

VU Research Portal

Measurement instruments for patients with rheumatic disorders: a clinimetric appraisal

Swinkels, R.A.H.M.

2005

document version

Publisher's PDF, also known as Version of record

[Link to publication in VU Research Portal](#)

citation for published version (APA)

Swinkels, R. A. H. M. (2005). *Measurement instruments for patients with rheumatic disorders: a clinimetric appraisal*.

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal ?

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

E-mail address:

vuresearchportal.ub@vu.nl

**MEASUREMENT INSTRUMENTS FOR PATIENTS
WITH RHEUMATIC DISORDERS:
A CLINIMETRIC APPRAISAL**

Drs. R.A.H.M. Swinkels

Financial support for the printing and distribution of the thesis has been provided by the Royal Dutch Association of Physiotherapists (KNGF), Reumafonds, St. Annafonds and the Vrije Universiteit Amsterdam. These institutions are gratefully acknowledged for the provided support.

Printing: Datawyse boekproducties

This study is supported by a grant from the Health Insurance Executive Board (CvZ), no. CURE/24101/97.

© R.A.H.M. Swinkels, Amsterdam, 2005

All rights reserved. No part of this book may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording or any information storage or retrieval system, without written permission from the author.

VRIJE UNIVERSITEIT

**Measurement instruments for patients with rheumatic
disorders:
a clinimetric appraisal**

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. T. Sminia,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de faculteit der Geneeskunde
op vrijdag 2 december 2005 om 13.45 uur
in de aula van de universiteit,
De Boelelaan 1105

door

Raymundus Antonius Henricus Maria Swinkels

geboren te Tilburg

promotoren: prof. dr. L.M. Bouter
prof. dr. R.A.B. Oostendorp

CONTENTS

Chapter 1 General introduction	7
Chapter 2 The ICF classification as a system for structuring outcome measurement	19
Chapter 3 Construct validity of impairment measures in rheumatic disorders: which constructs are selected for validation? A systematic review	35
Chapter 4 Impairment measures in rheumatic disorders for rehabilitation medicine and allied health care. A systematic review	53
Chapter 5 Which are the best instruments for measuring disabilities in gait and gait-related activities in patients with rheumatic disorders: a systematic review	75
Chapter 6 Reliability, validity and responsiveness of instruments to assess disabilities in personal care in patients with rheumatic disorders: a systematic review	93
Chapter 7 General discussion	111
Appendix 1 Summary literature search	131
Appendix 2 Criteria for methodological cut-off points	141
Appendix 3 List of abbreviations of instruments and sub-scales	147
Summary	151
Samenvatting	157
Dankwoord	163
About the author	167
Publications	169

Chapter 1

General introduction

Introduction

This thesis focuses on the quality clinimetric properties of measurement instruments for use in clinical practice by physiotherapists and rehabilitation physicians, in terms of reliability and construct validity. There will be specific emphasis in more detail on the assessment of patients with rheumatic disorders, defined in this thesis as rheumatoid arthritis, seronegative polyarthritis (including psoriatic arthritis), osteoarthritis, ankylosing spondylitis, polymyositis and fibromyalgia. It must be emphasized that it was our intention to make an inventory of measurement instruments and some of their methodological aspects in as far as they are relevant for physiotherapy and rehabilitation medicine, mainly in a clinical setting (i.e. not primarily for scientific research). The perspective of physiotherapists and physicians working in the field of rheumatology, who mainly focus on the consequences of the disease, may differ from that of rheumatologists, who focus mainly on the disease itself. As a result of this difference in perspective, some important methods to assess disease activity, such as laboratory tests (erythrocyte sedimentation rate, C-reactive protein) and radiological imaging are not included in our study. However, we realize that these assessments of disease activity might be of major importance for rheumatologists. The majority of the measurement instruments that we investigated were those that are used for the assessment of impairments (pain, range of movement, swelling, disease activity, etc.), disabilities in personal care activities, and disabilities in gait and gait-related activities. The reason for this choice was that these instruments reflect the most frequently formulated treatment goals in rehabilitation medicine and physiotherapy. Thus, since health-related quality of life is of major importance for patients with rheumatic disorders, the use of standardized measurement instruments in clinical settings (examination and treatment of patients) is important in order to objectify and to quantify clinical outcomes. However, in clinical settings the treatment goals mainly seem to focus on the treatment of impairments, and the primary focus is rarely on quality of life in general. For that reason, generic measurement instruments to assess quality of life are seldom used in clinical practice.

Prevalence and clinical features of rheumatic disorders

Rheumatic disorders include a wide range of chronic diseases, the most common of which are osteoarthritis, fibromyalgia and rheumatoid arthritis. Osteoarthritis of the knee, which is the most frequently affected joint, has a prevalence of 16-20 per 1,000 patients in general practice in the Netherlands (1;2). The prevalence of rheumatoid arthritis is relatively similar in many populations, ranging from 0.5-1.0% (3). In the United States, the prevalence of fibromyalgia is approximately 2%, which implies that

some 5 million people are currently suffering from this disease (4-6). No qualitative studies on the prevalence of fibromyalgia have yet been carried out in the Netherlands (7).

The prevalence of osteoarthritis and rheumatoid arthritis is steadily increasing, due to the aging of the population. The most prominent clinical feature of rheumatoid arthritis is systemic joint inflammation. This affects not only the joints, but also the surrounding soft tissues, and results in pain, swollen joints, unstable joints, fatigue, and a decline in general well-being. Osteoarthritis is characterized by a progressive loss of articular cartilage, in combination with increased metabolic activity in the underlying subchondral bone. As in rheumatoid arthritis, the pathological changes in cartilage and bone in osteoarthritis are followed by pathological changes in synovial membranes, capsules, ligaments, and the associated musculature. These rheumatic disorders lead to multiple capsular restrictions, progressive muscle weakness, stiffness of joints, swelling, crepitation, deformity of joints and, due to these clinical features, to inactivity in general. Impairments in body structures and function have an enormous impact on the activities of daily living, normal functioning and health-related quality of life, and in particular the activities involved in personal care, gait and gait-related activities, because these are essential for the ability of an individual to live independently, without assistance from other people.

Classification of rheumatic disorders

The following rheumatic disorders were included in our study: rheumatoid arthritis, seronegative polyarthritis (including psoriatic arthritis), osteoarthritis, ankylosing spondylitis, polymyositis and fibromyalgia. We realize that this selection is open for discussion, in particular with regard to the inclusion of the diagnosis 'fibromyalgia', which is sometimes defined as a 'diagnosis by exclusion of objective findings'. However, we included fibromyalgia because the American College of Rheumatology (ACR) developed operational criteria for the classification of fibromyalgia in 1990 (8). Later, in the 'Declaration of Copenhagen', fibromyalgia was established as a discriminating diagnosis, and the diagnosis was included in the 10th revision of the "International Statistical Classification of Diseases and Related Health Problems (ICD-10) by the World Health Organization in 1993 (9). Fibromyalgia is mentioned in the ICD-10 as "M79-0 Rheumatic, not clearly specified". Rheumatic disorders are frequently classified according to the ACR criteria, which are based on radiological and clinical criteria. The ACR criteria were first published in 1956 (10), and later revised in 1958 and again in 1987, in an attempt to improve their specificity and simplicity, (11). The ACR criteria focus mainly on diagnostic aspects. In the 1990s the European League Against Rheumatism (EULAR) also developed a definition of

response in patients with rheumatoid arthritis, based on a core set of variables (12). The EULAR response criteria are computed according to an index of activity in patients with rheumatoid arthritis, the *Disease Activity Scale* (DAS) (13). The DAS combines information from the Ritchie Articular Index, the swollen joint count, the erythrocyte sedimentation rate, and the patient's global self-assessment of the disease activity (14)(14). With regard to the validity of the ACR criteria and the EULAR response criteria, there is a high level of agreement between the classifications of improvement, and their validity is equivalent and acceptable (14). In the same decade (1992), the OMERACT (Outcome Measures in Rheumatoid Arthritis Clinical Trials) group was established. A key objective of the OMERACT group was to establish a core set of outcome measures for future clinical trials (15). This process was partly data-driven, and partly a combination of discussion, consensus and polling procedures. Their first goal was to determine the domains to be evaluated, and their second was to choose the most optimal outcome measures. This resulted in the following three domains to be evaluated in clinical trials of osteoarthritis in the knee, hip and hand: pain, physical functioning and global self-assessment, and for studies with a minimum follow-up of one year, also joint imaging (15). In 1999, the OMERACT group proposed a new paradigm to capture the essentials of an outcome measure, summarized in the OMERACT filter (16;17). This filter has three components: truth, discrimination, and feasibility. In the study presented in this thesis we have tried to follow the developments of the ACR, the EULAR and the OMERACT group in as far as these were relevant for physiotherapy and rehabilitation medicine.

Classification of the consequences of rheumatic disorders

The majority of interventions for rheumatic disorders, and for rheumatoid arthritis in particular, aim to reduce pain and disease activity, slowing down joint destruction and maintaining and improving the physical functioning of the patient. Certain disciplines, such as physiotherapy and rehabilitation medicine focus on the consequences of the disease. This can have an impact on the (in)dependence and the health-related quality of life of the patients. Self-care in patients with rheumatic disorders is of importance for their degree of independence in functioning, or dependence on family, friends or professional assistance (the domain of environmental factors). This also implies that for the planning of health care facilities, insight into self-care is important. Contrary to the approach of disciplines that diagnose the pathophysiological substrate, this leads to a focus on the clinical problems of patients with rheumatic disorders and on the appropriate treatment goals. The above-mentioned disciplines define treatment goals that do not focus on the disease itself, but mainly on the

consequences of the disease. The consequences of rheumatic disorders may affect all dimensions of well-being, all aspects of health-related quality of life, and also the activities of daily living. This means that the impact is multidimensional, so a classification system which makes it possible to describe these consequences should include signs and symptoms, as well as complaints and the patient's ability to function. This led to the development of the International Classification of Impairments, Disabilities and Handicaps (ICIDH) (18), which was later modified in 2001 as the International Classification of Functioning, Disability and Health (ICF) (19). The ICF will be described in more detail in Chapter 2, in which we also explain why the ICF has been used as the basis for this study.

Because of the multidimensional impact of the majority of rheumatic disorders, it is relevant to quantify this impact over all relevant dimensions of impairments in body functions and structures, limitations in activities, restrictions in participation, environmental factors and personal factors (19). This approach is conceptually different from the traditional classification of diseases, and results in a multidimensional description of the patient's health status and problems.

In relation to the setting of treatment goals, it is important to identify the underlying detrimental or beneficial factors, not only at the level of the biomedical problem, but also at the level of the patient (in terms of personal and environmental factors). These factors may influence the outcome of treatment in a positive or a negative way; they are therefore referred to as 'prognostic factors' for the course of the health problem. Prognostic factors can be generic (general) or non-generic, person-related or environment-related. Of course, prognostic factors such as age, gender, or course in time cannot be influenced. On the other hand, prognostic factors such as tuning between load and loading capacity at the level of organisms, activities and participation, or coping strategy, or avoidance behaviour, can influence the results of the treatment in a positive way. Divergence from the natural course can not only be caused by fluctuations in disease activity, but also by cognitive and/or behavioural factors, and/or by environmental factors. The above-mentioned aspects are important in the process of choosing the most optimal measurement instruments for the examination and/or treatment goals.

By making an inventory of all relevant data, based on a bio-psychosocial approach and incorporating the ICF as a starting point, it should be possible to place the patient's health problem in a wide context of influencing factors (prognostic factors), and to set the most optimal treatment goals.

Outcome assessment

'Clinimetrics' is a term that refers to outcome assessment in general. It was first used by Alvan Feinstein, who defined this term as 'the domain concerned with indexes, rating scales, and other expressions which are used to describe or measure symptoms, physical signs, and other distinctly clinical phenomena in clinical medicine' (20;22). The reasons for the use of measurement instruments are:

1. to quantify clinical signs and symptoms
2. to objectify clinical symptoms
3. to make a baseline assessment
4. to evaluate the clinical course.

By means of measurement instruments, and questionnaires in particular, the above-mentioned information can be gathered and objectified in a standardized way. Measurement instruments can be classified as anthropometrical instruments (e.g. sphygmomanometer, goniometer, spondylometer), questionnaires (to be completed by the patient), and observation lists (to be completed by the examiner). Anthropometrical instruments can be of additional value for the physical examination, while questionnaires can be of great additional value for the anamnesis. Health status measurement instruments usually include the dimensions of physical function, social function, emotional function, pain, and/or the perception of well-being. However, to complicate matters, in the majority of (multidimensional) measurement instruments there is no clear relationship with the ICF framework (21;23). In general, the use of standardized outcome measures is not common in clinical practice. A recent European study of the use of outcome measures in physical medicine and rehabilitation demonstrated that a considerable number of measures are being assessed in a small number of locations and with relatively few patients (22;24). With regard to rheumatoid arthritis, the three most frequently used measurement instruments reflect the ACR core set of variables for use in clinical trials: acute phase reactant, swollen joints and tender joints. The most frequently used questionnaire was found to be the Stanford Health Assessment Questionnaire (HAQ). In relation to the ICF framework, the great majority of instruments focus on the assessment of impairments (22;25).

The first qualitative requirement of a measurement instrument is good reliability. reproducibility. Reproducibility includes two concepts: 'agreement', which can be defined as lack of measurement error, and 'reliability', defined as the extent to which individuals can be distinguished from each other, despite measurement errors (23). In the case of anthropometrical instruments, reproducibility concerns both intra-observer and inter-observer agreement, and for questionnaires it concerns intra-individual reliability.

The second very important property of a measurement instrument is good validity. For diagnostic purposes, sensitivity and specificity are the most important properties that make it possible to discriminate between the presence or absence of disorders, and for evaluative purposes good responsiveness (ability to detect clinically relevant change over time) is of major importance. However, although very important, conclusions concerning responsiveness are not consistently mentioned in all chapters of this thesis. The reason for this is the fact that research into the responsiveness of measurement instruments is hampered by the lack of consensus regarding the preferred method, as well as the lack of data on responsiveness. This will be discussed in more detail in the General Discussion (Chapter 7).

Validity studies do not always differentiate between mono-dimensional measurement instruments and multi-dimensional instruments in the constructs that are used to validate an instrument. In many studies the instruments are validated against constructs that measure a different domain. This gives rise to the question of what impact this has on the numerical value of the calculated (construct) validity of the instruments. It is hypothesized that in studies validating instruments which measure the consequences of rheumatic disorders, validation against a divergent (less similar) construct will result in lower correlations than validation against a convergent (optimally comparable) construct. This is based on the assumption that the more similar the construct, the higher the expected correlation will be. The idea is that similarities in the domain of measurement will demonstrate a stronger relationship than differences in the domain of measurement. This hypothesis was the reason why we sub-divided all the available data regarding construct validity were sub-divided into validity based on correlation with more convergent constructs, and validity based on correlation with more divergent constructs. This will be explained in more detail in Chapter 3. In fact, a divergent construct validity is perfect when the correlation with the instrument to be validated is zero. However, this provides more information about the divergence of the construct than about the construct validity of the instrument to be validated.

Chapter 4 presents the results of a systematic review of measurement instruments for the assessment of impairments in patients with rheumatic disorders, Chapter 5 presents the results of a systematic review of measurement instruments for the assessment of disabilities in gait and gait-related activities in patients with rheumatic disorders, and Chapter 6 describes a systematic review of measurement instruments for the assessment of disabilities in personal care activities in patients with rheumatic disorders.

In summary, three aspects of our extensive literature study were innovative. Firstly, the classification of measurement instruments according to the ICF framework, secondly, differentiation at sub-scale level as much as possible (in order to detect the influence of a sub-scale on the total score), and thirdly, the method of data-analysis, because all data on construct validity are sub-divided into several levels of constructs, and pooled as far as possible within those levels.

Aim of the thesis

The primary aim of this thesis was to generate an overview of available instruments to assess patients with rheumatic disorders, and to qualify these instruments on the basis of their clinimetric properties.

