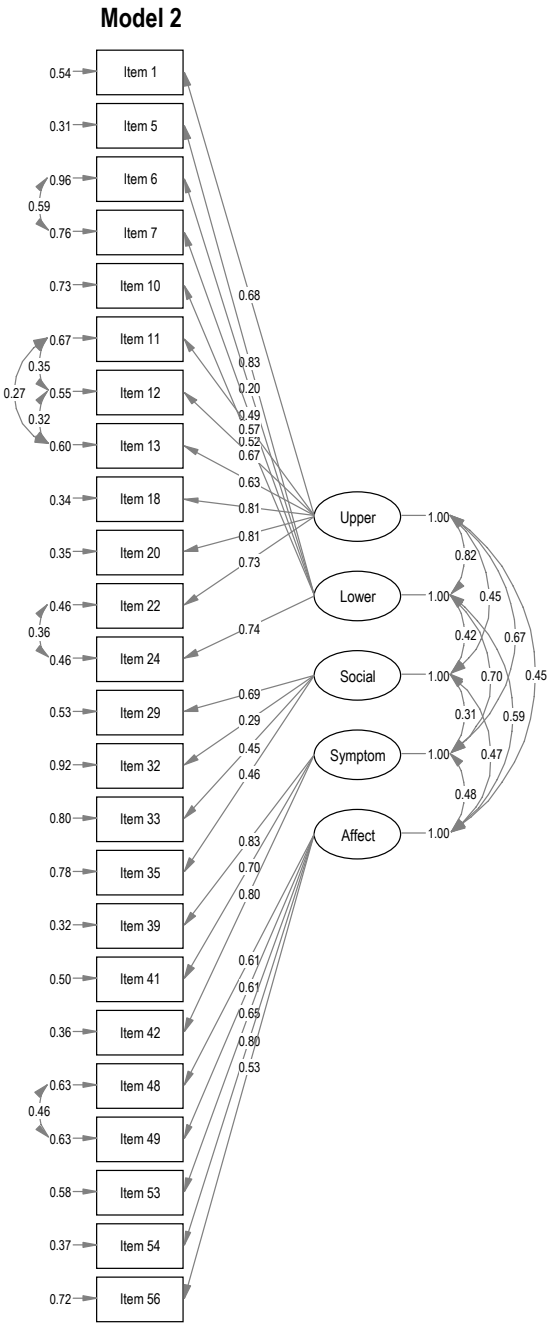

Figure 1. (Continued)

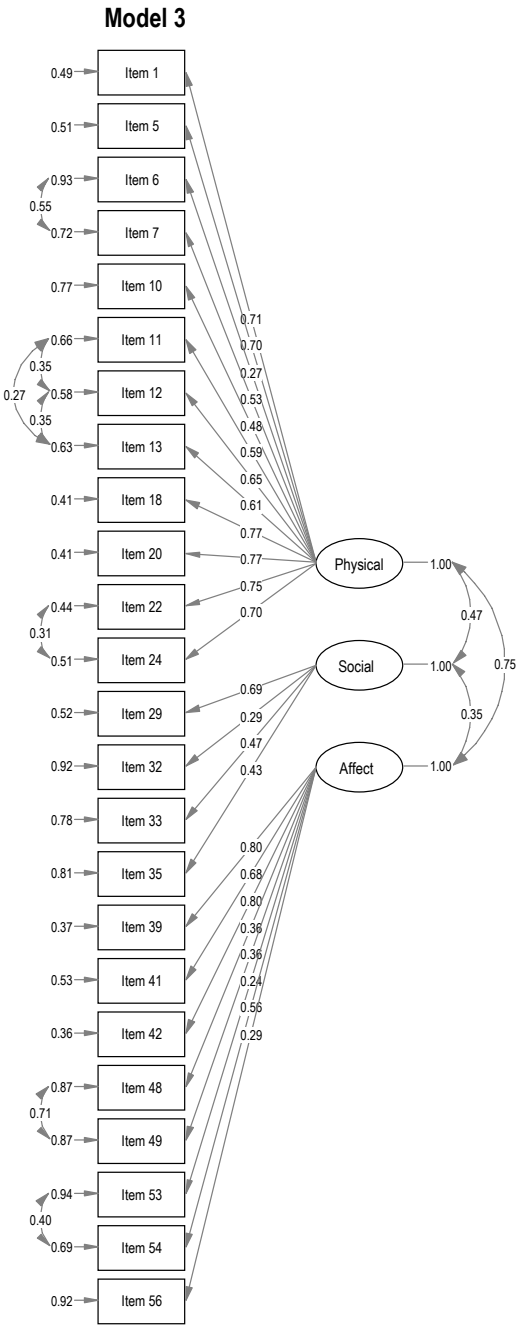

Figure 1. (Continued)

Discussion

The AIMS2-SF is usually scored by grouping its items into the same (second-order) components of the original long form AIMS2. Although this scoring procedure certainly has face validity, it has not been confirmed using appropriate statistical analyses. The current study used CFA to test and compare the goodness-of-fit of the current measurement model of the AIMS2-SF and 2 alternative models based on previous exploratory factor analyses. The results support the validity of the current measurement model of the AIMS2-SF, but suggest that an alternative model with separate upper and lower body limitations factors may be more meaningful.

The study confirms the findings by Guillemin et al[3] and Ren et al[4] that an upper body limitations factor should be distinguished from lower body limitations. This distinction was already suggested for version 1 of the long form AIMS[1,38] and is quite common in the assessment of physical function in rheumatic diseases. Despite the intuitive appeal of a distinction between upper and lower body limitations, the validity of this distinction depends critically on whether all items can be classified as involving upper or lower body functions. Although most of the items for physical function are clearly associated with either upper body activities (e.g., write with a pen or pencil) or lower body activities (e.g., trouble walking several blocks or climbing a few flights of stairs), the AIMS2-SF also contains some items for which this classification is not so clear. Items such as "need help to get dressed" or "trouble doing vigorous activities" may involve both upper and lower body functions or may be related to other factors such as physical condition or athleticism. Indeed, in the final model (model 2) the latter item loaded very poorly (0.20) on the specified lower body limitations factor. However, because this item demonstrated a similarly weak factor loading in both the current measurement model (model 1) and the alternative model without a symptom factor (model 3), it is suggested that this item may not represent the proposed physical function factor in general. All other items, however, demonstrated a sufficient loading on the specified upper or lower body limitations factor, supporting the use of separate scales for these factors.

The results also suggest the presence of unspecified factors or overlapping or redundant items, especially within the physical and psychological dimensions. The assumption of uncorrelated error terms did not hold for the items in the AIMS2-SF and several error covariances were added post hoc to the models. The high error correlation between items 11, 12, and 13 most likely indicates the presence of an additional factor related to restricted finger movement, which may be distinct from other physical limitations. Indeed, in the original long form AIMS2, these items were part of the hand and finger function subscale. Future studies could examine whether adding a separate factor for these items would improve the factor structure of the AIMS2-SF. Although

correlated error terms are generally considered indicative of the omission of 1 or more relevant factors, they can also point to other types of method effects such as overlapping item content or perceived commonality.[33] A review of the other items with high error term correlations indeed showed that these items used the same wordings (e.g., "need help") or assessed very similar feelings. Since high correlations between the error terms of such items are not uncommon, we considered it justifiable to allow error covariances between pairs of items within the same dimension. Nonetheless, this finding does deserve further attention and the final models need to be cross-validated to increase confidence in the replicability of the post hoc modifications. Moreover, further research could focus more closely on possible modifications in the wording of these items to improve the precision of the AIMS2-SF.

Also, it is possible that the different factor structures found in previous studies are the result of specific differences in the patient samples studied. For instance, Guillemin et al[3] developed the AIMS2-SF using data from a prospective cohort study of patients with RA with severe disease. Other studies, however, investigated its factor structure in OA,[4,6] which is a different and unique disease condition, or used cross-sectional data from patients with RA with moderate disease severity.[5] Although we did not find a clear association between sample characteristics and the reported factor structures in these studies, the results of the current study should be cross-validated in other samples, such as OA patients or RA patients with less severe disease.

An important limitation of this study is the omission of the proposed role dimension of the AIMS2-SF in the models. This is the result of a general problem associated with the use of both the long form and the short form of the AIMS2. Because patients who are unemployed, disabled, or retired at the time of the study are asked to skip the items from this dimension, missing values for these items are often high. Consequently, the items from this dimension are usually excluded from factor analyses and the actual presence of a separate role dimension has as yet not been confirmed.

In summary, this study supports the factorial validity of the AIMS2-SF and suggests the use of separate scales for upper and lower body limitations in scoring the questionnaire. The results also point to certain problems associated with some of the items that need further study.

Acknowledgments

We thank T. van Gaalen, W. Kievit and P. Welsing for their contribution to the organization of the study and data management. We thank the following rheumatologists and research nurses for their assistance in patient recruitment and data collection: J. Alberts, C. Allaart, A. ter Avest, P. Barrera Rico, T. Berends, H. Bernelot Moens, K. Bevers, C. Bijkerk, , A. van der Bijl, J. de Boer, A. Boonen, E. ter Borg, E. Bos, A. Branten, F. Breed-

veld, H. van den Brink, J. Bürer, G. Bruyn, H. Cats, M. Creemers, J. Deenen, C. De Gendt, K. Drossaers-Bakker, A. van Ede, A. Eijsbouts, S. Erasmus, M. Franssen, I. Geerdink, M. Geurts, E. Griep, E. de Groot, C. Haagsma, H. Haanen, J.Harbers, A. Hartkamp, J. Haverman, H. van Heereveld, A. van de Helm-van Mil, I. Henkes, S. Herfkens, M. Hoekstra, K. van de Hoeven, D.M. Hofman, M. Horbeek, F. van den Hoogen, P. M. Houtman, T. Huizinga, H. Hulsmans, P. Jacobs, T. Jansen, M. Janssen, M. Jeurissen, A. de Jong, M. Kleine Schaar, G. Kloppenburg, H. Knaapen, P. Koelmans, M. Kortekaas, B. Kraft, A. Krol, M. Kruijssen, D. Kuiper-Geertsma, I. Kuper, R. Laan, J. van de Laan, J. van Laar, P. Lanting, H. Lim, S. van der Linden, A. Mooij, J. Moolenburgh, N. Olsthoorn, P. van Oijen, M. van Oosterhout, J. Oostveen, P. van 't Pad Bosch, K. Rasing, K. Ronday, D. de Rooij, L. Schalkwijk, P. Seys, P. de Sonnaville, A. Spoorenberg, A. Stenger, G. Steup, W. Swen, J. Terwiel, M. van der Veen, M. Veerkamp, C. Versteegden, H. Visser, C. Vogel, M. Vonk, H. Vonkeman, A. Westgeest, H. van Wijk, N. Wouters.