More in detail, our objectives were:

1. To give a short overview of the content and development of the International Classification of Functioning Disability and Health (ICF), and to make an inventory of the degree to which measurement instruments measure the different domains of the ICF in patients with rheumatic disorders (Chapter 2).
2. To make an inventory of constructs, based on the three domains of the ICF against which impairment measures for rheumatic disorders are validated, and to determine whether validation against the same or a similar construct results in higher correlation values than validation against a different construct (i.e. testing convergent-divergent validity) (Chapter 3).
3. To perform a systematic review of the literature in order to make an inventory of available instruments and questionnaires for the assessment of impairments in body structures and function in patients with rheumatic disorders, and to investigate which of these instruments are of acceptable methodological quality with regard to reliability and validity. Furthermore, to investigate the assumption that the construct validity of impairment measures results in higher correlation values when validated against a more similar construct (Chapter 4).
4. To perform a systematic review of the literature in order to make an inventory of available instruments and questionnaires for the assessment of disabilities in gait and gait-related activities in patients with rheumatic disorders, and to investigate which of these instruments are of acceptable methodological quality with regard to reliability and validity. Furthermore, to investigate the hypothesis that the construct validity of impairment measures results in higher correlation values when validated against a more similar construct (Chapter 5).

5. To perform a systematic review of the literature in order to make an inventory of available instruments and questionnaires for the assessment of disabilities in personal care activities in patients with rheumatic disorders, and to investigate which of these instruments are of acceptable methodological quality with regard to reliability and validity. Furthermore, to investigate the hypothesis that the construct validity of impairment measures results in higher correlation values when validated against a more similar construct (Chapter 6).

In Chapter 7 the findings and conclusions of the preceding chapters are discussed in the context of the hypotheses that formed the starting point for this extensive literature review. Finally, in the summary, the findings and conclusions of the preceding chapters are summarized.

References

- 1 Thoonen B, Knottnerus JA: Actual and future prevalence of chronic joint complaints in general practice (in Dutch: Huidige en toekomstige prevalentie van chronische gewrichtsaandoeningen in de huisartsenpraktijk). *Huisarts en Wetenschap* 1991;35:369-73.
- 2 Schouten JS. Artrose. In: *Volksgezondheid Toekomst Verkenning. Health status in the Netherlands in the period 1950-2010* (in Dutch: De gezondheidstoestand van de Nederlandse bevolking in de periode 1950-2010. Ruwaard D, Kramers P, (Eds.). 461-5.1993. The Hague., SDU-Uitgeverij.
- 3 Silman AJ, Pearson JE: Epidemiology and genetics of rheumatoid arthritis. *Arthritis Res* 2002; 4 Suppl 3:S265-S272.
- 4 Arthritis Foundation. Fibromyalgia Syndrome. Retrieved November 8, 2001, from www.arthritis.org.
- 5 American College of Rheumatology 2000. Fibromyalgia. Retrieved November 8, 2001, from www.rheumatology.org.
- 6 Russell IJ: Fibromyalgia syndrome: approaches to management. *Bull Rheum Dis* 1996;45(3):1-4.
- 7 Bazelmans E, Vercoulen JH, Galama JM, van Weel C, van der Meer JW, Bleijenberg G: Prevalence of chronic fatigue syndrome and primary fibromyalgia syndrome in The Netherlands (in Dutch: Prevalentie van het chronisch vermoeidheidssyndroom en primaire fibromyalgie in Nederland). *Ned Tijdschr Geneesk* 1997;141(31):1520-3.
- 8 Wolfe F, Smythe HA, Yunus MB et al.: The American College of Rheumatology 1990 Criteria for the Classification of Fibromyalgia. Report of the Multicenter Criteria Committee. *Arthritis Rheum* 1990; 33(2):160-72.
- 9 Quintner J: Fibromyalgia: the Copenhagen declaration. *Lancet* 1992;340(8827):1103.
- 10 Bennett G, Cobb S, Jacox R, Jessar RA, Ropes MW: Proposed diagnostic criteria for rheumatoid arthritis. *Bull Rheum Dis* 1956;7(4):121-4.
- 11 Arnett FC, Edworthy SM, Bloch DA et al.: The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. *Arthritis and Rheumatism* 1988; 31(3):315-24.
- 12 Gestel AM van, Prevoo ML, Hof MA van 't, Rijswijk MH van, Putte LB van de, Riel PL van. 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(1):34-40.
- 13 Heijde DM van der, Hof HM van 't, Riel PL van, Putte LB van de. Development of a disease activity score based on judgment in clinical practice by rheumatologists. *J Rheumatol* 1993; 20(3):579-81.
- 14 Gestel AM van, Anderson JJ, Riel PL van, Boers M, Haagsma CJ, Rich B, Wells G, Lange ML, Felson DT. ACR and EULAR improvement criteria have comparable validity in rheumatoid arthritis trials. American College of Rheumatology European League of Associations for Rheumatology. *J Rheumatol* 1999;26(3):705-11.
- 15 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(4):799-802.
- 16 Boers M, Brooks PV, Vibeke Strand C, Tugwell P: The OMERACT filter for outcome measures in rheumatology. *J Rheumatol* 1998; 25(2):198-9.
- 17 Bellamy N: Clinimetric concepts in outcome assessment: the OMERACT filter. *J Rheumatol* 1999; 26(4):948-50.
- 18 World Health Organization. *International Classification of Impairments, Disabilities and Handicaps: a manual of classification relating to the consequences of disease.* 1980. Geneva, World Health Organization.
- 19 World Health Organization. *International Classification of Functioning, Disabilities and Health Problems.* 2001. Geneva, World Health Organization.

- 20 Feinstein AR: Clinimetrics. New Haven, Yale University Press, 1987.
- 21 Stucki G, Sigl T: Assessment of the impact of disease on the individual. *Best Pract Res Clin Rheumatol* 2003; 17(3):451-73.
- 22 Haigh R, Tennant A, Biering-Sorensen F et al.: The use of outcome measures in physical medicine and rehabilitation within Europe. *J Rehabil Med* 2001; 33(6):273-8.
- 23 Vet HCW de, Terwee CB, Bouter LM: Current challenges in clinimetrics. *J Clin Epidemiol* 2003; 56(12):1137-41.

Chapter 2

The ICF classification as a system for structuring outcome measurement

Published as: Swinkels RAHM. The ICF classification as a system for structuring outcome measurement. Physiotherapy Singapore 2004;(7):7-13

Abstract*Objectives:*

The aims of this article are threefold. First, a short overview is given of the content and development of the International Classification of Impairments, Disabilities and Handicaps (ICIDH), and particularly of the revision of the ICIDH classification, leading to the International Classification of Functioning Disability and Health (ICF). Second, the degree to which questionnaires measure the different domains of the ICF in chronic diseases is considered, in particular regarding rheumatic disorders. Third, a framework based on the ICF is proposed for the selection of questionnaires.

Methods:

For the second aim a computer-aided literature search (1982-2001) in several databases was performed to identify studies focusing on the clinimetric properties of instruments to assess impairments, disabilities and problems in participation in patients with rheumatic disorders.

Results:

Of the total of measurement instruments and questionnaires, 57.1% are inventories of (impairments in) body function and structure, 37.4% are inventories of (disabilities in) activities, and 5.5% measure problems in participation (handicap). A majority of 59,9% of instruments is intended as a diagnostic tool. 3.8% as a prognostic tool and 36,3% as an evaluative instrument. The distinction between biomedical and biopsychosocial models implies that clinical observation of a patient should distinguish the symptoms and signs of physical disease from those of distress and illness behaviour.

Conclusions:

It is concluded that there are no instruments or questionnaires available which cover all the domains of the ICF. Furthermore, the ICF provides a very useful conceptual framework as the starting point for analysis of the ICF-domains of available assessment instruments and questionnaires, particularly for patients with chronic diseases. However, many questionnaires do not fit the ICF and a framework is presented for selecting optimal instrument(s) in a clinical situation.

Introduction

Historically, medical concepts are primarily derived from the term “disease” and are predominantly characterized by terms like “etiology”, “pathology”, “disease”, and “manifestation”. This reflects a mainly biomedical analysis of a patient’s problems. The International Classification of Diseases and Related Health Problems (ICD 10) (1) is the most frequently used system for classification of diseases. This classification was introduced in 1893 and now in its tenth revision. The ICD is in use for so long because in general medicine it was (and still is) generally sufficient for recording the occurrence of a disease or the recovery and providing valuable and relevant data for studying the experience of health (2). However, the concept of disease as defined by the ICD should not be regarded as appropriate to health care in general, and rehabilitation medicine and physiotherapy in particular, for the classification of the health problems of patients suffering from disease, because the physiotherapist is more often confronted with the consequences of diseases than the disease itself. In rehabilitation medicine and allied health care (e.g. physiotherapy, occupational therapy), there is a need to describe the functioning of the patient in daily living or, in other words, the consequences of disease for the patient. This in addition to the medical diagnosis. A classification system which makes it possible to describe these consequences should include diseases and symptoms as well as complaints and patients’ ability to function, and led to the development of the International Classification of Impairments, Disabilities and Handicaps (ICIDH)(3), later modified as the International Classification of Functioning, Disability and Health (ICF)(4). With the new terminology, rehabilitation is seen as a coordinated process that enhances “activity” and “participation” (5), and many countries started ICF translation and implementation activities (6).

Since there seems to be more and more consensus concerning the acceptance and implementation of the ICF, the question arises as to whether this classification has consequences for outcome assessment. For many years health professionals have been convinced that the disease itself was the most important factor for the patient, but it appears more and more that the consequences of disease are also relevant to understanding how patients function. Also, treatment goals are derived from the consequences, not primarily from the disease itself.

Beside increasing knowledge about the natural course of chronic diseases, also the implementation of the biopsychosocial model in rehabilitation medicine in the last decade has been an important motor for the revision of the ICIDH. Because diagnostic assessment by physical therapists, other allied health care disciplines, and rehabilitation medicine is concerned mainly with functional disorders or

consequences arising as a result of disease (7) there was an increasing need for a system which could classify the problems of patients suffering from a disease in terms of symptoms and consequences of that disease as well as classifying the disease itself. This led to the development of the International Classification of Impairments, Disabilities and Handicaps (ICIDH), accepted as a supplement to the ICD by the World Health Assembly in 1976 and published in 1980 (3). The ICIDH has been revised twice and is now well established, providing the key to a more rational management of chronic disease (7,8). The final revision, the International Classification of Functioning, Disability and Health (ICF) (4) has been published in 18 languages and the bibliography, maintained by the WHO Collaborating Centre on the ICIDH in The Netherlands, lists almost 2000 references.

The aims of this article are threefold. 1) A short overview is given of the content and development of the ICF, and particularly of the revision of the ICIDH. 2) The degree to which measurement instruments¹ measure the different domains of the ICF is considered regarding to assessment in patients with chronic diseases, in particular regarding rheumatic disorders. 3) A framework is proposed, based on the ICF and the psychometric properties of measurement instruments for the selection of the most optimal questionnaires. These three goals will be discussed successively in separate paragraphs.

Content and development of the ICF

The original ICIDH classification (1980) was concerned with dimensions of health-related phenomena that extend the concept of disease. Initially it classified impairments, disabilities and handicaps.

- Impairment was defined as any loss or abnormality of psychological, physiological or anatomical structure or function. The use of the term “impairment” does not imply that disease is present or that individuals should be regarded as sick.
- Disability was defined as any restriction or lack (resulting from an impairment) of ability to perform an activity in the manner or within the range considered normal for a human being. The assessment of disability requires judgment of what is normal, which brings together expectations for functioning in physical, psychological and social terms.
- Handicap was defined as a disadvantage for a given individual resulting from an impairment or disability that limits the fulfilment of a role that is normal for that

¹ Measurement instruments include anthropometric instruments, questionnaires (to be completed by the patient) and observation lists (to be completed by the examiner).

person, related to age, sex, and social and cultural factors. According to this definition “handicap” represents the more social consequences that can arise as a result of impairment and disability. In the ICF the term “handicap” is replaced by the term “problems in participation”.

Actually, the ICIDH describes health problems on three levels. First, it offers a conceptual scheme whereby the consequences of diseases and disorders can be better understood. Second, it offers a classification scheme for the different domains of the consequences of diseases. Third, it proposes a theoretical framework within which to interrelate impairment, disability and handicap (9).

Over the years the ICIDH has been used in an increasing number of settings. However, criticisms of its limitations have increased as well, leading to a major revision of the original ICIDH. This finally resulted in the International Classification of Functioning, Disability and Health (ICF) (4). Table 2.1 presents an overview of the development of these classifications.

Table 2.1 Overview of the different classification systems and the domains they cover.

Classification	Year of publication	Domains
ICD	1893	Diseases
ICIDH	1980	Consequences of disease in terms of impairments, disabilities and handicaps
ICPC	1987	Diseases, symptoms, complaints
ICF	2001	Classification of functions, activities, participation, personal factors and external factors.

Legend: ICD = International Classification of Diseases.
 ICIDH = International Classification of Impairments, Disabilities and Handicaps.
 ICPC = International Classification of Primary Care
 ICF = International Classification of Functioning, Disability and Health.

There are three main criticisms of the original ICIDH: 1) the lack of personal and contextual factors in the ICIDH, where these factors play a key role in the development and prolongation of health problems related to diseases and disorders; 2) overlap among the three dimensions; and 3) a lack of clarity about the causal and temporal relationships among the three dimensions (10).

The results of discussion of these criticisms have been incorporated in the new ICF as environmental (external) factors and personal factors (Fig. 2.1).

In the revised version environmental (external) factors are defined as including the physical, social, and attitudinal environments that influence individual functioning (4). Environmental factors are organized under six headings: products and technology; natural environments and man-made changes to the environment; support and relationships; attitudes, values, and beliefs; services; and systems and policies. Personal factors are described as gender, age, other health conditions (comorbidity), fitness, lifestyle, habits, upbringing, coping styles, social background, education, profession, past and current experience, overall behaviour patterns and character styles, and individual psychological assets (10). A list of so-called personal and environmental factors has been added to the ICF.

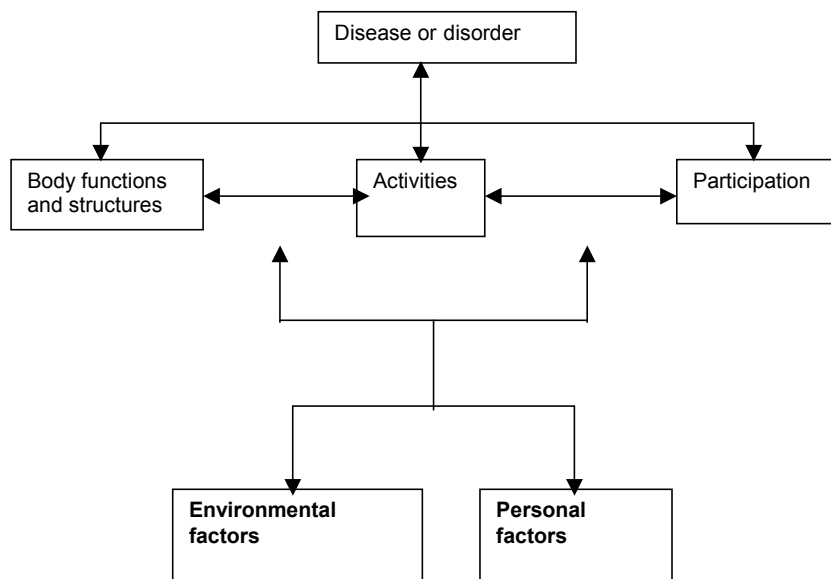


Figure 2.1 Current understanding of interactions and relationships between the dimensions of ICF (WHO, 2001)

In this version aspects of the Theory of Planned Behaviour are incorporated into environmental and personal factors, but the detail is not made explicit. The most recent revision of the ICIDH, the ICF, includes more dimensions than the initial 1980 version. The inclusion of the new dimensions provide a classification system that reaches beyond traditional body-centered descriptors of disability that include factors

that social and empowerment models of disability advocate as being important to understanding disability (11,12,13). The revision of the ICIDH into the ICF is in particular relevant for chronic diseases like rheumatic disorders, because (in particular in physiotherapy and occupational therapy) treatment of patients with chronic diseases is mainly focused on treatment of the consequences of the disease, not on the cause of the disease. Furthermore, in chronic patients the influence of the disease is not restricted to medical impairments, but is much more related to the psychosocial consequences. The ICF model integrates medical and rehabilitation models of disability with psychological models, and appears to integrate behavioural and social cognition models. Significant deviations or loss in body structures and functions are described as impairments. Activities are defined as performance of person-level tasks or activities undertaken by a person in the context of their culture. Because of its negative connotations the term “handicap” has been replaced with “participation” (14). Furthermore, the “participation” classification broadens the range of disablement by classifying most areas of human life. Following revisions in the 1990’s the ICF embodies what is termed the “biopsychosocial model”. Each dimension of disablement (the ICF lists about 250 individual disabilities!) is conceptualised as an interaction between intrinsic features of the individual and that person’s social and physical environment(15).

ICF and outcome measurements in assessment of chronic musculoskeletal disorders.

In general, assessment instruments can be categorized as anthropometrical instruments (goniometer, blood pressure meter, dynamometer, etc.), questionnaires, and observation inventories. There are no assessment instruments available which cover all the ICF domains. One of the problems in investigating assessment instruments is that the ICIDH (1980) was recently revised, while most instruments are developed before that time. This means that the majority of existing measurement instruments do not fit the domains of the ICF classification, in particular not relating the personal factors and environmental factors. However, it is surely possible to distinguish measurement instruments for impairments versus disabilities. In a recent systematic review of literature regarding rating scales for Parkinson’s disease, eleven rating scales are identified: three impairment scales, four disability scales and four scales evaluating both impairment and disability (16). To broaden this distinction to other chronic disorders we take patients with rheumatic disorders as example to use the ICIDH/ICF classification as a framework to analyse measurement instruments. To get an overview of all available measurement instruments for patients with rheumatic disorders we performed a systematic review of the literature (17) in which, among

other things, we analysed the contents of the instruments according the domains of the ICDH classification.

The review concerned the characteristics of measurement instruments for all impairments, disabilities and participation-problems that are relevant in physical therapy and occupational therapy for patients with rheumatic disorders. All studies had to focus on patients suffering from rheumatoid arthritis, seronegative polyarthritis, osteoarthritis, ankylosing spondylitis, polymyositis or fibromyalgia. Different versions of an instrument were considered as separate measurement instruments.

First the Medline database was searched for the period January 1982-April 2001, using search terms for the relevant rheumatic disorders, impairments, disabilities and participation-problems and various search terms for clinimetric properties like reliability, validity and responsiveness. The database of the Centre for Documentation of the Dutch National Institute of Allied Health Care was also searched for the period January 1988 –April 2001, using the same key words². Furthermore, the search in both databases was repeated with the names of the identified measurement instruments. English, French, German and Dutch literature was included. The search was subsequently augmented with a manual search based on the references of the relevant publications, and therefore the search also yielded some publications from before 1982.

The study generated 3405 records relating to 209 instruments. Some of them were anthropometrical instruments; the very great majority were questionnaires.

Measurement instruments can be used for diagnostics, or prognostic/predictive aims or to evaluate treatments and changes over time. Instruments are classified as 'diagnostic' if they can assess status praesens of the patient. They are classified as 'prognostic' or 'predictive' if they claimed to be able to identify patients at risk or to detect a probable pathologic course. Instruments are defined as 'evaluative' if they are able to detect minimal changes over time.

Table 2.2 reflects the properties of the 209 measure instruments among this division.

² The detailed search strategy is available on request from the first author.

Table 2.2 Clinical properties of the 209 measurement instruments for patients with rheumatic disorders regarding the goal of the instruments.

Aims of measurement instrument	%
Diagnostic	59,9%
Prognostic/predictive	3,8%
Evaluative	36,3%

The following criteria were applied:

- All studies had to focus on patients suffering from rheumatoid arthritis, seronegative polyarthritis (including psoriatic arthritis), osteoarthritis, ankylosing spondylitis, polymyositis or fibromyalgia.
- The studies had to contain information about the psychometric properties of instruments to assess relevant impairments, disabilities or problems in participation.
- Many questionnaires focus on more than one domain of the ICDH-classification, or measure more than one variable. Included were: 1) instruments which focus mainly (50% or more of the items) on the variable to be measured; 2) questionnaires with a sub-scale for the variable in question that can be interpreted separately as a single entity.
- Different versions of an instrument were considered as separate measurement instruments.
- Only instruments for the measurement of adult patients were included.

All identified publications were assessed independently on the basis of their title and abstract by two reviewers. In case of disagreement (3%) the article was also assessed by a third reviewer. The assessment was based on a standardized data-collection form (18). This form consists of four sections: general description (name, first author, etc.), assessment domain (according to the ICDH-classification), methodological aspects (concerning reliability, validity,) and aspects of utilization (language, required time to complete, costs, interpretation-criteria etc.). We investigated the following psychometric properties: intra-rater reliability, inter-rater reliability, construct validity and responsiveness.

For clinical use the majority of questionnaires is intended to measure status praesens of the patient (diagnostic; 59,9%). The second most frequent pretension is the evaluative capacity of instruments. However, responsiveness is investigated in 36,3% of all measure instruments (and not all of these 36,3% proved to be responsive).

The ICF domains in these instruments are summarized in Table 2.3.

As is may be observed in Table 2.3, a majority of 57,1% concerns assessment of impairments in function, and 37,4% concerns assessment of disabilities in activities. 5,5% of all analysed instruments/questionnaires concerned assessment of problems in participation. Not included are 'personal factors' and 'environmental factors' because of the fact that the literature inclusion is performed just before the final revision of the ICIDH to ICF, so the starting point for the analysis was the ICIDH.

Table 2.3 Representation of the ICIDH domains in measurement instruments and questionnaires for patients with rheumatic disorders.

ICIDH-domain	Number of instruments	%* (N=209)
Impairments in functions	104	57.1%
Disabilities in activities	68	37.4%
Problems in participation	10	5.5%
Other	27	12.9%

*The total is more than 100% because some questionnaires cover more than one domain.

An ICF-based framework for the selection of the most optimal questionnaires

A measurement instrument is required for each relevant domain in order to come to an objective evaluation of the treatment goals. The problem in clinical situations is not the question whether an instrument exists for a certain domain, but more how to select the best one, taking into account all relevant aspects for the selection of an instrument. For that reason five steps are suggested as a method in selecting the right instruments. The most important factors in the selection of optimal instruments are the treatment goals (based on ICF domains) on one hand, and methodological quality and feasibility on the other hand. Figure 2.2 gives a short overview of the five steps in the selection of optimal measurement instruments.

Step 1: To select the optimal instruments for a particular clinical situation the user first needs to identify exactly what he or she wants to know from the patient and to relate this to one or more domains, depending on the situation: body structures and function, activities, problems in participation, environmental factors or personal factors.

Step 2: Secondly, it is necessary to decide in advance why the information is needed. Is it diagnostic, is it for prognostic goals, or for evaluation of the clinical

course? This decision of the choice should also be related to the required methodological properties, such as reliability, validity, responsiveness, sensitivity and specificity.

Step 3: The third step is deciding what quality of information and analysis is appropriate. The user may be confident about the methodological properties of an instrument, or he may need to undertake research to find the appropriate one.

Those first three steps result in the selection of the optimal instrument.

Step 4: The fourth step is the availability and feasibility of the chosen instruments or questionnaires. Some are published commercially and are protected by copyrights, but the very great majority are freely available in the published literature, particularly in source articles (the first article in which a questionnaire is presented). Beside that, a questionnaire must be available in the concerning language and for use in clinical settings it may not take too much time to complete a questionnaire.

Step 5: The last step in using questionnaires in clinical settings is the correct interpretation of the results. Sometimes (particularly in the case of commercially published questionnaires) the key to the analysis is provided, but in many other cases an extensive literature search is necessary to find it. Furthermore, the frequent use of questionnaires generates large quantities of data, which makes correct interpretation easier.

Of those five steps, the relation to the ICF is in particular relevant in step 1 and step 2, where the health care provider has to define exactly what he/she wants to measure, expressed in the several domains as given by the ICF-classification. This has, of course, consequences for treatment goals on one hand and the preferred intervention on the other hand. Step 3 concerns the methodological quality. Steps 4 and 5 regard feasibility and problems in using the measurement instruments.

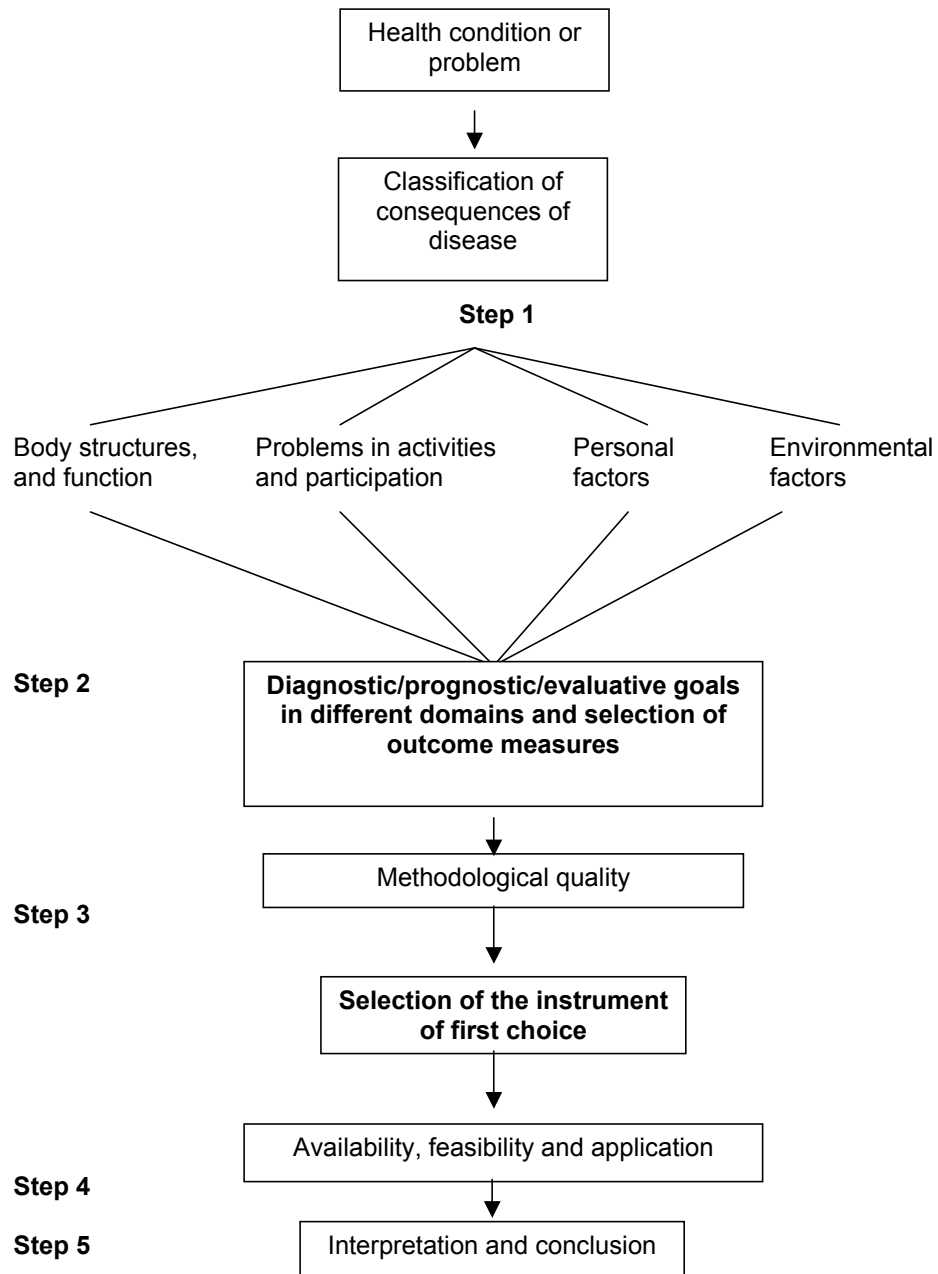


Figure 2.2: Path for selection of optimal clinimetric indexes in clinical settings

Discussion

In rehabilitation medicine and physiotherapy all interaction with patients is directed to the consequences of disease: intervention, treatment goals, and evaluation. Therefore, the ICF is a useful classification system because the ICF makes it possible to describe impairment, disabilities in activities and participation, and external and personal factors. Thus, the ICF makes it possible to classify all variables relevant to the patient's health problems.

The ICIDH, introduced in 1980, has been revised during the past decade. The most important revisions were the incorporation of external, environmental, and personal factors, leading to the integration of the biopsychosocial model into the conceptual framework of the ICIDH. The revision created more clarity about the causal and temporal relationships among the three dimensions. For example, the increasing evidence that dysfunction in chronic pain is poorly correlated with the severity of pain (19,20,21) demonstrates that the causal relationship between impairment and disability is weaker than previously assumed. However, the objections to the initial ICIDH mentioned by Gray et al. (10) , as mentioned in the section 'Content and development of the ICF', have not been upheld. Firstly the ICF model is incapable of explaining many of the findings in the field, including observed shared cumulative patterns of disability and systematic discrepancies between the ratings of disability made by different health professions (22). Not all aspects of health can be related and explained causally.

Secondly, the ICF model has not, of course, answered all the questions and problems regarding outcome assessment in rehabilitation medicine and physiotherapy. Measurement of the participation dimension is particularly difficult (23). However, the ICF was never developed with the intention of creating assessment instruments or questionnaires but was developed to classify the different dimensions of a health problem. That generates the question how valid it is to use the ICF-classification as a framework to analyse properties of measurement instruments. On the other hand it is quite possible to describe a patient's health status in ICF terminology and it is argued that in future questionnaires should be frames using this terminology. In fact, from the described steps to follow for selection of an optimal measurement instrument in particular step one and step two are strongly related to the ICF-classification: step one to select a relevant ICF-domain (what do you want to measure), and step two to couple the relevant domain(s) to the intervention. The intervention must be related to the variables of issue. For example in order to improve joint mobility, hot packs are not appropriate as intervention: the domain (mobility) is chosen incorrect, or the intervention is not related to that ICF-domain.

The distinction between the biomedical and biopsychosocial models implies the differentiation between disease and illness (in which illness is the complex interaction between physical, physiological, psychological and social variables). This must be translated to the clinical situation and to assessment. Clinical observation of a patient should distinguish the symptoms and signs of physical disease from those of distress and illness behaviour. As discussed earlier in this article, the starting point for examination and treatment in allied health care is the health status of the patients. The ICF provides a conceptual framework and a set of terms for the assessment of functional health status. However, it must be emphasized that the provided framework as presented in Figure 2.2 is in particular intended for selection of instruments for use in clinical settings, and to a lesser degree for research settings.

Fitting the ICF to functional health status also requires the incorporation of prognostic or predictive factors. This is one aspect where the ICF is inadequate and possibly one of the reasons why the ICF model is incapable of explaining many of the findings in the field, including observed shared cumulative patterns of disability. In fact history repeats itself: the implementation of functional and prognostic indicators of health status in clinical examination of patients took place during or after revision of the ICDH, and so not all elements are incorporated in the latest version of the ICF. However, the integration of external and personal factors is quite a step in the right direction: they include majority of prognostic factors, like work dissatisfaction, fear avoidance beliefs, coping style, catastrophizing etc. Finally, the revised ICDH did not solve problems in outcome assessment (although it was never developed for that purpose). The main reason for this is that majority of questionnaires were developed in the 1970s and 1980s, before the ICDH was adopted; and therefore many questionnaires do not fit its classification. A negative consequence of this is that multidimensional questionnaires in particular are at risk of mistakes in interpretation if there is only one overall outcome measure, as the underlying components are not clearly identified.