References

1. Meenan RF, Gertman PM, Mason JH. Measuring health status in arthritis: the Arthritis Impact Measurement Scales. Arthritis Rheum 1980;23:146–52.
2. Meenan RF, Mason JH, Anderson JJ, Guccione AA, Kazis LE. AIMS2. The content and properties of a revised and expanded Arthritis Impact Measurement Scales health status questionnaire. Arthritis Rheum 1992;35:1–10.
3. Guillemin F, Coste J, Pouchot J, Ghezail M, Bregeon C, Sany J. The AIMS2-SF: a short form of the Arthritis Impact Measurement Scales 2. French Quality of Life in Rheumatology Group. Arthritis Rheum 1997;40:1267–74.
4. Ren XS, Kazis L, Meenan RF. Short-form Arthritis Impact Measurement Scales 2: tests of reliability and validity among patients with osteoarthritis. Arthritis Care Res 1999;12:163–71.
5. Taal E, Rasker JJ, Riemsma RP. Psychometric properties of a Dutch short form of the Arthritis Impact Measurement Scales 2 (Dutch-AIMS2-SF). Rheumatology (Oxford) 2003;42:427–34.
6. Rosemann T, Korner T, Wensing M, Schneider A, Szecsenyi J. Evaluation and cultural adaptation of a German version of the AIMS2-SF questionnaire (German AIMS2-SF). Rheumatology (Oxford) 2005;44:1190–5.
7. Kievit W, Fransen J, Oerlemans AJ, et al. The efficacy of anti-TNF in rheumatoid arthritis, a comparison between randomized controlled trials and clinical practice. Ann Rheum Dis 2007;66:1473–8.
8. Archenholtz B, Bjelle A. Reliability, validity, and sensitivity of a Swedish version of the revised and expanded Arthritis Impact Measurement Scales (AIMS2). J Rheumatol 1997;24:1370–7.
9. Arkela-Kautiainen M, Kauppi M, Heikkila S, Kautiainen H, Malkia E, Leirisalo-Repo M. Evaluation of the Arthritis Impact Measurement Scales (AIMS2) in Finnish patients with rheumatoid arthritis. Scand J Rheumatol 2003;32:300–5.

10. Atamaz F, Hepguler S, Oncu J. Translation and validation of the Turkish version of the Arthritis Impact Measurement Scales 2 in patients with knee osteoarthritis. J Rheumatol 2005;32:1331–6.
11. Brandao L, Ferraz MB, Zerbini CA. Health status in rheumatoid arthritis: cross cultural evaluation of a Portuguese version of the Arthritis Impact Measurement Scales 2 (BRASIL-AIMS2). J Rheumatol 1998;25:1499–501.
12. Chu EM, Chiu KY, Wong RW, Tang WM, Lau CS. Translation and validation of Arthritis Impact Measurement Scales 2 into Chinese: CAIMS2. Arthritis Rheum 2004;51:20–7.
13. Neumann L, Dudnik Y, Bolotin A, Buskila D. Evaluation of a Hebrew version of the revised and expanded Arthritis Impact Measurement Scales (AIMS2) in patients with fibromyalgia. J Rheumatol 1999;26:1816–21.
14. Pouchot J, Guillemin F, Coste J, Bregeon C, Sany J. Validity, reliability, and sensitivity to change of a French version of the Arthritis Impact Measurement Scales 2 (AIMS2) in patients with rheumatoid arthritis treated with methotrexate. J Rheumatol 1996;23:52–60.
15. Salaffi F, Piva S, Barreca C, et al. Validation of an Italian version of the Arthritis Impact Measurement Scales 2 (ITALIAN-AIMS2) for patients with osteoarthritis of the knee. Gonarthrosis and Quality of Life Assessment (GOQOLA) Study Group. Rheumatology (Oxford) 2000;39:720–7.
16. Riemsma RP, Taal E, Rasker JJ, Houtman PM, van Paassen HC, Wiegman O. Evaluation of a Dutch version of the AIMS2 for patients with rheumatoid arthritis. Br J Rheumatol 1996;35:755–60.
17. Evers AW, Taal E, Kraaimaat FW, et al. A comparison of two recently developed health status instruments for patients with arthritis: Dutch-AIMS2 and IRGL. Br J Rheumatol 1998;37:157–64.
18. Taal E, Rasker JJ, Riemsma RP. Sensitivity to change of AIMS2 and AIMS2-SF components in comparison to M-HAQ and VAS-pain. Ann Rheum Dis 2004;63:1655–8.
19. Haavardsholm EA, Kvien TK, Uhlig T, Smedstad LM, Guillemin F. A comparison of agreement and sensitivity to change between AIMS2 and a short form of AIMS2 (AIMS2-SF) in more than 1,000 rheumatoid arthritis patients. J Rheumatol 2000;27:2810–6.
20. Jöreskog KG, Sörbom D, Du Toit S, Du Toit M. LISREL 8: new statistical features. Lincolnwood, IL: Scientific Software International; 2001.
21. Chou CP, Bentler PM, Satorra A. Scaled test statistics and robust standard errors for nonnormal data in covariance structure analysis: a Monte Carlo study. Br J Math Stat Psychol 1991;44:347–57.
22. Curran PJ, West SG, Finch JF. The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. Psychol Methods 1996;1:16–29.
23. Hu LT, Bentler PM, Kano Y. Can test statistics in covariance structure analysis be trusted? Psychol Bull 1992;112:351–62.
24. Jöreskog KG. Structural equation modeling with ordinal variables using LISREL. Lincolnwood, IL: Scientific Software International; 2002–2005.
25. Satorra A, Bentler P. A scaled difference chi-square test statistic for moment structure analysis. Psychometrika 2001;66:507–14.

26. Hu LT, Bentler PM. Fit indices in covariance structure modeling: sensitivity to underparameterized model misspecification. Psychol Methods 1998;3:424–53.
27. Hu LT, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Structural Equation Modeling 1999;6:1–55.
28. Bentler PM, Chou CP. Practical issues in structural modeling. Sociol Methods Res 1987;16:78–117.
29. Anderson JC, Gerbing DW. The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. Psychometrika 1984;49:155–73.
30. Boomsma A. Nonconvergence, improper solutions, and starting values in LISREL maximum likelihood estimation. Psychometrika 1985;50:229–42.
31. Byrne BM, Baron P, Larsson B, Melin L. The Beck Depression Inventory: testing and cross-validating a second-order factorial structure for Swedish nonclinical adolescents. Behav Res Ther 1995;33:345–56.
32. Byrne BM. Factor analytic models: viewing the structure of an assessment instrument from three perspectives. J Pers Assess 2005;85:17–32.
33. Byrne BM. Testing for the factorial validity, replication, and invariance of a measuring instrument: a paradigmatic application based on the Maslach Burnout Inventory. Multivariate Behav Res 1994;29:289–311.
34. Jöreskog KG. Testing structural equation models. In: Bollen KA, Long JS, eds. Testing structural equation models. Newbury Park (CA): Sage Publications; 1993:294–316.
35. Prevoo ML, van 't Hof MA, Kuper HH, van Leeuwen MA, van de Putte LB, van Riel PL. Modified Disease Activity Scores that include twenty-eight–joint counts: development and validation in a prospective longitudinal study of patients with rheumatoid arthritis. Arthritis Rheum 1995;38:44–8.
36. Fries JF, Spitz P, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. Arthritis Rheum 1980;23:137–45.
37. Siegert CE, Vleming LJ, Vandenbroucke JP, Cats A. Measurement of disability in Dutch rheumatoid arthritis patients. Clin Rheumatol 1984;3:305–9.
38. Mason JH, Anderson JJ, Meenan RF. A model of health status for rheumatoid arthritis. A factor analysis of the Arthritis Impact Measurement Scales. Arthritis Rheum 1988;31:714–20.

8 A Rasch analysis of the Dutch Health Assessment Questionnaire (HAQ) Disability Index and HAQ-II in patients with rheumatoid arthritis

P.M. ten Klooster
E. Taal
M.A.F.J. van de Laar

Arthritis & Rheumatism (Arthritis Care & Research), accepted pending revision.

Abstract

Objective. The Health Assessment Questionnaire Disability Index (HAQ-DI) is the most common self-reported measure of physical disability in rheumatoid arthritis (RA). Recently, the HAQ-II was developed in the US as a short, valid, and reliable alternative using Rasch analysis. Our objective was to compare the scaling properties of the HAQ-DI and HAQ-II in Dutch patients with RA.

Methods. We used data from 472 patients with confirmed RA. Internal construct validity of the HAQ versions was assessed using Rasch analysis. Additionally, external construct validity was assessed by examining correlates with other outcome measures.

Results. The HAQ-DI had a large floor effect with 9.5% of the patients indicating no disability compared with 4.3% for the HAQ-II. Both versions were unidimensional and adequately fitted the Rasch model, containing only one non-fitting item. Additionally, two HAQ-II items demonstrated overfit and a high residual correlation, suggesting overlap or redundancy in item content. The HAQ-II had an excellent scale length, indicating that it covered a wider range of physical function. Item difficulties were reasonably well spread for the HAQ-II, whereas the HAQ-DI items tended to cluster around similar difficulty levels. Both scales contained several items with DIF by gender, age, or disease duration. Both scales demonstrated the expected pattern of correlations with other outcome measures.

Conclusion. The results indicate that both the HAQ-DI and HAQ-II are psychometrically robust measures of physical function. The Rasch-developed HAQ-II, however, has several favourable scaling properties, including a better scale length and a reduced floor effect.