Conclusion

Since its most recent revision the ICF has provided a very useful conceptual framework as a starting point for the analysis of the domains, of available assessment instruments and questionnaires, especially for patients with chronic health problems such as neurological diseases, rheumatic disorders etc. However, many questionnaires do not fit the ICF classification. The fact that the majority of

measurement instruments focus on functional impairment is a logical consequence of the historical biomedical approach to patients and illness.

Beside the domains to be assessed, also the quality of methodological properties and the quality, availability and usefulness of an instrument are relevant aspects in clinical settings. For that reason a five-step-method is suggested to select the optimal instrument.

For clinical application responsiveness is a very relevant property. However, for majority of available instruments responsiveness is not investigated.

References

1. World Health Organization. International Statistical Classification of Diseases and Related Health Problems (ICD-10), 10th revision. Geneva: World Health Organization, 1993.
2. Thuriaux MC. The ICIDH: evolution, status, and prospects. *Disabil Rehabil* 1995;3;4:112-118.
3. World Health Organization. International Classification of Impairments, Disabilities and Handicaps: a manual of classification relating to the consequences of disease. Geneva: World Health Organization; 1980.
4. World Health Organization. International Classification of Functioning, Disabilities and Health Problems. Geneva: World Health Organization; 2001.
5. Disler PB, Cameron ID, Wilson SF. Rehabilitation medicine. A new series on fostering patients' abilities in the presence of disability. *Med J Aust* 2002;177(7):385-386.
6. Hirs WM, de Kleijn-de Vrankrijker MW. From ICF to FIC. *National Institute of Public Health and the Environment (RIVM), Newsletter* 2002,1(1):1-20.
7. Triet EF van, Dekker J, Kerssens JJ, Curfs EC. Reliability of the assessment of impairments and disabilities in survey research in the field of physical therapy. *Int Disabil Studies* 1990;12:61-65.
8. Harwood R, Ebrahim S. Quality of Life. *Lancet*, 1991;637.
9. Badley E. An introduction to the concepts and classifications of the international classification of impairments, disabilities, and handicaps. *Disabil Rehabil* 1993;4:161-178.
10. Gray DB, Hendershot GE. The ICIDH-2: Developments for a new era of outcomes research. Applying outcomes research to disability and health. *Arch Phys Med Rehabil* 2000;81 Suppl 2:S10-S14.
11. Hahn H. The political implications of disability definitions and data. *J Disabil Stud* 1993;4:41-52.
12. Gray DB, Hahn H. Achieving occupational goals: the social affects if stigma. In: Christiansen C, Baum C, editors. *Occupational therapy: enabling function and well-being*. 2nd ed. Thorofare (NJ): Slack; 1997:392-409.
13. Duckworth S. A British view on empowerment: embarking upon a disability career. *Science/Life* 1999;11:10-48.
14. Halbertsma J, Heerkens YF, Hirs WM, Kleijn-deVrankrijker MW, Ravensberg CD, Ten Napel H. Towards a new ICIDH. *Disabil Rehabil* 2000;3:144-156.
15. Bickenbach JE, Chatterji S, Badley EM, Üstün TB. Models of disablement, universalism and the international classification of impairments, disabilities and handicaps. *Soc Sci Med* 1999;48:1173-1187.

16. Ramakers C, Marinus J, Stiggelbout AM, Hilten BJ van. Systematic evaluation of rating scales for impairment and disability in Parkinson's disease. *Movement Disorders* 2002;5:867-876.
17. Swinkels RAHM, Visser E, Kappe YJ, Bouter LM, Oostendorp RAB. *Clinimetric in rheumatic disorders*. Amersfoort: Dutch Institute of Allied Health Care, 2000.
18. Oostendorp, RAB., Elvers, JWH., and Swinkels, RAHM. Judgement list for reading articles. Unpublished, 1997.
19. Tait RC, Margolis RB, Krause SJ, Liebowitz E. Compensation status and symptoms reported by patients with chronic pain. *Arch Phys Med Rehabil* 1988;69:1027-1029.
20. Report of the commission on the evaluation of pain. *Soc Secur Bull* 1987;50:13-44.
21. Loeser JD. What is chronic pain? *Theor Med* 1991;12:213-225.
22. Johnston M. Models of disability. *Physiotherapy Theory and Practice* 1996;12:131-141.
23. Post MWM, de Witte LP, Schrijvers AJP. Quality of life and the ICIDH: towards an integrated conceptual model for rehabilitation outcomes research. *Clin Rehabil* 1999;13:5-15.

Chapter 3

Construct validity of impairment measures in rheumatic disorders: which constructs are selected for validation? A systematic review.

Published as: Construct validity of impairment measures in rheumatic disorders: Which Constructs are selected for validation? A systematic review. Swinkels RAHM, Bouter LM, Oostendorp RAB, Swinkels-Meewisse EJCM, Dijkstra PU, Vet HCW de. Clin Exp Rheum 2005; in press.

ABSTRACT*Purpose:*

This paper focuses on the construct validity of instruments measuring impairments in body structures and function in rheumatic disorders. The objective is: 1) to make an inventory of constructs, based on the domains of the International Classification of Functioning, Disabilities and Health problems (ICF), against which instruments measuring impairments in body structures and function were validated; 2) to analyse whether validation against a similar construct resulted in higher correlation coefficients than validation against a dissimilar construct.

Methods:

In a systematic review papers were identified in which instruments measuring impairments in body structures and function for patients with rheumatic disorders were validated. The instruments identified were assessed on their methodological properties and quality regarding the constructs against which they were validated and (interclass)correlations were recorded and pooled. The pooled (interclass)correlations of similar constructs and dissimilar constructs against which was validated were compared. An instrument was decided to have good construct validity, if the correlation coefficient was 0.50 or higher, and the measurement instrument in question is validated against similar constructs.

Results:

In total 216 papers were identified analysing validity of 42 different instruments. Only 16% of these instruments were validated against instruments that represent the most similar construct. In general, estimates of construct validity were lower when validated against dissimilar constructs, except for instruments measuring impairments in mental functions.

Conclusion:

There is a trend that validation against a similar construct yields higher correlation coefficients than validation against a different construct. If an instrument measuring impairments is validated against the most similar construct, and a criterion of $r > 0.50$ is applied, only 10 out of 42 instruments turned out to be valid.

Introduction

This paper focuses on the construct validity of instruments measuring impairments in body structures and function in patients with rheumatic disorders. In the past decades the content and methodological quality of clinical outcome measures have increasingly become the focus of research. Methodological quality includes standardisation of measurement, reliability, responsiveness and validity.

Several kinds of validity are described in literature, e.g. face validity, criterion validity, content validity and construct validity. Face validity is the validity based on its appearance to the observer. Criterion validity is validity based on comparison with a gold standard. However, in the majority of measurement instruments there is no gold standard available (1). In these instances the focus is on the construct validity in which the correlation between the instrument under study with other instruments measuring this construct is assessed. The process of construct validation presents a considerable challenge to the researcher, because many constructs are multidimensional, (for example personal care, or quality of life) and therefore it is not easy to determine whether an instrument is actually measuring all aspects of the construct of interest (1). Surprisingly, this dilemma is seldom discussed in validity studies. In the majority of validation studies there is a lack of motivation why a certain construct is chosen to validate against and there is lack of reflection whether the instrument measuring the construct actually measures the construct it is intended to measure.

The kind of construct used for validation could cause considerable differences in the correlation with scores on the measure to be validated. For instance in patients with rheumatoid arthritis, the Visual Analogue Scale Pain has been validated against the Articular Index, but also against the Keitel Function Index and against the Haemoglobin-proportion. The resulting correlations were 0.86, 0.20, and 0.21 respectively, indicating that the choice of the construct may influence the strength of the relationship found (2-4).

Consequences of rheumatic disorders can be considered at different domains, at the level of tissues/organisms (impairments in body structures and function), at the level of limitations in daily activities (disabilities), and at the level of participation in social life, including work and hobby's. However, in the majority of validation studies concerning the consequences of rheumatic disorders the validated constructs were compared with various constructs, ignoring attention for the degree of similarity or the content of the domains. As shown in the example above, validating against a dissimilar construct may result in lower correlations than validating against a similar (most comparable) construct.

The International Classification of Functioning, Disabilities and Health problems (5) (ICF), which is often used in the field of physiotherapy and rehabilitation, differentiates between the domains mentioned above. Therefore the ICF is the basis

for our research to differentiate between constructs by classifying the constructs in the domains of the ICF[#].

By using the ICF-classification as a system to qualify constructs one might observe that, for example, a measure for impairments in body structure and function is validated against a measure for limitations in activities.

Besides the choice of the construct to validate against there is a lack of consensus about which correlation coefficients is acceptable for validity and which is not. Correlation coefficients between 0.35 and 0.45 have been considered acceptable (6). But a correlation of 0.40 has also been defined as 'reasonably high' (7). Whereas others state that a $\rho > 0.40$ **and** $p < 0.001$ is cut-off point for acceptable validity (8). However, a perfect correlation in validity studies is not desirable, because it would make the new instrument redundant (9).

Our hypothesis is that validating against a dissimilar construct results in lower correlation coefficients than validating against a similar, optimal comparable, construct. For more similar constructs it can be argued that the cut-off point for 'good' validity should be higher than that for dissimilar constructs. This should be taken into account if one wants to define a priori cut off points for acceptable construct validity.

The aims of this systematic review were: 1) to investigate the constructs against which measures of impairments in body structures or body functions for rheumatic disorders are validated; 2) to examine whether validation against the same or a similar construct results in higher correlation values than validation against a dissimilar construct.

Methods and materials

A systematic literature review of the methodological aspects of measurement instruments was performed for all relevant impairments in body structures and functions related to rheumatic disorders. An impairment in body structure and function was defined as 'relevant' if there was consensus among an expert panel (two experienced rheumatologists, clinically as well as in research, and two collaborators of the WHO-working group ICF) that the impairment is present in the majority of patients with rheumatic disorders.

In another paper, similar methodological aspects of instruments measuring limitations in activities and participation were analysed (10).

[#] The ICF was still under construction during the research period of this project; in fact we used the precursor of the ICF: The International Classification of Impairments, Disabilities and Handicaps (ICIDH). For this paper we used the ICF terminology: 'impairment' is an ICIDH-term which is substituted by the ICF-term 'impairment in body structures and function'; 'disability' (ICIDH) is substituted by 'limitations in activities' (ICF) and 'handicaps' (ICIDH) is substituted by 'participation' (ICF).

Literature search

The Medline database was searched for the period January 1982 - April 2001, using specific search terms for the relevant rheumatic disorders and various search terms for clinimetric properties. The 'rheumatic disorders' included in the search were: rheumatoid arthritis, seronegative polyarthritis (including psoriatic arthritis), osteoarthritis, ankylosing spondylitis, polymyositis and fibromyalgia. The database of the Centre for Documentation of the Dutch National Institute of Allied Health Professions was also searched for the period January 1988 – April 2001. In this search the keywords 'clinimetrics', 'assessment and methods', and all search terms for 'measurement instruments', were used. Additionally the names of the relevant impairments, and of the rheumatic disorders were used as free text words. Finally the names of measurement instruments identified in the first searches were used as free text words in additional searches in the databases. Papers written in English, French, German and Dutch were included. The search was subsequently augmented with a manual search based on the reference list of the identified papers. This final search yielded a small number of papers written before 1982.

Inclusion

Papers were included in the current review if they were performed in populations with the above mentioned rheumatic disorders and if they generated information about the clinimetric properties of the measurement instruments. No restrictions were applied with respect to study design. Further eligibility criteria were: 1) instruments should focus mainly (50% or more of the questions or sub-scales) on impairments in body structures and function; 2) instruments should contain a sub-scale for the impairment that might be interpreted separately as a single entity, independently from the other parts of the questionnaire or other sub-scales.

Data extraction

All papers selected were assessed by two reviewers independently according to a standardised scoring form (11), which was modified for rheumatic disorders. The assessment of the described measurement instruments included description of the construct against which was validated (based on the domains of the ICF: see Table 3.1), and recording of the correlation measures between the used constructs and the measurement instrument to be validated. Correlation coefficients were scored for entire measurement instruments, as well as for sub-scales, whenever separate information about the construct validity of the sub-scales was available. In case of disagreement between the observers (3% of cases), the paper was also assessed by a third reviewer.

All names of instruments and sub-scales used in this review are abbreviated (Appendix 3). References for each instrument are included in the tables to identify literature per measurement instrument.

Construct validity was classified into six levels of constructs, corresponding to the degree of (dis)similarity of the constructs (based on the ICF) against which a measurement instrument was validated (Table 3.1). A measurement instrument was assigned a numerical rating ranging from 1 to 6, based on this classification system.

Table 3.1 Definition of construct level

Level of construct	Definition:
Level 1	Validation against sub-scales or instruments that measure the most similar construct
Level 2	Validation against instruments that measure the same construct (impairment in body structures and function) as well as other impairments or disabilities
Level 3	Validation against instruments that measure other impairments in body structures and function than the ones to be validated
Level 4	Validation against instruments that measure limitations in activities instead of the domain to be validated (impairments in body structures and function)
Level 5	Validation against general measurement instruments that measure all domains of the ICF
Level 6	Validation against general characteristics like age, sex etc.

Level 1 is the most similar construct, (for example: a pain intensity questionnaire is validated against a Visual Analogue Scale for pain). Levels 5 and 6 constructs are the most dissimilar of the constructs. Level 5 construct is an imperfect construct to validate against because the construct includes all domains of the ICF instead of the domain the instrument intends to measure (for example: a pain intensity questionnaire is validated against disabilities in inter-personal relationships). In level 6, measures for assessment of impairments in body structures and function were validated with constructs concerning general characteristics such as age, gender, duration of complaints, etc. Therefore, validation with a level 6 construct means that it is validated against the most dissimilar construct. In this context, the words 'similar' and 'dissimilar' give no indication of the strength of the correlation coefficient (quantification); they qualify whether a measurement instrument is validated against a similar construct, or against a dissimilar construct. However, it is hypothesized that validation against a similar construct will result in a high correlation, and validation against a dissimilar construct will result in a low correlation.

Therefore, the strengths of the correlations in the validation studies were qualified as follows, a correlation coefficient < 0.50 was considered 'poor', a correlation coefficient between 0.50 and 0.65 was considered 'moderate', and a correlation coefficient ≥ 0.65 was considered 'good' (12).

A measurement instrument was decided to have good construct validity, if two conditions were met: 1) the measurement instrument in question is validated against similar constructs, (level 1 and level 2), and 2) the correlation coefficient was 0.50 or higher.

Data-analysis

Pooling of the data was performed within the construct level for each measurement instrument or subscale. The pooled index was :

$$X = (\Sigma[n_i * x_i])/N$$

where X = pooled index, n_i = number of persons included in the study, x_i = value of correlation coefficient (Pearson's r, Spearman's rho) in the study, N = total number of persons in all studies focusing on a specific measurement instrument in one of the construct levels. The pooled index was computed for the values of the correlation coefficients Pearson's r and Spearman's rho. ICC- values are pooled separately in the same way. In measurement instruments with various subscales values for the validity of instruments may be strongly influenced by the values of one or more subscales: if one of the subscales of a questionnaire is poorly correlated with the measure to be validated, it decreases the validity of the total questionnaire, whereas the other subscales might possess good validity. Therefore, the data were pooled for the separate sub-scales as far as possible and whenever they were available.

Results

In total 319 papers were identified in the databases. The references in these papers were checked for potential eligibility and relevance by reading title and abstract. This procedure generated an additional 405 papers resulting in 724 papers which were assessed. Of these 724 papers identified 216 fulfilled the inclusion criteria for detailed reviewing. The number (percentage) of validation studies of the different impairment constructs, together with the levels of the constructs against which they were validated, are shown in Table 3.2.

The most frequently validated instruments are those intended to assess impairments in mental functions (158 times) and those assessing sensory functions including pain (210 times). The majority (44%; n=263) of the instruments and/or subscales was validated against level 4 construct. Furthermore, 27% (n=160) of all investigated instruments was validated against level 6 construct. None of the instruments was validated against level 3 construct, and only a small number of instruments (16%; n=94) were validated against the most similar construct (level 1).