Introduction

Patient assessment of physical function is one of the core measures of clinical trials and observational studies of patients with rheumatoid arthritis (RA).[1–3] Over the years, the Health Assessment Questionnaire Disability Index (HAQ-DI) has become the measure of choice for assessing self-reported disability in RA.[4,5] Although the HAQ-DI has proven to be reliable, valid, and responsive,[6,7] it is not without its limitations. Especially its reduced sensitivity in patients with lower levels of disability due to a floor effect and the nonlinear nature of the scale have been repeatedly noted.[8–13]

In an effort to overcome these problems, Wolfe et al recently developed a revised version of the HAQ-DI, the HAQ-II.[14] Using Rasch analysis on a set of 31 items, including the 20 items from the HAQ-DI, they selected those 10 items that best balanced the concerns of item fit, scale length and evenly spaced items. The resulting HAQ-II showed excellent scaling properties, a reduced floor effect, and similar convergent validity and sensitivity to change as the original HAQ-DI. The aim of the present study was to examine the construct validity of the Dutch versions of the HAQ-DI and HAQ-II in a cross-sectional sample of patients with RA.

Patient and methods

Patients and study design

The data for this study were collected at the outpatient rheumatology clinic of Medisch Spectrum Twente in Enschede, the Netherlands. During three waves of data collection between 2005 and 2007, all patients visiting the clinic were asked to complete a questionnaire consisting of demographic questions and standard self-reported measures of disease activity and health status. In total, 1363 unique patients were included during the three study periods. From this sample, all patients with confirmed RA were selected resulting in a cross-sectional sample of 472 patients. For patients with multiple visits during the study periods, data from the first visit were used in the analysis.

Measures

The HAQ-DI contains 20 items measuring physical disabilities over the past week in 8 categories of daily living: dressing and grooming, rising, eating, walking, hygiene, reach, grip, and activities.[5] Each item of the HAQ-DI is scored on a 4-point rating scale from 0 (without any difficulty) to 3 (unable to do). The overall HAQ-DI score is calculated by summing and averaging the highest item score of each category when at least 6 categories are completed, essentially reducing the HAQ-DI to an 8-item scale.[14] The overall score ranges from 0 to 3 where scores of 0 to 1 are generally considered to represent mild to moderate disability, 1 to 2 moderate to severe disability, and 2 to 3

severe to very severe disability.[6] We used the standard scoring method, which corrects for the use of devices or assistance from others.[7] The validated Dutch version of the HAQ-DI used in this study is a literal translation of the current US version with one important modification.[15] In the Dutch version, the metric weight in the item "reach and get down a 5-pound object (such as a bag of sugar) from just above your head" has been reduced to 1 kg, making the task easier to complete.[16]

The HAQ-II consists of 10 items: 5 items from the original HAQ-DI and 5 additional items. All items are scored using the same 4-point response scale as the HAQ-DI. Following the original validation study of the HAQ-II,[14] we added 11 disability items to the patient questionnaire. The additional items were literally translated into Dutch by two bilingual individuals, using a forward-backward translation procedure. The HAQ-II is scored by simply taking the mean of the items when at least 8 items are completed, also resulting in a score from 0 to 3, with higher scores indicating more disability.

Besides the HAQ-DI and the additional items for the HAQ-II, patients completed the Medical Outcomes Study 36-Item Short Form Health Survey (SF-36, version 2)[17] and numerical rating scales for pain (NRS-P) and general health (NRS-GH). The SF-36 has eight scales which can be aggregated into a physical component summary (PCS) and mental component summary (MCS) score. The scales and summary scores range from 0 to 100, with higher scores representing better health status. The component summary scores are standardized using normative data from the 1998 US general population with a mean score of 50 and standard deviation of 10. The NRS-P and NRS-GH consisted of 11-point rating scales ranging from 0 ("no pain" or "very good") to 10 ("unbearable pain" or "very bad").

Statistical analysis

Internal construct validity of the HAQ-DI and HAQ-II was assessed using Rasch analysis, while external validity was assessed by testing for expected associations with other established outcome measures in RA. Rasch analyses were performed with Winsteps version 3.60 (Winsteps.com, Chicago, IL). All other analyses were performed using SPSS version 14.0 (SPSS Inc., Chicago, IL).

The Rasch model is a one-parameter item response theory model which assumes that the probability of a certain response to a questionnaire item is a function of the person's ability on the underlying dimension being measured by the scale and the difficulty of the item.[18,19] The model asserts that the easier the item (or task) is, the more likely it will be passed and the more able a person is, the more likely he or she will pass an item compared to a less able person.[20] When data are fitted to the Rasch model, both person ability and item difficulty are calibrated in log-odd units (logits) on a common interval-level scale.

Compared with classical test theory, Rasch analysis provides a more powerful method for evaluating the internal construct validity of a scale. First, Rasch analysis is useful in testing whether the items of a scale measure a single, unidimensional construct. Moreover, it can be used to evaluate whether the items are arranged hierarchically, with sufficient spread in difficulty to measure the full range of the underlying construct. Redundant items (items that are too easy or too difficult or items with the same item difficulty) and large gaps in difficulty between items can be identified to examine the efficiency and precision of the scale. Finally, another useful attribute of Rasch analysis is that it allows for the identification of items with differential item functioning (DIF, also called item bias), i.e. items that have different levels of difficulty across subgroups of patients after controlling for overall ability. In recent years, Rasch analysis has been increasingly and successfully used in the development and evaluation of functional disability questionnaires in rheumatology.[10,13,21–28]

Since the HAQ-DI and the HAQ-II consist of polytomously scored items with ordered response categories, the unrestricted partial credit model – which does not require the distance between item thresholds to be equal across items – was applied throughout the current analyses.[29] As all Rasch models assume that the items in a scale are unidimensional and locally independent, these assumptions were thoroughly tested within the Rasch analysis process. Unidimensionality and fit of the HAQ-DI and HAQ-II to the Rasch model was firstly assessed by examination of the information-weighted mean square (INFIT MNSQ) and outlier-sensitive mean-square (OUTFIT MNSQ) fit statistics for each item. MNSQ values are the ratio between observed and predicted variance and have an expected value of 1.0. Higher values suggest that the item is "noisy" or does not measure the same underlying dimension as the other items. Lower values indicate that the item measures redundant or overlapping item content. MNSQ values between 0.7 and 1.3 were considered acceptable.[30] Additionally, a principal component analysis of the standardized residuals was performed. Once the "Rasch factor" has been extracted there should be no secondary structures (factors) left in the data. The following rules of thumb were used to confirm unidimensionality: >60% of the variance explained by the Rasch factor and an eigenvalue and explained variance of the first residual factor <3.0 and <5%, respectively.[31] Finally, residual correlations between pairs of items were examined. A relatively high residual correlation (e.g., >0.5)[32] between two items indicates that these items are not locally independent and can also point to highly overlapping or redundant items or the existence of some other shared dimension.

The efficiency and precision in measuring the underlying disability construct was examined by inspection of the item difficulty calibrations of the scales. Ideally, item difficulty levels (in logits) should be spread across a wide range of ability. Additionally, person and item separation and reliability indices were examined. The item separation

index gives an estimate of the potential range of item difficulty covered by the scale (scale length), with larger values indicating a greater spread of items. The person separation index indicates the extent to which the items can distinguish between statistically different levels of person ability. Values >2.0 were considered acceptable as this corresponds with the ability of the scale to differentiate three distinct levels of ability (e.g., high, medium, and low ability). Person reliability is an indicator of the degree to which the items measure persons in a consistent manner and is analogous to Cronbach's alpha, where values >0.7 are required for group use and >0.85 for individual patient use.[33]

Possible DIF was evaluated between subgroups of patients based on gender, age, and disease duration. Age and disease duration were split at the median to create high and low subgroups. The presence of uniform DIF was assessed using the Rasch approach implemented in Winsteps.[31] Items were considered to display substantial DIF when the difference between the separate item calibrations was statistically significant as determined by the *t*-test and the size of difference was at least 0.5 logits.[34–36]

Additionally, we assessed the agreement and convergent validity of the scales. Agreement between the Rasch-transformed scores of the HAQ-DI and HAQ-II was assessed by the Bland–Altman approach.[37] Convergent validity was tested by correlating both the raw and the Rasch-transformed scores of the scales with the SF-36 PCS and MCS, the NRS-P, and the NRS-GH. It was hypothesized that the scales should be strongly associated (r >0.6) with the SF-36 PCS and moderately (r = 0.3–0.6) with the NRS-P, NRS-GH, and SF-36 MCS.

Results

The demographic and clinical characteristics of study sample are listed in Table 1. Physical disability of the included patients was generally mild to moderate with 46% of the patients scoring <1.0 on the HAQ-DI and 41% between 1.0 and 2.0.[6]

Both the HAQ-DI and HAQ-II scores tended towards normal distributions but were slightly skewed toward lower scores (Figure 1). The HAQ-DI had a relatively large floor effect with 9.5% of the patients scoring zero (no disability) compared to 4.3% of the patients on the HAQ-II.

In general, the HAQ-DI and HAQ-II items adequately fitted the unidimensional Rasch model (Tables 2 and 3). Both scales had only one noisy item with an INFIT or OUTFIT statistic >1.3. The hygiene category of the HAQ-DI did not fit with the overall construct of functional disability, whereas "walk outdoors on a flat ground" did not fit well with the other items of the HAQ-II. Additionally, the two most difficult items of the HAQ-II ("move heavy objects" and "lift heavy objects") had OUTFIT statistics <0.7,

Table 1. Demographic and clinical characteristics of the study sample

Gender (%)	
female	69.7
male	30.3
Age (years)	
mean (SD)	59.6 (14.2)
median (IQR)	59.0 (51.0–70.0)
Disease duration (years)	
mean (SD)	10.5 (11.2)
median (IQR)	7.0 (2.0–16.0)
HAQ-DI (range 0–3)	
mean (SD)	1.1 (0.7)
median (IQR)	1.0 (0.5–1.6)
HAQ-II (range 0–3)	
mean (SD)	1.0 (0.7)
median (IQR)	1.0 (0.5–1.5)
SF-36 PCS (range 0–100)	
mean (SD)	36.8 (9.2)
median (IQR)	36.8 (30.6–43.5)
SF-36 MCS (range 0–100)	
mean (SD)	47.8 (11.9)
median (IQR)	48.7 (39.5–58.2)
NRS-P (range 0–10)	
mean (SD)	4.5 (2.8)
median (IQR)	4.0 (2.0–7.0)
NRS-GH (range 0–10)	
mean (SD)	4.2 (2.5)
median (IQR)	4.0 (2.0–6.0)

IQR = interquartile range; HAQ-DI = Health Assessment Questionnaire Disability Index; HAQ-II = Health Assessment Questionnaire II; SF-36 = Medical Outcomes Study 36-Item Short Form; PCS = physical component summary; MCS = mental component summary; NRS-P = numerical rating scale for pain; NRS-GH = numerical rating scale for general health.

suggesting overfit or overlapping item content. The principal component analyses of the standardized residuals confirmed the unidimensionality of both scales. For the HAQ-DI, 62.9% of the variance was explained by the Rasch dimension, whereas 7.4% of the unexplained variance was accounted for by the first residual factor with an eigenvalue of 1.6. The Rasch dimension in the HAQ-II accounted for 74.8% of the variance and the first residual factor, with an eigenvalue of 2.4, explained only 6% of the variance. Finally, inter-item residual correlations were generally low for both measures. All residual correlations in the HAQ-DI were below 0.30. The HAQ-II, however, contained a high residual correlation of 0.59 between the items "move heavy objects" and "lift heavy objects". Inter-item residual correlations were low for the other items of the HAQ-II (r's <0.35).

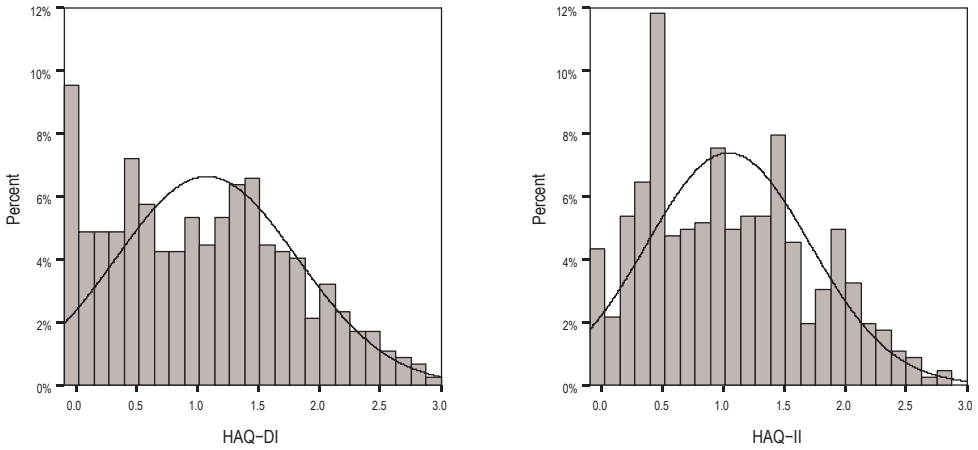

Figure 1. Distribution of the total HAQ-DI and HAQ-II scores. HAQ-DI: Skewness 0.32, Kurtosis –0.80; HAQ-II: Skewness 0.39, Kurtosis –0.75.

Table 2. Item difficulties and fit statistics of the HAQ-DI items ordered by difficulty level

	Item difficulty (logits)*	SE	INFIT MNSQ†	OUTFIT MNSQ†
Rising	2.00	0.09	1.01	0.94
Walking	0.47	0.08	1.06	1.07
Dressing and grooming	0.34	0.08	0.86	0.84
Reach	0.21	0.08	0.87	0.82
Eating	–0.03	0.08	0.96	0.98
Grip	–0.80	0.08	1.19	1.26
Activities	–1.03	0.08	0.89	0.88
Hygiene	–1.17	0.07	1.23	**1.37**

SE = standard error; MNSQ = mean square; INFIT = information-weighted fit; OUTFIT = outlier-sensitive fit.
Person separation index 2.49, Person reliability 0.86, Item separation index 11.45.
* More negative scores indicate more difficult items.
† MNSQ values >1.30 (noisy items or items not measuring the underlying construct) are shown in bold; No MNSQ values <0.70 (overlapping or redundant items).

The HAQ-II had an excellent scale length with an item separation index of 21.63 compared to 11.45 for the HAQ-DI, indicating that the HAQ-II covered a much wider range of the functional disability construct. Inspection of the item difficulty calibrations showed that the items of the HAQ-II were reasonably well spread across a wide range of difficulty. Besides its relatively limited scale length, the items of the HAQ-DI tended to cluster around similar difficulty levels around the middle of the scale hierarchy, with

Table 3. Item difficulties and fit statistics of the HAQ-II items ordered by difficulty level

	Item difficulty (logits)*	SE	INFIT MNSQ†	OUTFIT MNSQ†
Get on and off the toilet?	2.89	0.12	1.02	0.88
Stand up from a straight chair?	2.33	0.10	1.10	1.19
Open car doors?	1.87	0.10	0.96	1.08
Walk outdoors on flat ground?	1.28	0.09	1.18	**1.70**
Reach and get down a 1-kg object (such as a bag of sugar) from just above your head?	0.71	0.09	1.07	1.12
Go up 2 or more flights of stairs?	–0.63	0.08	1.07	1.11
Wait in a line for 15 minutes?	–0.86	0.08	1.02	1.20
Do outside work (such as yard work)?	–1.56	0.08	0.95	0.91
Move heavy objects?	–2.92	0.08	0.74	*0.66*
Lift heavy objects?	–3.11	0.08	0.77	*0.64*

See Table 2 for abbreviations.

Person separation index 2.74, Person reliability 0.88, Item separation index 21.63.

* More negative scores indicate more difficult items.

† MNSQ values >1.30 (noisy items or items not measuring the underlying construct) are shown in bold; MNSQ values <0.70 (overlapping or redundant items) are shown in italics.

relatively few items at the extremes. Both scales had person separation and reliability indices >2.0 and >0.85, respectively, indicating that both can adequately discriminate between levels of physical disability and are sufficiently reliable for individual patient use.

Three items of the HAQ-DI exhibited substantial uniform DIF between subgroups of patients (Figure 2). After controlling for overall disability, women had less difficulty with dressing (DIF contrast = 0.93, $P < 0.001$), but more difficulty with grip (DIF contrast = 0.75, $P < 0.001$). Hygiene was less difficult for younger patients (DIF contrast = 0.86, $P < 0.001$) and patients with shorter disease duration (DIF contrast = 0.78, $P < 0.001$). The HAQ-II also had three items with DIF. Standing up from a straight chair was more difficult for men (DIF contrast = 0.62, $P = 0.007$), younger patients (DIF contrast = 0.55, $P = 0.008$), and patients with shorter disease duration (DIF contrast = 0.75, $P < 0.001$). Getting on and of the toilet was more difficult for younger patients (DIF contrast = 0.75, $P = 0.001$) and patients with shorter disease duration (DIF contrast = 0.54, $P < 0.020$). Finally, opening car doors was more difficult for younger patients (DIF contrast = 0.53, $P = 0.010$).

The raw and Rasch-transformed HAQ-DI and HAQ-II scores were highly intercorrelated (Table 4) and the absolute difference in raw mean scores was only 0.04 units. Additional Bland–Altman analysis of the Rasch-transformed scores showed that the mean HAQ-II scores were systematically biased towards worse scores on the HAQ-II

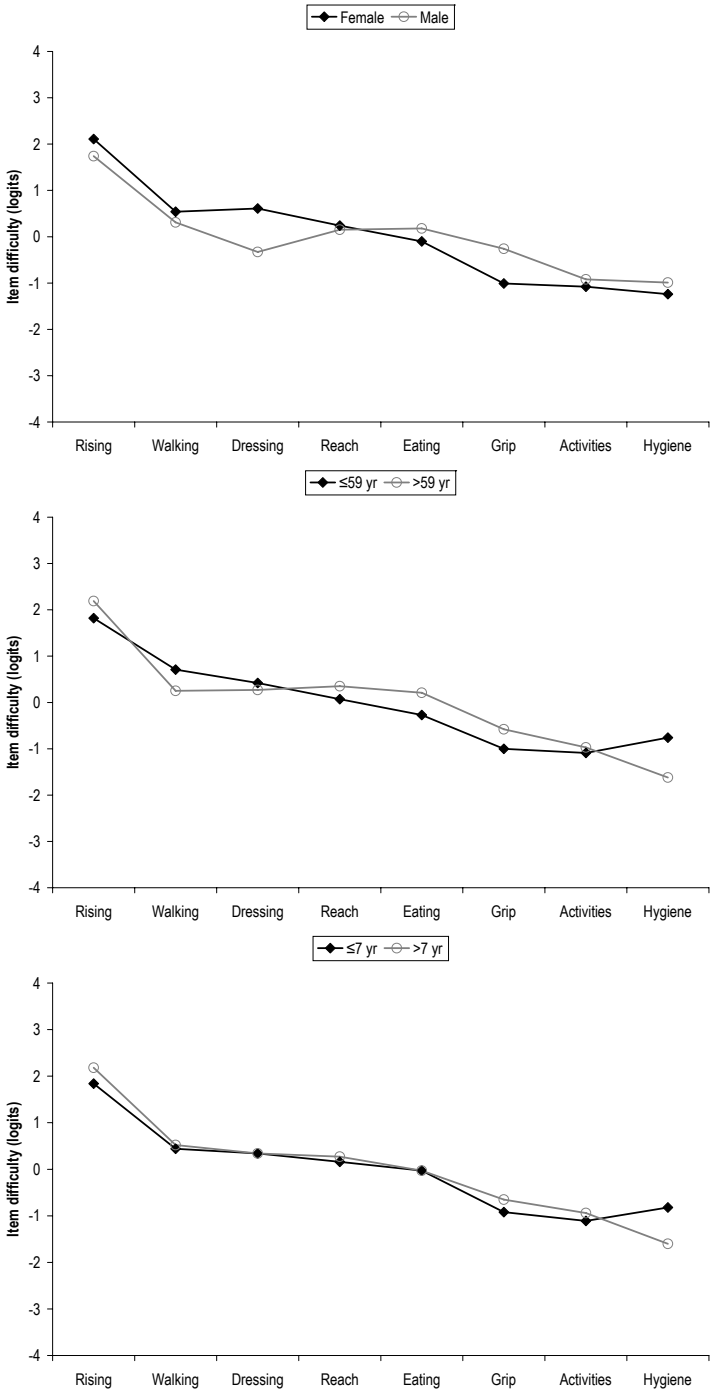

Figure 2. Differential item functioning plots of the HAQ-DI (left) and HAQ-II (right) between patient groups based on gender (top), age (middle), and disease duration (top). Age and gender are split at the median.