Correlation coefficients per measurement instrument and subscales, when available, are presented for each group of impairments per level of construct in Table 3.3. In the Table the correlations for the measurement instruments and/or sub-scales are only included if data concerning at least level 1 and level 2 construct, or level 1 construct as well as level 4,5 or 6 construct were available. This presentation enables comparison of the validity of the most similar construct with the more dissimilar construct. Only for the measurement instruments of mobility (Table 3.3) all correlations are presented, as no comparison was possible between the most similar and the most dissimilar construct validity because of lack of studies providing these data. If an instrument measuring impairments is validated against the most similar construct, and a criterion of $r > 0.50$ is applied, only 10 out of 42 instruments turn out to be valid.

Comparison between level 1 or 2 construct and level 4, 5 and/or 6 was possible for 22 instruments. In 10 out of these 22 instruments (45%) correlations (Pearson or Spearman's rho) were stronger when validated against level 1 or 2 construct than when validated against level 4-5-6 construct (Table 3.3, indicated in dark grey). For 7 (32%) instruments correlation coefficients were similar or weaker (Pearson or Spearman's rho) when validated against level 1 or 2 construct compared to validation against level 4-5-6 construct (Table 3.3, indicated in light grey). For 5 instruments the results were conflicting (Table 3.3, indicated in middle grey).

Table 3.3 Correlation coefficients for construct validity of instruments for the assessment of mental functions, stiffness, pain, mobility, muscle force and swelling

Instrument and sub-scale @	Level 1 Construct	Level 2 Construct	Level 4 Construct	Level 5 Construct	Level 6 Construct	Ref.
Mental functions						
AHI	0.46*		0.69*	0.23	0.35	16;17
AIMS-anxiety	0.43*		0.19*	0.45*	0.13*	18-40
AIMS-depri	0.43*		0.24*	0.45*	0.23*	see AIMS
AIMS-EmoF	0.57*		0.25*			see AIMS
AIMSS-anxiety	0.41*		0.47	0.45*		35
AIMSS-depri	0.47*		0.57	0.51*		35
BDI	0.67*				0.82	17;41
SSAI	0.57*					17
STAI	0.68*					17
Stiffness						
BASDA	0.69*					42-44
MS-D	0.26		0.16*	0.19*	-0.27	4;24;37;45;46
MS-S	0.51		0.37*	0.28*		4;24;37;45;46
VAS-S			0.91*			47
Pain						
AI	0.83*		0.24*	0.42	0.38*	2-4;24;37;48-51;51-56
AIMS	0.58*			0.75*	0.54*	17-21;21-23;25-40
AIMSD			0.65	0.65	0.43	22;31;34;38;57-60
AIMS-pain	0.40*	0.54*	0.39*	0.60*	0.44*	see AIMS
AIMSD-pain		0.63*		0.81	0.49*	see AIMSD
AIMS2-pain		0.49	0.38*		0.56	26;61;62
AIMS2Dpain	0.66	0.21	0.33*	0.40	0.31	62
AIMSSpain	0.41*		0.50*	0.61*	0.46	35
Dol	0.79*					63-72
EI	0.67		0.36*			73
MPQ	0.37					30;74-77
SAJ	0.55*		0.63			51;56
VAS-P	0.83*		0.26*	0.73		4;37;46;63;78-80
Mobility						
BASMI	0.93*					43;44;81
Chest			0.60			82-84
EPM			0.54	0.54*		85-87
Gonio	0.92					88
MobSpine-CCD			0.37			82-84;89
MobSpine-OWD			0.49			82-84;89
MobSpine-VitCp			0.23			82-84;89
Shob			0.66*		0.19	83;84;89;90
Spond	0.92					88;91
Stest						61;92
FFD			0.18			83;84;89;91
Muscle force						
Gripp			0.31*		-0.33	24;46;93-95
MSI			0.24			96
Sphy	0.87*		0.35*	0.87*	0.21*	37;97-102
Swelling						
AI	0.88*		0.35*	0.43*		2-4;24;37;48-55;103
SAJ		0.64*	0.63			51;56

Legend Table 3.3

@ for explanation of abbreviations see Appendix 3

Level 1 construct: level of the construct (see Table 3.1) against which the instrument is validated. All values expressed in Pearson's r or Spearman's r

*: pooled value

Ref: reference numbers

Dark grey = stronger correlations if validated against level 1 or 2 construct than when validated against level 4-5-6 construct

Light grey = similar or weaker correlations if validated against level 1 or 2 construct compared to validation against level 4-5-6 construct

Middle grey = conflicting results in correlations between validation against level 1-2 construct versus level 4-5-6 construct

Discussion

Instruments measuring impairments in body structures and functions were most often (44%) validated against a level 4 construct (instruments that measure other impairments in body structures and functions than the one to be validated), while 27% was validated against a level 6 construct. On the average only 16% of the instruments was validated against sub-scales or instruments that measure the most optimal comparable (level 1) construct. We wanted to compare correlations of the instruments when validated against different levels of construct. It is therefore possible that some well known and valid instruments

were not included in this review, for the reason that comparison data were not available in literature. Furthermore, some frequently used questionnaires, like the Health Assessment Questionnaire (HAQ), or the Western Ontario McMaster Osteoarthritis Index (WOMAC) were not included, because the main focus of these instruments is not on impairments in body structures and functions. Beside that, instruments that did not include or present separate data for a subscale impairment are excluded from analysis.

The aim of the current study was not to identify the most valid instruments, but to investigate whether the level of constructs against which an instrument is validated influences the outcome of the study.

In about half of the studies, the correlations for all measurement instruments were stronger if validated against the most similar construct. Sometimes the correlations of instruments that were validated against constructs 5-6 level (general; multidimensional) are stronger. However, this phenomenon mainly occurred in the multidimensional AIMS, in particular AIMS-anxiety, AIMS-depri and AIMS-EmoF. On the basis of our systematic review validity of the AIMS subscales 'depression' and 'anxiety' might be reconsidered. These subscales might not exactly cover the construct they are intended to measure. In fact, in general construct validity is a way of hypothesis testing, where the hypothesis is hidden in the hypothetical (theoretical) constructs. The hypothetical constructs contain proposed underlying factors (which we tried to classify in Table 3.1). Possibly the proposed underlying factors (which are

also incorporated in the different items of a questionnaire) are not fully correct, what might explain the found differences in correlations.

From the results of the current study, it occurred that the correlations for construct validity are, in general, weaker when validated against the most dissimilar construct (level 6 construct), compared to the correlations if validated against level 1 construct or level 2 construct. On the other hand, the correlations for the construct validity of instruments to measure impairments in body structures and function were in seven out of 22 cases similar or stronger, when validated against construct 4-5-6 (in general multidimensional instruments), compared to validation against constructs 1 or 2. An explanation for this phenomenon might be that weak correlations of certain sub-scales are compensated by strong correlations of other sub-scales that are more similar with the general construct against which they are validated. Based on this assumption it can be concluded that multidimensional questionnaires need to be validated against multi-dimensional constructs, or separately for each sub-scale against similar constructs.

Another explanation for the discrepancy found might be the correlation between impairments in body structures and function on one hand, and limitations in activities on the other hand. For example, the correlation between pain and disabilities in daily living is $r=0.39$ (13). Probably more than one impairment is present in patients with a rheumatic disorder. Beside pain high scores on measures of disease activity, stiffness and swollen joints might result in considerable disability in activities of daily living. Finally it is possible that the construct of the measurement instrument to be validated actually covers another (or broader) construct than it is supposed to do and therefore possibly results in a weaker correlation than expected if validated against a dissimilar construct.

As seen in this systematic review the reported validity of a measurement instrument depends on the level of construct against which it is validated. If the differentiation in levels of construct as suggested in this paper is applied, the scores for validation against construct 1 are highest. For that reason, only correlation coefficients > 0.50 were considered acceptable, between 0.50 and 0.65 were considered moderate, and if they were greater than 0.65 they were considered to be good (if validated against level 1 or level 2 construct). If this criterion is applied to instruments of this systematic review, the number of valid instruments decreased considerably. For the measurement of impairments in mental functions, only four instruments meet this validity criterion. For the measurement of stiffness two out of four instruments met this validity criterion. For the assessment of pain all instruments except the MPQ and AIMS-pain met this criterion. For the assessment of mobility of joints only BASMI, Gonio and Spond met the validity criterion, and for the assessment of muscle force only the Sphy met the validity criterion (Table 3.3).

Validation against a dissimilar (imperfect) construct resulted in lower validity values, thus different 'cut-off points' should be used to decide whether to accept or to reject a certain validity. Currently, for assessing construct validity it is advocated to formulate hypotheses about the correlation of a scale under study with other instruments (14). In fact, in general construct validity is a way of hypothesis testing, where the hypothesis is hidden in the hypothetical (theoretical) constructs. The hypothetical constructs contain proposed underlying factors (which we have tried to classify in Table 3.1).

In these hypotheses one should take into account that high correlations can only be expected with instruments measuring a similar construct in the same domain (level 1 or 2). Useful hypotheses are, for example: 'the correlation of an impairment scale is higher with another impairment scale than with an activity scale (a more dissimilar construct)', or 'the correlation of a specific subscale of an instrument with a similar subscale of another instrument is higher than with the other subscale of that instrument' (15).

There is clearly a need for further research, in which hypotheses are formulated about the correlation to be expected, taking into account the extent of similarity of the constructs to be compared.

Conclusions

- 1) A majority of 80% of the investigated instruments are validated against dissimilar constructs that measure another impairment and/or an activity, or other aspects than the domain to be validated.
- 2) An average of 16% of instruments measuring impairments is validated against sub-scales or instruments that measure the most similar construct. The available data show that in about 50% the values for construct validity are much lower if validated against a dissimilar construct, compared to validation against the most similar construct or the same impairment combined with (an)other impairment(s) in body structures and function.
- 3) Multidimensional questionnaires need to be validated multi-dimensionally, or separately for each sub-scale against other constructs.

Acknowledgements

We kindly thank Jan Pool for his critical comments on previous versions of this manuscript.

References

1. Portney LG, Watkins MP. Foundations of clinical research. Applications to practice. Appleton and Lange, 2000.
2. Gaston-Johansson F, Gustafsson M. Rheumatoid arthritis: determination of pain characteristics and comparison of RAI and VAS in its measurement. *Pain* 1990;41:35-40.
3. Kalla AA, Kotze TJ, Meyers OL, Parkyn ND. Clinical assessment of disease activity in rheumatoid arthritis: evaluation of a functional test. *Ann Rheum Dis.* 1988;47:773-9.
4. Mallaya RK, Mace BE. The assessment of disease activity in rheumatoid arthritis using a multivariate analysis. *Rheumatol Rehabil.* 1981;20: 14-7.
5. World Health Organization. International Classification of Functioning, Disabilities and Health Problems. 2001. Geneva, World Health Organization.
6. Ghiselle EE. The measurement of occupational aptitude. Los Angeles: Berkely, 1966.
7. Huiskes CJAE, Kraaimaat FW, Bijlsma JWJ. Development of a self-report questionnaire to assess the impact of rheumatic diseases on health and lifestyle. *J Rehabil Sciences* 1990;3:65-70.
8. Martin DP, Engelberg R, Agel J, Swiontkowski MF. Comparison of the Musculoskeletal Function Assessment Questionnaire with the Short Form-36, the Western Ontario and McMaster University Osteoarthritis Index, and the Sickness Impact Profile Health-status measures. *J Bone Joint Surg.* 1997;79A.
9. Deyo RA, Inui TS. Toward clinical applications of health status measures: sensitivity of scales to clinically important changes. *Health Services Research* 1984;19:275-89.
10. Swinkels RAHM, Oostendorp RA, Bouter LM. Which are the best instruments for measuring disabilities in gait and gait-related activities in patients with rheumatic disorders. *Clin Exp Rheumatol* 2004;22:25-33.
11. Oostendorp RAB, Elvers JWH, Swinkels RAHM. Quality scoring list for reading articles (unpublished). 1997.
12. Eliasziw M, Young SL, Woodbury MG, Fryday-Field K. Statistical Methodology for the Concurrent Assessment of Interrater and Intrarater Reliability: Using Goniometric Measurements as an Example. *Phys Ther* 1994;74:777-88.
13. Waddell G. The back pain revolution. Edinburgh: Churchill Livingstone, 1998.
14. Streiner DL, Norman GR. Health Status Measurement Scales. A practical guide to their development and use. Oxford: Oxford University Press, 2003.
15. Roorda LD, Roebroek ME, Lankhorst GJ, van Tilburg T, Bouter LM. Measuring functional limitations in rising and sitting down: development of a questionnaire. *Arch Phys Med Rehabil* 1996;77:663-9.
16. Nicassio PM, Wallston KA, Callahan LF, Herbert M, Pincus T. The measurement of helplessness in rheumatoid arthritis. The development of the arthritis helplessness index. *J Rheumatol* 1985;12:462-7.
17. Hagglund KJ, Roth DL, Haley WE, Alarcon GS. Discriminant and convergent validity of self-report measures of affective distress in patients with rheumatoid arthritis. *J Rheumatol* 1989;16:1428-32.
18. Meenan RF, Anderson JJ, Kazis LE, Egger MJ, Altz-Smith M, Samuelson CO, Jr. *et al.* Outcome assessment in clinical trials. Evidence for the sensitivity of a health status measure. *Arthritis Rheum* 1984;27:1344-52.
19. Meenan RF, Gertman PM, Mason JH. Measuring health status in arthritis. The arthritis impact measurement scales. *Arthritis Rheum* 1980;23:146-52.
20. Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. *Med Care* 1989;27: S178-S189.
21. Anderson JJ, Firschein HE, Meenan RF. Sensitivity of a health status measure to short-term clinical changes in arthritis. *Arthritis Rheum* 1989;32:844-50.
22. Hidding A, Santen M van, Klerk Ed, Gielen X, Boers M. Discordance between self-report measures and clinical observations of functional disability in ankylosing spondylitis, rheumatoid arthritis and fibromyalgia. Group physical therapy in ankylosing spondylitis (thesis). p94-104. Meppel: Krips Repro, 1993.