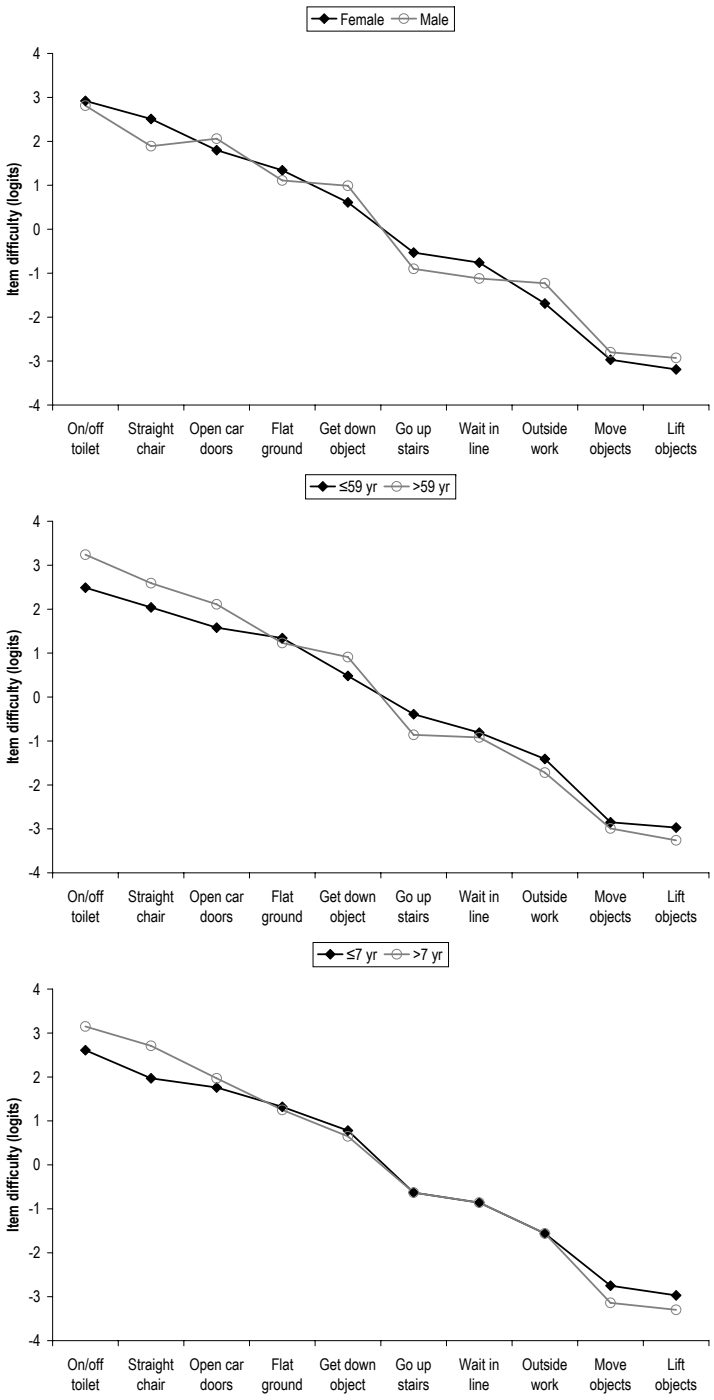

Figure 2. (Continued)

(mean difference 0.216 logits, paired *t*-test, $P < 0.001$) with the 95% limits of agreement ranging from –2.156 to 2.588. Both scales demonstrated the expected pattern of correlations with other outcome measures, where HAQ-II correlates tended to be slightly stronger (Table 4).

Table 4. Pearson intercorrelations between self-reported outcome measures

	HAQ-DI	HAQ-II	SF-36 PCS	SF-36 MCS	NRS-P	NRS-GH
HAQ-DI	1.00					
HAQ-II	0.92 (0.89)	1.00				
SF-36 PCS	–0.65 (–0.65)	–0.71 (–0.71)	1.00			
SF-36 MCS	–0.32 (–0.32)	–0.34 (–0.34)	0.09	1.00		
NRS-P	0.46 (0.46)	0.46 (0.47)	–0.59	–0.21	1.00	
NRS-GH	0.40 (0.41)	0.41 (0.44)	–0.54	–0.28	0.69	1.00

See Table 1 for abbreviations.
Correlations with the Rasch-transformed scores of the HAQ versions are presented in parentheses.

Discussion

This study used Rasch analysis to examine the construct validity of the HAQ-DI and the HAQ-II in patients with RA. The results suggest that both scales are psychometrically robust measures of physical disability. Compared with the HAQ-DI, the HAQ-II has favourable scaling properties, as demonstrated by a better scale length and a reduced floor effect. The results further point to some improvements that could be made with respect to misfitting or redundant items and items with DIF which were present in both scales.

In general, both the HAQ-DI and the HAQ-II showed an acceptable fit to the Rasch model as judged by the information-weighted mean square INFIT and OUTFIT statistics. Following previous studies examining the HAQ using Rasch analysis,[10,13,14] we applied the common critical range of 0.7 to 1.3 for reasonable fit. However, it should be noted that these statistics are sensitive to sample size and may lead to an unacceptable Type I error rate with large samples.[38] In the HAQ-DI, only the hygiene category demonstrated substantial underfit as indicated by a relatively high OUTFIT value. The finding that this category is noisy or does not measure the same dimension as the other categories is in accordance with previous Rasch analyses of the scale.[13,14,25] Wolfe et al[14] have suggested that this noisiness may be caused by patients guessing at their ability to answer the underlying item "take a tub bath", since many people use showers instead of bathtubs. The HAQ-II also contained one item with high OUTFIT. The finding that "walk outdoors on a flat ground" did not closely relate to the overall construct of disability is somewhat surprising and has not been reported in previous studies. Given that the corresponding INFIT statistic, which is less sensitive to unexpected responses to items far from a person's level of ability, is acceptable, the misfit of this item may be inflated by a few unexpected responses of patients with high disability.

Besides this underfitting item, the HAQ-II additionally contained two items with low INFIT and OUTFIT statistics, which usually indicates overlap or redundancy in the pattern of responses. This overlap between "move heavy objects" and "lift heavy objects" was also apparent by an unacceptably high residual correlation between these items. Indeed, simple inspection of the item content does suggest that the items assess very similar and interdependent tasks as people will usually try to lift objects in order to move them. Omission of one of these items or, even better, replacement with a slightly less difficult item in future studies could lead to better overall scaling properties of the HAQ-II.

As would be expected from a Rasch-developed measure, the HAQ-II demonstrated better distributional and scaling properties than the traditional HAQ-DI. Total HAQ-II scores showed a substantially lower floor effect than total HAQ-DI scores. The floor effect of the HAQ-DI, in which patients report a normal score but nonetheless experi-

ence functional limitations, is a well-known problem of the HAQ-DI[8–11,13,14] and was one of the main reasons for the development of the HAQ-II. This smaller floor effect is achieved by a better scale length and item difficulty calibration of the HAQ-II. The HAQ-II measures a wider ranger of disability and has specifically more items probing relatively difficult activities. In fact, according to the current results, the 5 most difficult items in the HAQ-II were the ones that were added to the scale by Wolfe et al. This resulted in a high item separation index for the HAQ-II, which was almost twice the size of that for the HAQ-DI.

The present study design did not allow for a direct examination of the responsiveness of the scales. However, the high person separation and reliability indices indicate that both scales can identify several statistically distinct levels of person ability. This lends support to the sensitivity of both scales to changes in physical disability, where the HAQ-II would theoretically be somewhat more sensitive.

One concern with both scales is the presence of items with DIF between patient groups. Especially older and younger patients, who are at the same level of disability, appear to respond somewhat differently to several items. Although the absolute magnitude of DIF was generally small and may average out across the items in a scale, future studies should continue to examine the presence of DIF and its influence on the total scale scores.

Although the actual item difficulty estimates of the HAQ-DI in this study were somewhat different from those reported in previous studies in RA,[10,13,25] the difficulty hierarchy (rank order) was quite similar to the ones most recently found by Wolfe et al[25] and Taylor and McPherson.[13] Also, both the actual difficulty estimates and the difficulty hierarchy of the HAQ-II items in this study were very similar to those of the original US version.[14] This finding provides some preliminary support for both the comparability of the scales across different RA cohorts and the robustness of the present findings. Future research, using Rasch analysis on pooled data from different countries and cohorts, could assess more thoroughly whether the scales are equivalent across cultures and different cohorts.

An important, and possibly related issue, remains the divergent translation of the item "reach and get down a 5-pound object (such as a bag of sugar) from just above your head" in the Dutch HAQ-DI and HAQ-II. In the versions we used in this study, this item is made "easier" by reducing the object weight to 1 kg. In the Rasch analysis, this was reflected in substantially lower item difficulty estimates for this item of the HAQ-II and the corresponding reach category of the HAQ-DI compared with previous studies using the original wording.[10,13,14,25] This difference in item difficulty should be kept in mind when comparing the present results with previous (Rasch) analyses of the HAQ-DI and HAQ-II. Recently, a new consensus version of the Dutch HAQ-DI was proposed which should overcome this problem.[16]

The present findings should also be considered within the context of the strict assumptions of the Rasch model. Although the one-parameter Rasch model is probably the most commonly used item response theory model for analyzing (functional disability) questionnaires, it may not always be the most suitable model. For instance, a recent study of a similar measure of functional disability showed that a two-parameter model fitted the data significantly better than the one-parameter model.[39] Future studies could examine whether applying more general models will lead to different results.

Finally, the HAQ-DI and HAQ-II were highly intercorrelated and demonstrated a similar pattern of associations with other validated outcomes, suggesting that both scales assess the same underlying construct. Off course, this is not very surprising since 5 out of the 10 HAQ-II items stem directly from the HAQ-DI. Although the absolute mean difference between the raw HAQ-DI and HAQ-II scores was negligible, Bland–Altman analysis of the Rasch-transformed scores showed that the HAQ-DI and HAQ-II scores were significantly different and characterized by high intra-individual variation. Wolfe et al[14] have suggested conversion formulae for transforming group-level data from HAQ-DI to HAQ-II and vice versa. The current results, however, confirm their finding that the HAQ-DI and HAQ-II cannot be used interchangeably at the individual patient level.

In conclusion, this study suggests that the HAQ-DI and HAQ-II are both adequately valid measures of physical disability in patients with RA, but confirm that the Rasch-developed HAQ-II has better distributional and scaling properties. Moreover, given that the HAQ-II is much shorter – particularly when the aids and devices section of the HAQ-DI are considered – and easier to score, the HAQ-II appears to be more suitable for use in clinical care.