23. 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.
24. Anderson JJ, Chernoff MC. Sensitivity to change of rheumatoid arthritis clinical trial outcome measures. *J Rheumatol* 1993;20:535-7.
25. Potts MK, Brandt KD. Evidence of the validity of the Arthritis Impact Measurement Scales. *Arthritis Rheum* 1987;30:93-6.
26. 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.
27. Fitzpatrick R, Ziebland S, Jenkinson C, Mowat A, Mowat A. A comparison of the sensitivity to change of several health status instruments in rheumatoid arthritis. *J Rheumatol* 1993;20:429-36.
28. Liang MH, Larson MG, Cullen KE, Schwartz JA. Comparative measurement efficiency and sensitivity of five health status instruments for arthritis research. *Arthritis Rheum* 1985;28:542-7.
29. Meenan RF, Gertman PM, Mason JH, Dunaif R. The arthritis impact measurement scales. Further investigations of a health status measure. *Arthritis Rheum* 1982;25: 1048-53.
30. Keefe FJ, Caldwell DS, Queen KT, Gil KM, Martinez S, Crisson JE *et al.* Pain coping strategies in osteoarthritis patients. *J Consult Clin Psychol.* 1987;55: 208-12.
31. Taal E, Seydel ER, Jacobs J, Wiegman O, Rasker JJ. The Dutch Arthritis Impact Measurement Scales (DUTCH-AIMS): measuring the physical and psychological health aspects of rheumatoid arthritis (in Dutch). *Gedrag & Gezondheid* 1989;17: 69-74.
32. Liang MH, Fossel AH, Larson MG. Comparisons of five health status instruments for orthopedic evaluation. *Med Care* 1990;28:632-42.
33. Goepfing J, Doyle MA, Charlton SL, Lorig K. A nursing perspective on the assessment of function in persons with arthritis. *Res Nurs Health* 1988;11:321-31.
34. Jacobs JW, Oosterveld FG, Deurbouts N, Rasker JJ, Taal E, Dequeker J *et al.* Opinions of patients with rheumatoid arthritis about their own functional capacity: how valid is it? *Ann Rheum Dis* 1992;51:765-8.
35. Wallston KA, Brown GK, Stein MJ, Dobbins CJ. Comparing the short and long versions of the Arthritis Impact Measurement Scales. *J Rheumatol* 1989;16: 1105-9.
36. Duffy CM, Watanabe Duffy KN, Gladman DD, Brubacher BB, Buskila D, Langevitz P *et al.* The utility of the arthritis impact measurement scales for patients with psoriatic arthritis. *J Rheumatol* 1992;19:1727-32.
37. Hawley DJ, Wolfe F. Sensitivity to change of the health assessment questionnaire (HAQ) and other clinical and health status measures in rheumatoid arthritis: results of short-term clinical trials and observational studies versus long-term observational studies. *Arthritis Care Res* 1992;5:130-6.
38. Vliet Vlieland TP, Wijk TP van der, Jolie IM, Zwiderman AH, Hazes JM. Determinants of hand function in patients with rheumatoid arthritis. *J Rheumatol* 1996;23: 835-40.
39. Weinberger M, Samsa GP, Tierney WM, Belyea MJ, Hiner SL. Generic versus disease specific health status measures: comparing the sickness impact profile and the arthritis impact measurement scales. *J Rheumatol* 1992;19:543-6.
40. Husted J, Gladman DD, Farewell VT, Long JA. Validation of the revised and expanded version of the Arthritis Impact Measurement Scales for patients with psoriatic Arthritis. *J Rheumatol* 1996;23:1015-9.
41. Callahan LF, Kaplan MR, Pincus T. The Beck Depression Inventory, Center for Epidemiological Studies Depression Scale (CES-D), and General Well-Being Schedule depression subscale in rheumatoid arthritis. Criterion contamination of responses. *Arthritis Care Res* 1991;4:3-11.

42. Garrett S, Jenkinson T, Kennedy LG, Whitelock H, Gaisford P, Calin A. A new approach to defining disease status in ankylosing spondylitis: the Bath Ankylosing Spondylitis Disease Activity Index. *J Rheumatol* 1994;21:2286-91.
43. Calin A, Nakache JP, Gueguen A, Zeidler H, Mielants H, Dougados M. Defining disease activity in ankylosing spondylitis: is a combination of variables (Bath Ankylosing Spondylitis Disease Activity Index) an appropriate instrument? *Rheumatology (Oxford)* 1999;38:878-82.
44. Spoorenberg A, Heijde D van der, de Klerk E, Dougados M, de Vlam K, Mielants H et al. A comparative study of the usefulness of the Bath Ankylosing Spondylitis Functional Index and the Dougados Functional Index in the assessment of ankylosing spondylitis. *J Rheumatol* 1999;26:961-5.
45. Vliet-Vlieland TPM, Zwiderman AH, Breedveld FC, Hazes JMW. Measurement of morning stiffness in rheumatoid arthritis clinical trials. *J Clin Epidemiol* 1997; 50.
46. Rhind VM, Bird HA, Wright V. A comparison of clinical assessments of disease activity in rheumatoid arthritis. *Ann Rheum Dis* 1980;39:135-7.
47. Oosterhof J. Research on the reproducibility of the VAS for stiffness in patients with rheumatoid arthritis (in Dutch). (unpublished)1992.
48. Ritchie DM, Boyle JA, McInnes JM, Jasani MK, Dalakos TG, Grieveson P et al. Clinical studies with an articular index for the assessment of joint tenderness in patients with rheumatoid arthritis. *Q J Med* 1968;37:393-406.
49. Camp AV. An articular index for the assessment of rheumatoid arthritis. *Orthopaedics* 1971;4:39-45.
50. Dougados M, Gueguen A, Nakache JP, Nguyen M, Mery C, Amor B. Evaluation of a functional index and an articular index in ankylosing spondylitis. *J Rheumatol* 1988;15:302-7.
51. Prevoo ML, Kuper IH, Hof MA van't, Leeuwen MA van, Putte LB van de, Riel PL van. Validity and reproducibility of self-administered joint counts. A prospective longitudinal followup study in patients with rheumatoid arthritis. *J Rheumatol* 1996;23:841-5.
52. Hart LE, Tugwell P, Buchanan WW, Norman GR, Grace EM, Southwell D. Grading of tenderness as a source of interrater error in the Ritchie articular index. *J Rheumatol* 1985;12:716-7.
53. Doyle DV, Dieppe PA, Scott J, Huskisson EC. An articular index for the assessment of osteoarthritis. *Ann Rheum Dis* 1981;40:75-8.
54. Fuchs HA, Brooks RH, Callahan LF, Pincus T. A simplified twenty-eight-joint quantitative articular index in rheumatoid arthritis. *Arthritis Rheum* 1989;32: 531-7.
55. Smolen JS, Breedveld FC, Eberl G, Jones I, Leeming M, Wylie GL et al. Validity and reliability of the twenty-eight-joint count for the assessment of rheumatoid arthritis activity. *Arthritis Rheum* 1995;38:38-43.
56. Calvo FA, Calvo A, Berrocal A, Pevez C, Romero F, Vega E et al. Self-administered joint counts in rheumatoid arthritis: comparison with standard joint counts. *J Rheumatol* 1999;26:536-9.
57. Oosterveld FGJ, Jacobs JWG, Deurbouts N, Rasker JJ, Dequeker J, Taal E. The DUTCH-AIMS and RA: a good measure for physical functioning (in Dutch). *Ned Tijdschr Fysioth* 1990;100:317-9.
58. Vlieland TP, Ende CH van den, Breedveld FC, Hazes JM. Evaluation of joint mobility in rheumatoid arthritis trials: the value of the EPM-range of motion scale. *J Rheumatol* 1993;20:2010-4.
59. Taal E, Jacobs JW, Seydel ER, Wiegman O, Rasker JJ. Evaluation of the Dutch Arthritis Impact Measurement Scales (DUTCH-AIMS) in patients with rheumatoid arthritis. *Br J Rheumatol* 1989;28:487-91.
60. Fitzpatrick R, Newman S, Lamb R, Shipley M. A comparison of measures of health status in rheumatoid arthritis. *Br J Rheumatol* 1989;28:201-6.
61. Molendijk M, Colard I, Jacobs JWG, Oosterveld FGJ. Drukpunten bij het fibromyalgie syndroom; de betrouwbaarheid in de dagelijkse praktijk. *Ned Tijdschr Fysioth* 1991;101:31-6.

62. Riemsma RP, Taal E, Rasker JJ, Houtman PM, Paassen HC van, Wiegman O. Evaluation of a Dutch version of the AIMS2 for patients with rheumatoid arthritis. *Br J Rheumatol* 1996;35:755-60.
63. Wigers SH, Skrondal A, Finset A, Gøtestam KG. Measuring change in fibromyalgic pain: the relevance of pain distribution. *J Musculoskeletal Pain* 1997;5:29-41.
64. McCarthy DJ, Jr., Gatter RA, Phelps P. A dolorimeter for quantification of articular tenderness. *Arthritis Rheum* 1965;44:551-9.
65. Wessel J. The reliability and validity of pain threshold measurements in osteoarthritis of the knee. *Scand J Rheumatol* 1995;24:238-42.
66. Puttick M, Schulzer M, Klinkhoff A, Koehler B, Rangno K, Chalmers A. Reliability and reproducibility of fibromyalgic tenderness, measurement by electronic and mechanical dolorimeters. *J Musculoskeletal Pain* 1995;3:3-14.
67. Atkins CJ, Zielinski A, Klinkhoff AV, Chalmers A, Wade J, Williams D et al. An electronic method for measuring joint tenderness in rheumatoid arthritis. *Arthritis Rheum* 1992;35:407-10.
68. Guthrie NG, Edworthy SM, Martin L, Fritzler MJ, Kinsella TD, Ryan JP. Comparing three techniques for the evaluation of fibromyalgia. *Physiotherapy Canada* 1994;46:94-7.
69. Daltroy LH, Larson MG, Roberts NW, Liang MH. A modification of the Health Assessment Questionnaire for the spondyloarthropathies. *J Rheumatol* 1990;17: 946-50.
70. Smythe HA, Buskila D, Urowitz S, Langevitz P. Control and "fibrositic" tenderness: comparison of two dolorimeters. *J Rheumatol* 1992;19: 768-71.
71. Tunks E, Crook J, Norman G, Kalaher S. Tender points in fibromyalgia. *Pain* 1988;34:11-9.
72. Langley GB, Fowles M, Sheppeard H, Wigley RD. A simple pressure dolorimeter for the quantification of joint tenderness in inflammatory arthritis. *Rheumatol Int* 1983;3:109-12.
73. Mander M, Simpson JM, McLellan A, Walker D, Goodacre JA, Dick WC. Studies with an enthesitis index as a method of clinical assessment in ankylosing spondylitis. *Ann Rheum Dis* 1987;46:197-202.
74. Lankveld W van, Pad-Bosch P van't, Putte LBA van de, Staak C van de, Näring G. Pain from rheumatoid arthritis measured with the Visual Analogue Scale and the Dutch version of the McGill pain questionnaire. *Ned Tijdschr Geneesk.* 1992;136:1166-70.
75. Nordenskiöld U, Grimby G. Assessments of disability in women with rheumatoid arthritis in relation to grip force and pain. *Disabil Rehabil* 1997;19: 13-9.
76. Burckhardt CS, Bjelle A. A Swedish version of the short-form McGill Pain Questionnaire. *Scand J Rheumatol* 1994;23:77-81.
77. Melzack R. The McGill Pain Questionnaire: major properties and scoring methods. *Pain* 1975;1:277-99.
78. Downie WW, Leatham PA, Rhind VM, Wright V, Branco JA, Anderson JA. Studies with pain rating scales. *Ann Rheum Dis* 1978;37:378-81.
79. Langley GB, Sheppaerd H. Problems associated with pain measurement in arthritis: comparison of the visual analogue and verbal rating scales. *Clin Exp Rheumatol* 1984;2:231-4.
80. Hurst NP, Kind P, Ruta D, Hunter M, Stubbings A. Measuring health-related quality of life in rheumatoid arthritis: validity, responsiveness and reliability of EuroQol (EQ-5D). *Br J Rheumatol* 1997;36:551-9.
81. Jenkinson TR, Mallorie PA, Whitelock HC, Kennedy LG, Garrett SL, Calin A. Defining spinal mobility in ankylosing spondylitis (AS). The Bath AS Metrology Index. *J Rheumatol* 1994;21:1694-8.
82. Moll JM, Wright V. The pattern of chest and spinal mobility in ankylosing spondylitis. An objective clinical study of 106 patients. *Rheumatol Rehabil* 1973;12:115-34.
83. Heikkila S, Viitanen JV, Kautiainen H, Kauppi M. Sensitivity to change of mobility tests; effect of short term intensive physiotherapy and exercise on spinal, hip, and shoulder measurements in spondyloarthropathy. *J Rheumatol* 2000;27: 1251-6.

84. Viitanen JV, Heikkila S, Kokko ML, Kautiainen H. Clinical assessment of spinal mobility measurements in ankylosing spondylitis: a compact set for follow-up and trials? *Clin Rheumatol* 2000;19:131-7.
85. Vlieland TP, Ende CH van den, Breedveld FC, Hazes JM. Evaluation of joint mobility in rheumatoid arthritis trials: the value of the EPM-range of motion scale. *J Rheumatol.* 1993;20:2010-4.
86. Ferraz MB, Oliveira LM, Araujo PM, Atra E, Walter SD. EPM-ROM Scale: an evaluative instrument to be used in rheumatoid arthritis trials. *Clin Exp Rheumatol* 1990;8:491-4.
87. Ende CHM van den, Vliet-Vlieland TPM, Breedveld FC, Hazes JM. The EPM-ROM scale: a valuable measurement instrument for the mobility of the joints of patients with rheumatoid arthritis (in Dutch). *Ned Tijdschr Fysioth* 1994;104: 86-91.
88. Reynolds PM. Measurement of spinal mobility: a comparison of three methods. *Rheumatol Rehabil* 1975;14:180-5.
89. Viitanen JV, Kautiainen H, Suni J, Kokko ML, Lehtinen K. The relative value of spinal and thoracic mobility measurements in ankylosing spondylitis. *Scand J Rheumatol* 1995;24:94-7.
90. Viitanen JV. Thoracolumbar rotation in ankylosing spondylitis. A new noninvasive measurement method. *Spine* 1993;18: 880-83.
91. Sturrock RD, Wojtulewski JA, Hart FD. Spondylometry in a normal population and in ankylosing spondylitis. *Rheumatol Rehabil* 1973;12:135-42.
92. Roberts WN, Liang MH, Pallozzi LM, Daltroy LH. Effects of warming up on reliability of anthropometric techniques in ankylosing spondylitis. *Arthritis Rheum* 1988;31:549-52.
93. Nordenskiöld UM, Grimby G. Grip force in patients with rheumatoid arthritis and fibromyalgia and in healthy subjects. A study with the Grippit instrument. *Scand J Rheumatol* 1993;22:14-9.
94. Dellhag B, Bjelle A. A Grip Ability Test for use in rheumatology practice. *J Rheumatol* 1995;22:1559-65.
95. Dellhag B, Hosseini N, Bremell T, Ingvarsson PE. Disturbed grip function in women with rheumatoid arthritis. *J Rheumatol* 2001;28:2624-33.
96. Stucki G, Schonbachler J, Bruhlmann P, Mariacher S, Stoll T, Michel BA. Does a muscle strength index provide complementary information to traditional disease activity variables in patients with rheumatoid arthritis? *J Rheumatol* 1994;21: 2200-5.
97. Hamilton GF, McDonald C, Chenier TC. Measurement of grip strength: validity and reliability of the sphygmomanometer and Jamar Grip Dynamometer. *J Orthop Sports Phys Ther* 1992;16:215-19.
98. Helewa A, Goldsmith CH, Smythe HA. The modified sphygmomanometer-an instrument to measure muscle strength: a validation study. *J Chronic Dis* 1981;34:353-61.
99. Helewa A, Goldsmith CH, Smythe HA. Patient, observer and instrument variation in the measurement of strength of shoulder abductor muscles in patients with rheumatoid arthritis using a modified sphygmomanometer. *J Rheumatol* 1986;13: 1044-49.
100. Spiegel JS, Paulus HE, Ward NB, Spiegel TM, Leake B, Kane RL. What are we measuring? An examination of walk time and grip strength. *J Rheumatol* 1987;14:80-6.
101. Lee P, Baxter A, Dick WC, Webb J. An assessment of grip strength measurement in rheumatoid arthritis. *Scand J Rheumatol* 1974;3:17-23.
102. Pincus T, Brooks RH, Callahan LF. Reliability of grip strength, walking time and button test performed according to a standard protocol. *J Rheumatol* 1991;18: 997-1000.
103. Prevoo ML, Riel PLCM van, Hof MA van, Rijswijk MH van, Leeuwen MA van, Kuper HH et al. Validity and reliability of joint counts. A longitudinal study in patients with recent onset rheumatoid arthritis. In Prevoo ML, ed. *Clinical assessment in rheumatoid arthritis: aspects of validity and reproducibility*, p 19-31. Nijmegen: Catholic University Nijmegen, 1995.

Chapter 4

Impairment measures in rheumatic disorders for rehabilitation medicine and allied health care. A systematic review.

*Published as: Swinkels RAHM, Bouter LM, Oostendorp RAB, Ende CHM van den.
Impairment measures in rheumatic disorders for rehabilitation medicine and allied
health care. A Systematic review. Rheum Int. 2005;25(7):501-512*

ABSTRACT*Objective:*

The objective of this study is to provide a critical overview of available instruments to assess impairments in patients with rheumatic disorders, and to recommend reliable and valid instruments for use in allied health care and rehabilitation medicine.