Acknowledgements

The authors thank Christine Bellmann, Ilse Bosgra, Petra Hagens, Nicolette Kupper, Julia Rulle, Lisanne Schmit, Amrah Schotanus, Johan Steehouder, Katharine Steentjes, Lidewij van Gessel, and Anouk van der Heij for collecting the data and Christina Bode and Andre Brands for their help in organizing the study.

References

1. Boers M, Tugwell P, Felson DT, et al. World Health Organization and International League of Associations for Rheumatology core endpoints for symptom modifying antirheumatic drugs in rheumatoid arthritis clinical trials. J Rheumatol 1994;21:86–9.
2. Felson DT, Anderson JJ, Boers M, et al. The American College of Rheumatology preliminary core set of disease activity measures for rheumatoid arthritis clinical trials. Arthritis Rheum 1993;36:729–40.

3. Wolfe F, Lassere M, van der Heijde D, et al. Preliminary core set of domains and reporting requirements for longitudinal observational studies in rheumatology. J Rheumatol 1999;26:484–9.
4. Fries JF, Spitz P, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. Arthritis Rheum 1980;23:137–45.
5. Fries JF, Spitz PW, Young DY. The dimensions of health outcomes: the Health Assessment Questionnaire, disability and pain scales. J Rheumatol 1982;9:789–93.
6. Bruce B, Fries JF. The Stanford Health Assessment Questionnaire: Dimensions and Practical Applications. Health Qual Life Outcomes 2003;1:20.
7. Bruce B, Fries JF. The Stanford Health Assessment Questionnaire: a review of its history, issues, progress, and documentation. J Rheumatol 2003;30:167–78.
8. Stucki G, Stucki S, Bruhlmann P, Michel BA. Ceiling effects of the Health Assessment Questionnaire and its modified version in some ambulatory rheumatoid arthritis patients. Ann Rheum Dis 1995;54:461–5.
9. Pincus T, Swearingen C, Wolfe F. Toward a multidimensional Health Assessment Questionnaire (MDHAQ): assessment of advanced activities of daily living and psychological status in the patient-friendly health assessment questionnaire format. Arthritis Rheum 1999;42:2220–30.
10. Tennant A, Hillman M, Fear J, Pickering A, Chamberlain MA. Are we making the most of the Stanford Health Assessment Questionnaire? Br J Rheumatol 1996;35:574–8.
11. Uhlig T, Haavardsholm EA, Kvien TK. Comparison of the Health Assessment Questionnaire (HAQ) and the modified HAQ (MHAQ) in patients with rheumatoid arthritis. Rheumatology (Oxford) 2006;45:454–8.
12. Wolfe F. The psychometrics of functional status questionnaires: room for improvement. J Rheumatol 2002;29:865–8.
13. Taylor WJ, McPherson KM. Using Rasch analysis to compare the psychometric properties of the Short Form 36 physical function score and the Health Assessment Questionnaire disability index in patients with psoriatic arthritis and rheumatoid arthritis. Arthritis Rheum 2007;57:723–9.
14. Wolfe F, Michaud K, Pincus T. Development and validation of the Health Assessment Questionnaire II: a revised version of the Health Assessment Questionnaire. Arthritis Rheum 2004;50:3296–305.
15. Zandbelt MM, Welsing PM, van Gestel AM, van Riel PL. Health Assessment Questionnaire modifications: is standardisation needed? Ann Rheum Dis 2001;60:841–5.
16. Boers M, Jacobs JW, van Vliet Vlieland TP, van Riel PL. Consensus Dutch Health Assessment Questionnaire. Ann Rheum Dis 2007;66:132–3.
17. Ware JE, Kosinski M, Dewey JE. How to score version 2 of the SF-36 Health Survey. Lincoln, RI: QualityMetric Incorporated; 2000.
18. Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. Med Care 2000;38:II28–42.
19. Reeve BB, Fayers P. Applying item response theory modelling for evaluating questionnaire item and scale properties. In: Fayers PM, Hays RD, eds. Assessing quality of life in clinical trials: Methods and practice. Oxford: Oxford University Press; 2005:55–73.

20. Tennant A, McKenna SP, Hagell P. Application of Rasch analysis in the development and application of quality of life instruments. Value Health 2004;7 Suppl 1:S22–6.
21. Wolfe F, Kong SX. Rasch analysis of the Western Ontario MacMaster questionnaire (WOMAC) in 2205 patients with osteoarthritis, rheumatoid arthritis, and fibromyalgia. Ann Rheum Dis 1999;58:563–8.
22. Roorda LD, Jones CA, Waltz M, et al. Satisfactory cross cultural equivalence of the Dutch WOMAC in patients with hip osteoarthritis waiting for arthroplasty. Ann Rheum Dis 2004;63:36–42.
23. Ryser L, Wright BD, Aeschlimann A, Mariacher-Gehler S, Stucki G. A new look at the Western Ontario and McMaster Universities Osteoarthritis Index using Rasch analysis. Arthritis Care Res 1999;12:331–5.
24. Wolfe F, Hawley DJ, Goldenberg DL, Russell IJ, Buskila D, Neumann L. The assessment of functional impairment in fibromyalgia (FM): Rasch analyses of 5 functional scales and the development of the FM Health Assessment Questionnaire. J Rheumatol 2000;27:1989–99.
25. Wolfe F. Which HAQ is best? A comparison of the HAQ, MHAQ and RA-HAQ, a difficult 8 item HAQ (DHAQ), and a rescored 20 item HAQ (HAQ20): analyses in 2,491 rheumatoid arthritis patients following leflunomide initiation. J Rheumatol 2001;28:982–9.
26. Pouchot J, Ecosse E, Coste J, Guillemin F. Validity of the Childhood Health Assessment Questionnaire is independent of age in juvenile idiopathic arthritis. Arthritis Rheum 2004;51:519–26.
27. Kucukdeveci AA, Sahin H, Ataman S, Griffiths B, Tennant A. Issues in cross-cultural validity: example from the adaptation, reliability, and validity testing of a Turkish version of the Stanford Health Assessment Questionnaire. Arthritis Rheum 2004;51:14–9.
28. Durez P, Fraselle V, Houssiau F, Thonnard JL, Nielens H, Penta M. Validation of the ABILHAND questionnaire as a measure of manual ability in patients with rheumatoid arthritis. Ann Rheum Dis 2007;66:1098–105.
29. Masters G. A Rasch model for partial credit scoring. Psychometrika 1982;47:149–74.
30. Wright BD, Linacre JM, Gustafson JE, Martin-Lof P. Reasonable mean-square fit values. Rasch Measurement Transactions 1994;8:370.
31. Linacre JM. A user's guide to WINSTEPS MINISTEP Rasch-model computer programs. Chicago IL: Winsteps.com; 2006.
32. Davidson M, Keating JL, Eyres S. A low back-specific version of the SF-36 Physical Functioning scale. Spine 2004;29:586–94.
33. Tennant A, Conaghan PG. The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a Rasch paper? Arthritis Rheum 2007;57:1358–62.
34. Draba RE. The identification and interpretation of item bias (Research memorandum No. 25). Chicago, IL: University of Chicago; 1977.
35. Lai JS, Teresi J, Gershon R. Procedures for the analysis of differential item functioning (DIF) for small sample sizes. Eval Health Prof 2005;28:283–94.
36. Linacre M. Sample size and item calibration stability. Rasch Measurement Transactions 1994;7:328.

37. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;i:307–10.
38. Smith RM, Schumacker RE, Bush MJ. Using item mean squares to evaluate fit to the Rasch model. J Outcome Meas 1998;2:66–78.
39. Holman R, Lindeboom R, Vermeulen M, de Haan RJ. The AMC Linear Disability Score project in a population requiring residential care: psychometric properties. Health Qual Life Outcomes 2004;2:42.

9 Summary and discussion

Summary

Patient-reported outcomes (PROs) have taken a central role in the evaluation of disease status and treatment effects in the rheumatic diseases. The aim of this thesis was to explore several contemporary issues related to the patient's perspective and the use of modern psychometric analysis of PROs in rheumatology.

Chapter 2 of the thesis presents a study of the 1-year course of patients' priorities for improvement in health in 226 patients with active RA starting treatment with tumor necrosis factor blocking agents. The aims of the study were to identify those aspects of health that patients with RA themselves would most like to see improved and to examine whether these priorities for improvement were sensitive to changes in health status. The results of the study showed that, upon entering the cohort, patients most commonly selected pain, hand and finger function, walking and bending, mobility, and work as priorities for improvement. Despite major improvements in almost all of these aspects of health during treatment, this priority ranking remained mostly unchanged at the group level. Pain was the only priority area that was selected significantly less often after 3 and 12 months of treatment, but it remained the most often selected priority. Within individual patients, however, priorities for improvement frequently changed. Since changes in the priority of pain were related to the achieved level of patient-perceived pain and disease activity, it appears that individual patient priorities are not stable over time and can change as a result of effective treatment. This finding gives some preliminary support to the idea that the importance of particular outcomes to individual patients varies as a result of effective treatment and that existing measures may be enhanced by taking account of these variations in priorities.

Chapter 3 describes the development and evaluation of a new concept for measuring meaningful change in pain on a visual analog scale (VAS) from the individual patient's perspective. The concept of "patient-perceived satisfactory improvement" (PPSI) was constructed as a 5-point categorical rating of change scale and applied in a 2-week prospective study of 181 arthritis patients treated with local corticosteroid injections, a treatment with known efficacy. The optimal cut-off for PPSI on the VAS was a minimal reduction of 30 mm or 55%. The results also showed that the absolute change in pain associated with a satisfactory improvement was highly dependent on baseline pain. As a result, percent change scores performed significantly better in classifying satisfactory improved patients. The cut-off for PPSI was consistent over the course of treatment and reasonably consistent across different diagnostic groups. Overall, this study suggests that PPSI is a clinically relevant and stable concept for interpreting truly meaningful improvements in pain from the individual perspective.

Chapter 4 focuses on a more general concern with PROs, namely the validity of patient's retrospective reports of symptoms or health. The aim of the study in this chapter was to study the agreement between patients' actual baseline assessments of pain and global health on a VAS before treatment and retrospective assessments collected 2 weeks after treatment. Data were used from the study as described in chapter 3. The results of this study showed that 2 weeks after treatment with a local injection, patients slightly overestimated the severity of pain and global health before treatment. Although actual and retrospective assessments were adequately correlated and fairly accurate at the group level, there was unacceptable intra-individual variation between actual and retrospective assessments. The study concludes that even over relatively short time intervals, retrospective assessments should not be used as substitutes for individual baseline status or to measure individual changes over treatment in clinical trials.