Methods:

A computer-aided literature search (1982-2004) in several databases was performed to identify studies focusing on the clinimetric properties of instruments to assess impairments in function in patients with rheumatic disorders. Data on intra-rater reliability, inter-rater reliability and construct validity were extracted in a standardized way. Explicit criteria were applied for reliability and validity.

Results:

The search identified a total of 49 instruments to assess impairments in functions in patients with rheumatic disorders; nineteen met the criteria for reliability, 22 met the criteria for validity, and eleven out of the 49 appeared to meet the criteria for both reliability and validity. In summary, evidence of both reliability and validity was only found for eleven out of 49 instruments for the assessment of impairments in patients with rheumatic disorders.

Conclusion:

Only a limited number of the identified instruments for the assessment of impairments is both reliable and valid. Allied health care professionals should be cautious in the selection of measurement instruments to assess their patients.

Introduction

In rehabilitation medicine and allied health care for patients with rheumatic disorders, attention has shifted from disease severity to impairments, disabilities and problems in participation (1). In particular, the development of the International Classification of Impairments, Disabilities and Handicaps (ICIDH) (2) and after its revision in the past decade now renamed the *International Classification of Functioning, disabilities and Health Problems* (ICF) (3) has encouraged this shift. The ICF distinguishes the disease itself from disease consequences and influencing contextual factors. In rehabilitation medicine and allied health care the treatment goals can not directly be derived from the severity of the disease. If the treatment goals are expressed in terms of impairment, disability or problems in participation (as in the ICF) these must be assessed and objectified (4;5).

From this point of view it is essential to choose adequate measurement instruments, based on evidence of their methodological quality (4;6-11).

This study focuses on the psychometric properties of instruments to measure impairments in functions in body and structures of patients with rheumatic disorders, a population which is frequently treated with physiotherapy and occupational therapy. 'Impairment' is defined as any significant loss or abnormality of psychological, physiological or anatomical structure or function. Impairment is part of a health condition, but does not necessarily indicate that a disease is present or that the individual should be regarded as sick (12). 'Disabilities' are defined as limitations in the performance of a task or action by an individual (12). We recently demonstrated that 57% of all available instruments to assess people with rheumatic disorders focus on impairments (13). We started to focus on instruments to measure impairments because of this dominance. In majority of clinical situations (60%) instruments are used to measure status praesens of the patient (13).

Inter-rater and intra-rater reliability is the first characteristic of an instrument that is required to be adequate; if a measurement is not reproducible it will not be useful. Furthermore, an important requirement of a measurement scale is the unidimensionality (of subscales). If sufficient reliability has been demonstrated, it is also relevant to assess the validity.

In the literature, several aspects of validity are described, but construct validity is the most commonly assessed aspect. An instrument can be validated by studying its correlation with an optimally comparable construct (for instance the same impairment) or by correlation with an imperfect construct, such as gender, age, etc. In this study we distinguished various levels of construct validity, because we expected to find higher correlation values if measurement instruments were validated against an optimally comparable construct than if validated against an imperfect construct (14). If so, this could be relevant for interpretation of validity studies.

The aim of this study was threefold. The first objective was to make an inventory of available instruments and questionnaires for the assessment of impairments in patients with rheumatic disorders. The second aim was to investigate which of these instruments have acceptable methodological quality with regard to reliability and validity. The third aim was to investigate the assumption that construct validity results in higher correlation values when validated against a more suitable construct.

Method and materials

We performed a systematic review of the literature on the methodological characteristics of measurement instruments for all impairments that are relevant in allied health care for patients with rheumatic disorders. The term 'measurement instrument' includes questionnaires (to be completed by the patient), observation lists (to be completed by the researcher) and anthropometric measurement instruments (technical measurement tools). Not all reported properties of instruments are appropriate for all kind of instruments. For example inter-rater reliability (the degree to which two or more observers can obtain the same ratings) is inappropriate for questionnaires that are completed by the patient.

Inclusion procedure

The following criteria were applied:

- All studies had to focus on patients suffering from rheumatoid arthritis, seronegative polyarthritis (including psoriatic arthritis), osteoarthritis, ankylosing spondylitis, polymyositis or fibromyalgia.
- The studies had to contain information about the psychometric properties of instruments to assess impairments in mental functions, stiffness, pain, joint mobility (in terms of Range of Motion), muscle force and swelling (ICIDH-classification (15)).
- Many questionnaires focus on more than one domain of the ICIDH-classification, or measure more than one variable. Included were: 1) instruments which focus mainly (50% or more of the items) on the impairment to be measured; 2) questionnaires with a sub-scale for the impairment in question that can be interpreted separately as a single entity.
- Different versions of an instrument were considered as separate measurement instruments.
- Only instruments for the measurement of adult patients were included.

Because of the method of data-reduction as described above, some frequently used questionnaires, like the Health Assessment Questionnaire (HAQ), or the Western Ontario McMaster Osteoarthritis Index (WOMAC), the EuroQoL and Nottingham Health Profile are not included. The condition to be included was that 50% or more of

the items should focus on the impairment to be measured. Questionnaires like HAQ, WOMAC and Nottingham Health Profile are dominantly focussing on disabilities.

Literature search

First the Medline database was searched for the period January 1982 - May 2004, using search terms for the relevant rheumatic disorders and various search terms for psychometric properties³. The database of the Centre for Documentation of the Dutch Institute of Allied Health Care was also searched for the period January 1988– May 2004, using the same key-words. Furthermore, the search in both data-bases was repeated with the names of the identified measurement instruments. English, French, German and Dutch literature was included. The search was subsequently augmented with a manual search based on the references of the relevant publications, and therefore the search also yielded in some publications from before 1982. In total 156 measurement instruments were identified.

Data-extraction

All identified publications were assessed independently on the basis of their title and abstract by two reviewers (RS and YK). In case of disagreement (3%) the article was also assessed by a third reviewer. The assessment was based on a standardized data-collection form (16). This form consists of four sections: general description (name, first author, etc.), assessment domain (according to the ICDH-classification), methodological aspects (concerning reliability, validity,) and aspects of utilization.

Methodological criteria for psychometric properties

We investigated the following psychometric properties: intra-rater reliability, inter-rater reliability, construct validity and responsiveness. To interpret the data on reliability and validity we used criteria based on De Jong et al. (17), Eliasziw et al. (18) and Doeglas et al. (19) (Table 4.1).

The cut-off points for 'good' reliability are supported by Weiner and Stewart, who also suggested 0.85 as criterion (20;21).

³ The detailed search strategy is available on request from the first author.

Table 4.1.

Applied cut-off points for intra-rater reliability, inter-rater reliability and construct validity.

	Intra-rater reliability	Inter-rater reliability	Optimal convergent construct validity	Least convergent construct validity
good	$x \geq 0.85$	$x \geq 0.80$	$x \geq 0.65$	$x \geq 0.50$
moderate	$0.65 < x < 0.85$	$0.60 < x < 0.80$	$0.50 < x < 0.65$	$0.40 < x < 0.50$
poor	$x < 0.65$	$x < 0.60$	$x < 0.50$	$x < 0.40$

Legend: x = Pearson's r , Spearman's ρ or IntraClass Correlation Coefficient (ICC)

To investigate the influence of validating against different constructs (varying from optimally comparable to imperfectly comparable) the construct validity was divided into five clusters, in which the constructs against which a measurement instrument is validated are defined according to their anticipated degree of similarity to the instrument at issue (Table 4.2).

Table 4.2. Categorization of comparators utilized in assessing construct validity.

Construct level:	Definition:
Construct 1	Validation against sub-scales or instruments that measure the same impairment
Construct 2	Validation against instruments that measure the same impairment as well as other impairments
Construct 3	Validation against instruments that measure other impairments than the instrument to be validated
Construct 4	Validation against instruments that measure disabilities instead of the domain (impairments) to be validated
Construct 5	Validation against generic instruments that measure impairments as well as disabilities and participation problems

'Construct 1' is the most convergent construct, which means that the validity is measured against a variable which is very similar to the variable to be validated (for example: a pain-intensity questionnaire (variable pain) that is validated against a Visual Analogue Scale for pain). Validity is defined as 'construct 2' if the instrument is validated against instruments that measure the same construct as well as other impairments. 'Construct 3' indicates that the construct relates to other impairments than the impairment to be validated. 'Construct 4' relates to disabilities, and not to impairments. 'Construct 5' is the least convergent validity, which means that the construct that is used to validate a variable relates to other domains than the variable that is to be validated (for example: a pain-intensity questionnaire (variable pain) that is validated against disabilities in inter-personal relationships).

We clustered 'construct 1' and 'construct 2' as 'optimal convergent validity' and 'construct 3', 'construct 4' and 'construct 5' as 'imperfect convergent validity'. This distinction is also reflected in the criteria for validity in the two last columns of Table 4.1. The argument for this distinction is the fact that optimal convergent construct validity comes closest to the gold standard, and is therefore expected to result in higher values than the imperfect convergent validity (14). An explanation of all the abbreviations used in the Result section is given in Appendix 3.

Data-analysis

The data were analysed in the Statistical Package for Social Sciences (SPSS) 8.0. A classification of instruments was first made according to the type of impairments, based on the domain of assessment (impairments) (12).

The values of different studies of the same instrument were pooled for each psychometric property that was assessed (reliability and construct validity, related to sample size and correlation coefficients). By means of statistical pooling taking sample sizes into account the results of separate studies on reliability and constructed validity, expressed in correlation coefficients, were combined into a single index. When the relevant information was available, mathematical pooling of the data was performed if the measurement instrument was validated against the same construct. The values were pooled per construct. The pooled index is composed of measurements, appropriately weighted:

$$X = (\sum[n_i * x_i])/N$$

where X = pooled index, n_i = number of persons included in the study, x_i = value of methodological aspect (Pearson's r, Spearman's rho or Intra Class Correlation Coefficient (ICC)) in the study, N = total number of persons in all studies included in the pooling) (22). The pooled index was computed for Pearson's-/Spearman's correlation coefficients and the ICC values separately. Values for the construct validity of multidimensional instruments can be strongly influenced by values of one or more sub-scales. Therefore, whenever possible, the data were also pooled for the separate sub-scales.

Results

Table 4.3 presents the impairments in the ICDH-classification that are considered to be relevant for people with rheumatic disorders in allied health care settings. The number of identified instruments is also shown for each group of relevant impairments.

In total the search identified 49 measurement instruments for the assessment of impairments in functions in patients with rheumatic disorders (Table 4.3). Several of

these 49 instruments measure more than one impairment. Two adapted versions of the AIMS, a multidimensional questionnaire, are included: the AIMS2 and the AIMSS (for all abbreviations see the Appendix 3).

As can be observed in Table 4.3, the categories of impairments that are most frequently assessed by measurement instruments are pain (n=20; 41%), and impairments in musculoskeletal and movement related functions (n=16; 33%).

Table 4.3. Number of measurement instruments for each category of impairments.

	Number of measurement instruments/subscales
Impairments in mental functions	11
Impairments in sensation (stiffness)	3
Impairments in sensory functions (pain)	20
- experience of pain	19
- pain behaviour	1
Impairments in neuro musculo- skeletal- and movement related functions	15
- in mobility of joints	11
- in muscle force	3
- joint swelling	2

With regard to stiffness, the Visual Analogue Scale-Stiffness and Morning Stiffness-duration measure *only* stiffness. The BASDAI also measures other aspects of disease activity, such as (joint) pain, swelling and tiredness. Impairments in sensory functions are mainly assessed by instruments that focus on pain, with only one exception (the AIMS). Fifteen measurement instruments that were identified only measure impairments in musculoskeletal and movement-related functions. Six of these instruments have been specifically developed for assessing patients with ankylosing spondylitis.

Measurement of impairments in mental functions of patients with rheumatic disorders

Most instruments that measure impairments in mental functions of patients with rheumatic disorders (they all concern questionnaires) do not only measure these impairments, but also other impairments and/or disabilities. Four instruments that were identified focus only and specifically on impairments in mental functions (AHI, BDI, SSAI and STAI); the SF-36 and the AIMS and its modifications have only a sub-scale for measuring mental health and focus mainly on disabilities.

The AIMS, in different modifications, is the only questionnaire that meets the criterion for reliability (Table 4.4). For the most convergent construct validity (validated against an optimally comparable construct), two questionnaires meet the criterion: BDI

(pooled $r=0.67$), and STAI (pooled $r=0.68$). For the imperfect construct validity, another two questionnaires can be qualified as 'good': the AHI (pooled $r=0.69$) and the AIMS-depri (pooled $r=0.74$). None of the identified instruments had good reliability as well as good validity.

Measurement of impairments in sensation: stiffness

There are three available instruments to measure stiffness: BASDAI, MS-D, and VAS-S (Table 4.4). In all three the intra-rater reliability is good. There is a great variety in the validity of these instruments, with values that range from 0.26 to 0.70 if validated against instruments that measure the same impairment (Table 4.4). The highest value is found for the BASDAI ($r=0.70$).

Only the BASDAI and the VAS-S have good reliability as well as good validity (shown in grey in Table 4.4).

Measurement of impairments in sensory functions: pain

A total of twenty measurement instruments were identified for the assessment of pain in patients with rheumatic disorders (two modifications of the AIMS included). Eight out of these twenty only measure (modalities of) pain: AI, Dol, MPQ, OPB, P-NRS, RAPS, RPS and VAS-P. All the others are multidimensional instruments, or represent a separate sub-scale for pain in a multidimensional instrument. The results are presented in Table 4.4.

Reliability studies were performed for all twenty measurement instruments, with the exception of the EI, the P-NRS and the RPS. Five of the instruments showed good reliability: AIMS (pooled $r=0.86$), AIMS2-pain (ICC=0.89), ASES (pooled $r=0.88$), AUSCAN (ICC=0.89), J-MAP (internal consistency = 0.90) and RAPS (internal consistency = 0.92).

Seven instruments showed good convergent construct validity, with the highest values for the AI, the Dol and the VAS-P. Seven instruments were found to have imperfect construct validity, all by correlation with composite measurement instruments, combining impairments, disabilities and problems in participation in one measure. If reliability and construct validity are both required to be 'good', the AIMS and the AUSCAN-OHI are the only suitable instruments for pain assessment.

Measurement of impairments in joint mobility

For the assessment of joint mobility in patients with rheumatic disorders, eleven measurement instruments were identified (Table 4.4). On the basis of the available data, all instruments meet the criterion for intra-rater reliability, with the exception of the EDI. For the BASMI and the spondylometer there is no available information concerning reliability. The same was found for the inter-rater reliability, with the exception of Chest-expansion (ICC=0.53), and MobSpine-CCD (ICC=0.72). The

BASMI (pooled $r=0.93$), the goniometer ($r=0.92$) and the spondylometer ($r=0.92$) prove to be valid in patients with rheumatic disorders if validated against an optimally comparable construct. If the validity is based on correlation with an imperfect construct, the EPM and the Shober-test can also be qualified as valid. If both reliability and validity are required to be 'good', four measurement instruments meet the criteria: the chest-expansion, the EPM, the goniometer and the Shober-test.

Measurement of impairments in muscle force

Three instruments were identified to measure impairments in muscle force in patients with rheumatoid arthritis. The Grippit and the MSI are only intended to measure hand muscles, but the sphygmomanometer can also be used for other muscle groups. (Table 4.4).

Only the sphygmomanometer meets the criteria for both reliability and validity (reliability $r=0.93$; construct validity $r=0.87$ (pooled)).

Measurement of impairments in swelling

Two instruments are described for the assessment of (joint)swelling, both of which are subscales of an instrument to assess disease activity (Table 4.4).

The AI meets the criterion for inter-rater reliability (pooled $r =0.82$). The validity of the AI ($r=0.88$) and the SAJ ($r=-0.65$) was investigated. The AI was found to be the best instrument to assess joint swelling in patient with rheumatic disorders.

Table 4.4. Results for reliability and validity concerning measurement instruments per impairment.