Chapters 5 and 6 describe the results of a cross-cultural study of rheumatic pain intensity in patients from an Arabic and Western culture. Participants were 42 young Egyptian women with RA and 30 Dutch women with RA, matched for age and disease duration. **Chapter 5** compares the validity and reliability of a graphic rating scale (GRS) and a verbal rating scale (VRS) for measuring pain between both countries. The study showed that both the GRS and VRS were reliable and valid in the total study cohort. Within the individual countries, however, the GRS seemed to perform better than the VRS. In **chapter 6**, ethnocultural differences in pain intensity reports between the Egyptian and Dutch patients and the influence of possible confounding variables are examined. The analysis showed that, although the progression of RA and radiographic damage were not significantly different between both groups, the Egyptian population reported significantly worse pain and physical function and demonstrated higher disease activity. After controlling for the differences in disease activity and socioeconomic and clinical variables, country of residence remained a significant independent predictor of pain intensity. The study confirms that there are ethnocultural differences in the pain reports between Egyptian and Dutch women with RA and indicates the need for more studies to explore the mechanisms that may underlie these differences.

In **chapter 7**, the factorial validity of the short form Arthritis Impact Measurement Scales 2 (AIMS2-SF) is evaluated using structural equation modeling. Three a-priori defined measurement models, based on previous exploratory factor analyses, were tested and compared in a sample of 279 patients with active RA who completed the long form AIMS2 before starting treatment with TNF-blocking agents. The analyses showed that both the currently applied measurement model and an alternative model

with the physical dimension divided into upper and lower body limitations adequately fitted the data. The latter model, however, performed significantly better. The study concludes that the AIMS2-SF has acceptable factorial validity and suggests the use of separate scores for upper and lower body limitations. The results also point to several specific problems associated with the content or wording of some of the items that need further study.

Finally, **chapter 8** presents a Rasch analysis, a basic form of item response theory modeling, of the Health Assessment Questionnaire Disability Index (HAQ-DI) and a more recent revised version (HAQ-II) in a cross-sectional sample of 472 patients with confirmed RA. Results of this study showed that the HAQ-DI and the HAQ-II adequately fitted the Rasch model, with both scales containing one item that did not appear to measure the underlying dimension of physical disability. Additionally, two HAQ-II items demonstrated overfit and a high residual correlation, suggesting overlap or redundancy in item content. Both scales contained several items with differential item functioning by gender, age, or disease duration. Compared with the HAQ-DI, the HAQ-II had a smaller floor effect and a much longer scale length. Overall, the study suggests that both the HAQ-DI and the HAQ-II are valid, unidimensional measures of physical disability. The Rasch-developed HAQ-II, however, has several favorable scaling properties which favor its use over the traditional HAQ-DI.

General discussion and future directions

The patient's perspective and the use of modern psychometrics to evaluate and improve PROs have found their way into outcome assessment in the rheumatic diseases. Both have shown their merits in improving the quality of health research, especially when used in combination, and both will undoubtedly receive even more attention in the near future. However, the basic nature and assumptions of both paradigms are very different and may even conflict with one another.

The patient's perspective focuses on improving health outcomes research by advocating more use of the knowledge, values, and experiences of patients with the disease themselves in research. This concept is very broad and has expanded from merely having patients rate their own health status or symptoms using PROs to diverse issues such as identifying patients' views on meaningful outcomes and improvements in outcomes and including patients as research partners in all stages of a research project. As such, this paradigm is mainly concerned with the face validity of outcome measures, i.e. do they measure what is important to patients in a meaningful manner? Psychometrics in general, and modern psychometrics in particular, focuses on the statistical properties of items and on maximizing the precision and efficiency of out-

come measures. As a result, it is very well possible that the increasing use of modern psychometrics will lead to a loss of face validity to the patients. For instance, IRT analysis can result in the omission of items that are important from the patient's perspective, but that do not meet certain statistical characteristics (e.g., items with poor fit to a specific IRT model).

As with classical psychometrics, modern psychometrics is not the sole solution in developing and evaluating good PROs. The challenge for clinicians and researchers is to increase the measurement precision and efficiency of PROs using sophisticated statistical methods, without losing their face validity to individual patients. Qualitative techniques and continuous input from patients themselves will therefore remain essential tools to ensure the validity of PROs now and in the future.

Finally, two promising developments in the light of PROs and modern psychometrics are the introduction of more elaborate IRT models and computerized adaptive tests in health research. The use of 2- and 3-parameter and multidimensional IRT models may prove to be more appropriate for analyzing health-related domains due to their more realistic assumptions about the nature of the underlying dimensions. Computerized adaptive tests, where each patient is administered a unique set of items from a large IRT-based item bank tailored to his or her specific situation, will make it possible to directly compare test scores across individuals, countries, and diseases.

Samenvatting

In de klinische praktijk en bij onderzoek naar het ziekteverloop en de effectiviteit van behandelingen bij patiënten met een reumatische aandoening wordt in toenemende mate gebruik gemaakt van gestandaardiseerde vragenlijsten die door de patiënt zelf worden ingevuld. In dit proefschrift worden verschillende vraagstukken verkend die voortkomen uit twee paradigmaverschuivingen in het huidige onderzoek naar deze patiënt-gerapporteerde uitkomstmaten (*Patient reported outcomes*, PROs): de toenemende aandacht voor het in kaart brengen van het individuele patiëntenperspectief met behulp van PROs en de toepassing van moderne psychometrische technieken voor het ontwikkelen en evalueren van PROs.

In **hoofdstuk 2** wordt een onderzoek beschreven naar het verloop van prioriteiten voor gezondheidsverbetering bij 226 patiënten met actieve reumatoïde artritis (RA) gedurende 1 jaar behandeling met tumor necrose factor α blokkerende geneesmiddelen. Het doel van deze studie was het identificeren van díe gezondheidsaspecten die patiënten met RA zelf het liefst verbeterd zouden zien en het onderzoeken in hoeverre deze prioriteiten gevoelig zijn voor veranderingen in gezondheid. De resultaten van dit onderzoek lieten zien dat patiënten bij aanvang van de behandeling het liefst verbeteringen zouden zien in pijn, hand- en vingerfuncties, lopen en buigen, bewegingsmogelijkheden en werk. Ondanks sterke verbeteringen in bijna alle aspecten van gezondheid, bleef deze rangschikking van prioriteiten op groepsniveau grotendeels onveranderd gedurende de behandeling. Alleen pijn werd significant minder vaak als prioriteit voor verbetering gekozen na 3 en 12 maanden behandeling, maar bleef wel de meest gekozen prioriteit. Bij individuele patiënten veranderde de prioriteitstelling echter regelmatig. De bevinding dat veranderingen in het al dan niet selecteren van pijn als prioriteit voor verbetering geassocieerd waren met het bereikte niveau van pijn en ziekteactiviteit, suggereert dat individuele prioriteiten van patiënten niet stabiel zijn en kunnen veranderen als gevolg van effectieve behandeling. Dit biedt enige ondersteuning voor de veronderstelling dat het belang van specifieke uitkomsten voor individuele patiënten varieert als gevolg van effectieve behandeling en dat bestaande meetinstrumenten verbeterd kunnen worden door rekening te houden met deze veranderingen in prioriteiten.

Hoofdstuk 3 beschrijft de ontwikkeling en evaluatie van een nieuw concept voor het meten van relevante verbeteringen in pijn op een visueel analoge schaal (VAS) vanuit het perspectief van de individuele patiënt. Het concept van voldoende verbetering (*patient-perceived satisfactory improvement*, PPSI) bestond uit een 5-punts veranderingsschaal en werd toegepast in een prospectief onderzoek van 2 weken onder 181 reumapatiënten die behandeld werden met lokale injecties met corticosteroïden, een behandeling met bewezen effectiviteit. De optimale afkapwaarde voor PPSI op de VAS was een absolute afname van 30 mm of een relatieve afname van 50%. Uit het onderzoek kwam bovendien naar voren dat de absolute afname in pijn die nodig was voor een voldoende verbetering sterk afhankelijk was van de uitgangswaarde op de VAS. Hierdoor presteerden relatieve veranderingen significant beter in het classificeren van voldoende en onvoldoende verbeterde patiënten. De afkapwaarden voor voldoende verbetering bleken consistent over tijd en redelijk consistent over groepen met verschillende diagnoses. Geconcludeerd wordt dat PPSI een klinisch relevant en stabiel concept is voor het interpreteren van relevante verbeteringen in pijn vanuit het individuele patiëntenperspectief.

Hoofdstuk 4 richt zich op een meer algemeen vraagstuk met betrekking tot patiëntgerapporteerde uitkomstmaten, namelijk de validiteit van het retrospectief meten van symptomen en gezondheid. Het doel van dit onderzoek was het bepalen van de overeenstemming tussen daadwerkelijke beoordelingen van pijn en algemene gezondheid op een VAS door patiënten bij aanvang van een behandeling en retrospectieve beoordelingen 2 weken na de behandeling. Voor dit onderzoek werden data gebruikt uit de studie zoals beschreven in hoofdstuk 3. Twee weken na de behandeling bleken patiënten geneigd de ernst van hun pijn en gezondheid voor de behandeling enigszins te overschatten. Hoewel de daadwerkelijke en retrospectieve *baseline* metingen sterk gecorreleerd waren en retrospectieve metingen op groepsniveau redelijk nauwkeurig bleken, was er sprake van een onacceptabel hoge intra-individuele variatie tussen de daadwerkelijke en retrospectieve metingen. De conclusie van deze studie is dat, zelfs over relatief korte tijdsperiodes, retrospectieve metingen op individueel niveau niet gebruikt kunnen worden ter vervanging van daadwerkelijke baseline metingen of voor het meten van individuele veranderingen in klinische onderzoeken.

De hoofdstukken 5 en 6 beschrijven de resultaten van een crosscultureel onderzoek naar reumatische pijn bij patiënten uit een Arabische en een westerse cultuur. Deelnemers aan dit onderzoek waren 42 jonge Egyptische vrouwen met RA en 30 Nederlandse vrouwen met RA, die gematched waren op leeftijd en ziekteduur. **Hoofdstuk 5** vergelijkt de validiteit en betrouwbaarheid van de gebruikte grafische schaal (*graphic rating scale*, GRS) en verbale schaal (*verbal rating scale*, VRS) voor het meten van pijn in

beide landen. Deze studie liet zien dat zowel de GRS als de VRS betrouwbaar en valide waren in de totale studiegroep. Binnen de individuele landen bleek de GRS echter beter te presteren. In **hoofdstuk 6** wordt gekeken naar verschillen in pijnintensiteit tussen de Egyptische en Nederlandse patiënten en de invloed van mogelijke achterliggende factoren op deze verschillen. Hoewel er geen verschillen waren in ziekteprogressie en radiografisch waarneembare gewrichtsschade tussen beide patiëntgroepen, rapporteerden de Egyptische patiënten significant meer pijn en fysieke beperkingen en hadden ze een hogere ziekteactiviteit. Na correctie voor de verschillen in ziekteactiviteit en socio-economische en klinische variabelen, bleef het land van verblijf een significante, onafhankelijke voorspeller van pijn intensiteit. De studie bevestigt dat er etno-culturele verschillen zijn tussen Egyptische en Nederlandse vrouwen met RA en wijst op de behoefte aan meer onderzoek naar mogelijke onderliggende mechanismen van deze verschillen.

In **hoofdstuk 7** wordt de factoriële validiteit van de verkorte versie van de Arthritis Impact Measurement Scales 2 (AIMS2-SF) geëvalueerd met behulp van structurele vergelijkingsmodellen. Drie vooraf vastgestelde meetmodellen, gebaseerd op voorgaande exploratieve factoranalyses, werden getoetst en vergeleken in een groep van 279 patiënten met actieve RA die de AIMS2 invulden voorafgaande aan hun eerste behandeling met tumor necrose factor α blokkerende geneesmiddelen. De analyses lieten zien dat zowel het huidige meetmodel als een alternatief meetmodel waarin de fysieke dimensie was opgesplitst in bovenste en onderste extremiteiten voldoende *fit* vertoonden met de data. Het alternatieve model presteerde echter significant beter. De studie concludeert dat factoriële validiteit van de AIMS2-SF acceptabel is en stelt voor om aparte scores te gebruiken voor beperkingen van het boven- en onderlichaam. De resultaten wijzen ook op verschillende specifieke problemen met betrekking tot de inhoud of verwoording van sommige items die verder onderzoek behoeven.

Hoofdstuk 8 beschrijft tenslotte een Rasch analyse, een basisvorm van item response theorie, van de Health Assessment Questionnaire Disability Index (HAQ-DI) en een meer recente versie (de HAQ-II) in een cross-sectionele groep van 472 patiënten met RA. De resultaten van deze studie lieten zien dat de HAQ-DI en de HAQ-II voldoende fit vertoonden met het Rasch model, waarbij beide vragenlijsten 1 item bleken te bevatten dat niet paste bij de onderliggende dimensie van fysiek functioneren. Daarnaast demonstreerden 2 items van de HAQ-II *overfit* en een hoge correlatie tussen de residuen, wat duidt op overlap of overtolligheid in de inhoud van deze items. Beide vragenlijsten bevatten meerdere items die verschillend functioneerden binnen subgroepen gebaseerd op leeftijd, geslacht of ziekteduur. Vergeleken met de HAQ-DI, had de HAQ-II minder last van een bodemeffect en een veel bredere schaallengte voor het

meten van de onderliggende dimensie. Al met al wijst dit onderzoek erop dat zowel de HAQ-DI als de HAQ-II valide, unidimensionale vragenlijsten zijn voor het meten van fysiek functioneren. De HAQ-II, die werd ontwikkeld met behulp van Rasch analyse, heeft echter meerdere gunstige schaaleigenschappen waardoor het gebruik van deze vragenlijst wordt geprefereerd boven de traditionele HAQ-DI.

Dankwoord

Graag wil ik iedereen bedanken die direct of indirect heeft bijgedragen aan dit proefschrift. Een aantal personen wil ik hierbij in het bijzonder noemen.

Allereerst natuurlijk de patiënten die aan de verschillende onderzoeken hebben deelgenomen en belangeloos de vragenlijsten hebben ingevuld.

Mijn promotor Mart van de Laar en assistent-promotor Erik Taal wil ik bedanken voor hun goede begeleiding. De duidelijkheid over het onderwerp en de richting van het onderzoek en het vertrouwen en de vrijheid die ik kreeg om hier zelf invulling aan te geven, is me erg goed bevallen.

Wiepke Drossaers-Bakker heeft me begeleid bij het schrijven van de eerste artikelen en me de *do's* en *don'ts* van het publiceren in medische tijdschriften bijgebracht. Gezien mijn niet-medische achtergrond, was dit een enorme hulp.

De data voor het onderzoek onder Egyptische en Nederlandse vrouwen met RA werden verzameld door Alexander Vlaar tijdens zijn co-schap op de afdeling reumatologie. Alexander, bedankt voor de prettige samenwerking bij het opzetten van dit onderzoek en het schrijven van de artikelen.

Martine Veehof stond als één van de onderzoekers aan de wieg van de REMI-TRACT / DREAM studie. Bedankt dat ik ook van deze data gebruik mocht maken en voor de plezierige samenwerking. Heel veel succes met het verdedigen van jouw proefschrift over een paar weken!

Ook wil ik de volgende studenten psychologie bedanken die een groot deel van de vragenlijsten op de polikliniek hebben afgenomen: Christine Bellmann, Ilse Bosgra, Petra Hagens, Nicolette Kupper, Julia Rulle, Lisanne Schmit, Amrah Schotanus, Johan Steehouder, Katharine Steentjes, Lidewij van Gessel en Anouk van der Heij.

Mijn kamergenoot Nelly van Uden-Kraan wil ik graag bedanken voor de erg gezellige tijd in Cubicus.

Verder natuurlijk dank aan alle collega's van communicatiewetenschap en de reumatologen en medewerksters van het secretariaat reumatologie van het MST, de laatsten ook voor de gezellige pauzes tijdens het eerste onderzoek.

Arnold Veldhoen wil ik bedanken voor de opmaak en vormgeving van de omslag van dit proefschrift.

Martijn Visser en Kim de Jong, bedankt dat jullie mijn paranimfen willen zijn.

Rest me nog mijn familie, vrienden en lieve vriendin Jessica te bedanken voor alle steun in de afgelopen jaren.

Curriculum vitae

Peter ten Klooster werd geboren op 3 juni 1976 in Hasselt (Overijssel). Na het behalen van de Mavo (Christelijke Mavo Zwartsluis) in 1992 en de Havo (Carolus Clusius College Zwolle) in 1995, studeerde hij Personeel en Arbeid aan de Hogeschool Windesheim te Zwolle, waar hij in 1999 afstudeerde. In datzelfde jaar begon hij met de studie Toegepaste Communicatiewetenschap aan de Universiteit Twente. Enkele maanden na zijn afstuderen in 2002, startte hij bij de afdeling Communicatiewetenschap van deze universiteit als onderzoeksmedewerker. Eind 2004 kreeg hij de kans te starten met een promotietraject op het gebied van patiënt-gerapporteerde uitkomstmaten binnen de reumatologie onder leiding van prof. dr. Mart van de Laar en dr. Erik Taal. De onderzoeken beschreven in dit proefschrift zijn het resultaat van dit promotietraject.

Peter ten Klooster was born on June 3, 1976 in Hasselt, The Netherlands. After completing his secondary education, he studied Human Resources Management at the Windesheim College for higher professional education in Zwolle, where he graduated in 1999. Subsequently, he studied Communication Science at the University of Twente. Shortly after his graduation in 2002, he started working at the department of Communication Studies of this university as a research assistant. At the end of 2004, he got the opportunity to start a PhD project on patient-reported outcomes in rheumatology under the supervision of prof. dr. Mart van de Laar and dr. Erik Taal. The studies presented in this thesis are the result of this PhD project.

Publications

ten Klooster PM, Drossaers-Bakker KW, Taal E, van de Laar MA. Patient-perceived satisfactory improvement (PPSI): interpreting meaningful change in pain from the patient's perspective. Pain 2006;121:151–157

ten Klooster PM, Vlaar AP, Taal E, Gheith RE, Rasker JJ, El-Garf AK, van de Laar MA. The validity and reliability of the graphic rating scale and verbal rating scale for measuring pain across cultures: a study in Egyptian and Dutch women with rheumatoid arthritis. The Clinical Journal of Pain 2006;22:827–830.

ten Klooster PM, Drossaers-Bakker KW, Taal E, van de Laar MA. Can we assess baseline pain and global health retrospectively? Clinical and Experimental Rheumatology 2007;25:176–181.

ten Klooster PM, Veehof MM, Taal E, van Riel PL, van de Laar MA. Changes in priorities for improvement in patients with rheumatoid arthritis during 1 year of anti-tumour necrosis factor treatment. Annals of the Rheumatic Diseases 2007;66:1485–1490.

ten Klooster, PM, Visser M, de Jong MD. Comparing two image research instruments: the Q-sort method versus the Likert attitude questionnaire. Food Quality and Preference 2008;19:511-518.

ten Klooster PM, Veehof MM, Taal E, van Riel PL, van de Laar MA. Confirmatory factor analysis of the Arthritis Impact Measurement Scales 2 Short Form in patients with rheumatoid arthritis. Arthritis & Rheumatism (Arthritis Care & Research), In press.

Veehof MM, ten Klooster PM, Taal E, van Riel PL, van de Laar MA. Comparison of internal and external responsiveness of the generic Medical Outcome Study Short Form-36 (SF-36) with disease-specific measures in rheumatoid arthritis. The Journal of Rheumatology 2008;35:610–617.

Veehof MM, ten Klooster PM, Taal E, van Riel PL, van de Laar MA. Psychometric properties of the Rheumatoid Arthritis Disease Activity Index (RADAI) in a cohort of consecutive Dutch RA patients starting anti-TNF treatment. Annals of the Rheumatic Diseases, In press.

Vlaar AP, ten Klooster PM, Taal E, Gheith RE, El-Garf AK, Rasker JJ, van de Laar MA. A cross-cultural study of pain intensity in Egyptian and Dutch women with rheumatoid arthritis. The Journal of Pain 2007;8:730–736.