Measurement instrument / subscale@	Intra-rater reliability	Inter-rater reliability	Construct validity if validated against instruments measuring		
			the same impairment	the same impairment as well as other impairments	disabilities impairments as well as disabilities and participation problems, or general aspects like, gender, age
Measurement of impairments in mental functions					
AHI ^(23;24)	0.53		0.46*		0.69*
AIMS ⁽²⁵⁻⁴⁶⁾	0.86*				0.23 ⁽²³⁾
AIMS-anxiety ⁽²⁵⁻⁴⁶⁾			0.43*		0.33 ⁽³⁴⁾
AIMS-depri ⁽²⁵⁻⁴⁶⁾					0.37 ⁽³⁶⁾
AIMS-Emof ⁽²⁵⁻⁴⁶⁾			0.43*		0.28 ⁽³⁶⁾
AIMSS-anxiety ⁽⁴⁰⁾			0.57*		0.19*
AIMSS-depri ⁽⁴⁰⁾			0.41*		0.74*
BDI ^(24;47)			0.47*		0.45*
SF36 ⁽⁴⁸⁻⁵²⁾	0.66#		0.67*		0.41*
SSAI ⁽²⁴⁾			0.36*		0.14*
STAI ⁽²⁴⁾			0.57*		0.27*
			0.68*		0.64*
					0.35*
					0.47 ⁽²⁴⁾
					0.57 ⁽²⁴⁾
					0.51*
Measurement of impairments in sensation (stiffness)					
EASDAI ⁽⁵³⁻⁵⁵⁾	0.74 ⁽⁵³⁻⁵⁵⁾		0.70 ⁽⁵³⁻⁵⁵⁾		0.57 ⁽⁵⁴⁾
MS-D ^(31;43;56;57)			0.26 ⁽⁵⁸⁾		0.67 ⁽⁵⁴⁾
VAS-S ^(31;43;56-59)	0.93		0.51 ⁽⁵⁸⁾		0.16*
					0.91*
Measurement of impairment in sensory functions (pain)					
ADLps ⁽⁶⁰⁾	0.31 ⁽⁶⁰⁾				
AI ⁽⁶¹⁻⁷²⁾	0.84*		0.83*		0.24*
	0.83##*				0.42 ⁽⁷²⁾
AIMS ⁽²⁵⁻⁴⁶⁾	0.86*		0.58*		0.75*
AIMS-pain ⁽²⁵⁻⁴⁶⁾	0.62		0.40*	0.54*	0.60*
AIMS2-pain ^(32;46;73)	0.89#			0.49 ⁽⁷³⁾	0.38*
AIMSS-pain ⁽⁴¹⁾	0.57 ⁽⁴¹⁾		0.41*		0.50*
ASES ⁽⁷⁴⁾	0.88*				0.61*
USCAN-OHI ⁽⁷⁵⁾	0.84		0.65		
DoI ⁽⁷⁶⁻⁸⁵⁾	0.81*		0.79*		
	0.74##*				
EI ⁽⁸⁶⁾			0.67 ⁽⁸⁶⁾		0.36*
FFI ⁽⁸⁷⁾	0.84##*				
J-MAP ⁽⁸⁸⁾	0.90		0.62		0.41
MPQ ^(37;89-92)	0.61*		0.37 ⁽⁹⁰⁾		
OPB ^(93;94)	0.78 ⁽⁹⁴⁾				
P-NRS ^(95;96)					0.32*
FAPS ⁽⁹⁷⁾	0.92		0.52		
			0.68		
RPS ⁽⁹⁸⁾			0.53		-0.43
SAJ ⁽⁷⁰⁾	0.82*		0.55*		
Stest ⁽⁹⁴⁾	0.67*				
VAS-P ^(31;43;57;76;95;96;99;100)	0.80*		0.82*		0.26*
	0.73##*				0.73 ⁽¹⁰⁰⁾

Measurement of impairments in joint mobility

BASMI ⁽¹⁰¹⁾			0.93*		
Chest ⁽¹⁰²⁾	0.95 ⁽¹⁰²⁾	0.53#		0.60 ⁽¹⁰²⁾	
EDI-abd ⁽¹⁰³⁾	0.63#				
EPM ⁽¹⁰⁴⁻¹⁰⁶⁾	0.87*	0.87*		0.54	0.54*
Conio ⁽¹⁰⁷⁾	0.89 ⁽¹⁰⁷⁾		0.92 ⁽¹⁰⁷⁾		
MKI ⁽¹⁰⁸⁾	0.90		-0.59		
MobSpine CCD ^(109;110)	0.93#	0.72#		0.37 ⁽¹¹⁰⁾	
MobSpine OWD ^{(109;110)96}	0.93#	0.92#		0.49 ⁽¹¹⁰⁾	
Shob ^(107;109)	0.95#	0.88#*		0.66*	
Spond ^(111;112)			0.92 ⁽¹¹¹⁾		
Stest ^(112;113)	0.97 ⁽¹¹³⁾				

Measurement of impairments in muscle force

Gripp ^(31;114-116)	0.86*				
MSI ⁽¹¹⁷⁾	0.84 ⁽¹¹⁷⁾	0.78 ⁽¹¹⁷⁾			
Shy ⁽¹¹⁸⁻¹²³⁾		0.93 ⁽¹²¹⁾	0.87*		

Measurement of joint swelling

AI ⁽⁶¹⁻⁷²⁾	0.80*	0.82*	0.88*		
	0.83#*				
SAJ ⁽⁷⁰⁾	0.77 ⁶³			-0.65 ⁽⁷⁰⁾	

Legend:

@ for explanation of abbreviations see Appendix 3.

All values expressed in Pearson's-r of Spearman's-r.

Intraclass Correlation Coefficient-value.

* pooled value.

In grey: instruments that meet the criteria for reliability as well as for validity.

The column 'if validated against instruments that measure other impairments than the instrument to be validated' is lacking in this Table because those studies were not appropriate here, because no data are retrieved on this.

Consequences of distinguishing convergent construct validity (validated against an optimal comparable construct) versus divergent construct validity (validated against an imperfect construct)

Summarizing the data on measurement instruments for relevant impairments, data on the validity of 40 instruments and/or sub-scales are available. For 21 out of these 40 measurement instruments, there are data on the optimally comparable construct validity as well as the imperfect construct validity. In fourteen of those 21 instruments the correlation values for the optimally comparable constructs proved to be higher than the values for validation against the imperfect constructs. In the studies, reporting that the validity based on the imperfect constructs was better, in four out of six it concerned the AIMS, or modifications or sub-scales of the AIMS.

Discussion

We identified 49 instruments for the assessment of impairments in mental functions, stiffness, pain, joint swelling, mobility and muscle strength in patients with rheumatic disorders. Sixteen of these instruments were found to have good reliability, and 24 were found to have good validity. Only nine out of the 49 measurement instruments had good reliability as well as good validity. None of the identified instruments for the assessment of impairments in mental functions were both reliable and valid. With regard to the other impairments, only 1-3 instruments per impairment met the criteria for reliability as well as validity. For assessment of stiffness the best instruments are the BASDAI and the VAS-S. For assessment of pain the best instrument is the AIMS. For joint mobility Chest, EPM and Gonio are the best available instruments, for muscle force the Sphy, and finally, for assessment of (joint)swelling the best instrument is the AI. The best available instruments are reflected in the grey rows in Table 4.4.

Clinical consequences of this study

Many measurement instruments are developed for the assessment of impairments in people with rheumatic disorders. For clinical use, unidimensionality, good reliability and good validity are required for a useful application. The question is not what kind of new instruments should be developed, but rather which of the existing measurement instruments should be rejected because of insufficient quality or lack of information about their psychometric properties. For at least seven instruments there is a lack of information concerning their reliability. This might be a consequence of the fact that this review was restricted to studies focusing on populations with rheumatic disorders. The reliability and validity of the instruments might well have been investigated in other patient populations. The results of this review show that six instruments had insufficient reliability as well as insufficient validity, so there is evidence to justify rejection of those six instruments for populations with rheumatic disorders: MS-D, ADLps, MPQ, P-NRS, RPS and EDI-abd.

Many questionnaires that are intended to be used by people with chronic diseases are multidimensional in order to get an overview of the impact of the disease as complete as possible. However, this multidimensionality makes validation of the questionnaire more complex. For unidimensional questionnaires it may be easier to find an optimal comparable construct to validate against than for a multidimensional questionnaire. For that reason we analysed data for the separate subscales if possible in order to study the influence of the validity of subscales on the validity of the total questionnaire. The assumption is that each subscale is unidimensional, which is the first requirement of a measurement scale and requires also a good internal consistency.

It is remarkable that unsatisfactory validity was found for several sub-scales of the AIMS, namely the AIMS-pain, AIMSS-pain, AIMS-anxiety, AIMS-depri and AIMS-emof. Those (sub-)scales were especially developed to assess a specific impairment, so one would expect them to have high correlation values if validated against an optimally comparable construct. The reliability of sub-scales of the AIMS has not been investigated. Research on this topic is restricted to the AIMS as a whole. On the basis of our results it could be concluded that the AIMS as a whole is reliable and valid, but that there is insufficient evidence for 'good' psychometric properties of separate sub-scales of the AIMS. Furthermore, restrictions to this conclusion must be made because of the fact that the conclusion regarding the validity of the AIMS as a whole is based on comparison with an imperfect construct. In fact, in general construct validity is a way of hypothesis testing, where the hypothesis is hidden in the hypothetical (theoretical) constructs. The hypothetical constructs contain proposed underlying factors (which we tried to classify in Table 4.2). Possibly the proposed underlying factors (which are also incorporated in the different items of a questionnaire) are not fully correct, what might also explain the found differences in correlations. In fact, the results of construct validity must be interpreted in a broader context than only the strength of correlation with other measures. It is often advocated in the literature that validity is also related to reliability and internal consistency, as well as to the proposed underlying factors in the hypothetical constructs. The results of our study should be interpreted in the light of this.

Despite the classification of constructs in 'convergent constructs' versus 'imperfect convergent constructs' and our expectation to find stronger correlations in validation against 'convergent constructs', our results did not confirm this for 100%. In particular some impairments-measures showed strong correlations when validated against disabilities. This could possibly be explained by the fact that some impairments are good predictors for certain disabilities. For example 'depression' (AIMS-depri) will have a relative strong impact on disabilities and participation problems (correlation $r=0.74$; Table 4.4).

The majority of instruments to measure joint mobility were found to have good intra- and/or inter-rater reliability. Some of the instruments are designed for specific groups of patients: for measuring mobility in patients with ankylosing spondylitis the most reliable and valid available instruments are the Chest-expansion measurement and the Shobert's test. For patients with rheumatoid arthritis the best option is the EPM-ROM. For all rheumatic disorders in general the best options are the goniometer, the spondylometer and the BASMI.

The sphygmomanometer is the most reliable and valid instrument to assess muscle strength. It is mainly used for the assessment of grip strength, but can also be

applied to assess the strength of other muscles (119). Particularly for the assessment of grip-strength the sphygmomanometer can measure isometric muscle strength reliably and quickly.

In this study we did not report the responsiveness of the identified instruments. So far, research into the responsiveness of measurement instruments is hampered by the lack of consensus about the preferred method. Further research is needed to investigate the responsiveness of the identified instruments in the assessment of impairments.

In this study, several levels of construct validity were distinguished. We only accepted a value of ≥ 0.50 as 'good' if validated against an imperfect construct, and a value of ≥ 0.65 as 'good' if validated against an optimally comparable construct. Our cut-off points are based on the literature, but in fact many authors seem to deviate from literature when interpreting the results of their own study (106;124-126). To our knowledge, different cut-off points for optimal convergent validity versus imperfect convergent validity have never before been applied in validation studies. The results of our study demonstrate that the use of different cut-off points is justified; most measurement instruments showed better correlations with optimally comparable constructs than with imperfect constructs, and this information is relevant for the interpretation of validity studies in general. However, the levels of the cut-off points applied in this study remain arbitrary. Regarding correct interpretation of the cut-off points for 'good' reliability, it must be emphasized that these are also depending on the sample size, since a sample size of 1000 can tolerate a much less reliable instrument than a sample of 10 (21). The aim of our study was to give an overview of reliability of instruments, intended for individual use in clinical situations. As a consequence of pooling data the total number of patients generating our results is large, therefore a high cut-off point is required to justify extrapolation to level of individuals.

The workgroup Outcome Measures in Rheumatoid Arthritis Clinical Trials (OMERACT) also emphasizes the importance of the methodological quality of measurement instruments. OMERACT established a core set of eight end-points that should be evaluated in clinical trials (disease activity; disability; pain; patient global assessment; physician global assessment; swollen joint; tender joint; joint imaging) (127;128). The instruments discussed in this article focus on two of these domains (pain and the number of tender joints). The results of our study might also be useful in selecting the most appropriate measurement instrument(s) to assess joint swelling and pain.

In summary, evidence of both reliability and validity was only found for eleven out of 49 instruments for the assessment of impairments in patients with rheumatic disorders. Those eleven best instruments are reflected in grey in Table 4.4. In six instruments neither the reliability nor the validity was found to be adequate. Evidence concerning the reliability and validity of sub-scales of the AIMS is lacking.

Acknowledgment

We wish to thank Mrs. Y. Kappe (YK) for reviewing the majority of the articles, together with the first author.

Reference List

1. Rimmer JH. Health promotion for people with disabilities: the emerging paradigm shift from disability prevention to prevention of secondary conditions. *Physical Therapy*; 1999;79:495-502.
2. World Health Organization. *International Classification of Impairments, Disabilities and Handicaps: a manual of classification relating to the consequences of disease*. 1980 Geneva, World Health Organisation.
3. World Health Organization (2001). *International Classification of Functioning, Disabilities and Health Problems*. 2001 Geneva, World Health Organization.
4. Feinstein AR. *Clinimetrics*. 1987 New Haven: Yale University Press.
5. Hendriks HJM, Brandsma JW, Heerkens YF, Oostendorp RAB, Nelson RM. Intraobserver and interobserver reliability of assessments of impairments and disabilities. *Phys Ther* 1997;77:1097-106.
6. Kirschner B, Guyatt GH. A methodological framework for assessing health indices. *J Chron Dis* 1985;38:27-36.
7. Bombardier C, Tugwell P. Methodological considerations in functional assessment. *J Rheumatol* 1987;14 Suppl 15:6-10.
8. Tugwell P, Bombardier C. A methodologic framework for developing and selecting endpoints in clinical trials. *J Rheumatol* 1982;9:758-62.
9. Beckerman H, Bouter LM. Efficacy of physiotherapy for musculoskeletal disorders: what can we learn from research? *Br J Gen Pract* 1993;43:483-91.
10. McDowell, I. and Newell, C. *Measuring health: a guide to rating scales and questionnaires*. 1987 New York, Oxford University Press.
11. Streiner DL, Norman GR. *Health Measurement Scales*. In: Streiner DL, Norman GR, eds. 1999 Oxford: Oxford University Press 1-68.
12. World Health Organization. *International Classification of Functioning and Disability. Beta-2 draft, short version*. 1999 Geneva, World Health Organization.
13. Swinkels RAHM. The ICF-classification as a system for structuring outcome assessment. *Physiotherapy Singapore* 2004;7(3):7-13.
14. Swinkels RAHM, Bouter, Swinkels-Meewisse EJCM, LM, Oostendorp RAB, Dijkstra PU, Vet HCW de. Construct validity of impairment measures in rheumatic disorders: which constructs are selected for validation? *Clin Exp Rheum* 2005; in press.
15. World Health Organization. *Adaptation of the classification of impairments for five Dutch allied health professions*. 1995 Geneva. World Health Organization.
16. Oostendorp RAB, Elvers JWH, and Swinkels RAHM. *Judgement list for reading articles*. 1997, Unpublished.
17. DeJong A, Molenaar IW. An application of Mokken's model for stochastic cumulative scaling in psychiatric research. *J.Psychiatr Res* 1987;21:137-49.
18. Eliasziw M, Young SL, Woodbury MG, Fryday-Field K. *Statistical Methodology for the Concurrent Assessment of Interrater and Intrarater Reliability: Using Goniometric Measurements as an Example*. *Physical Therapy* 1994;74:777-88.
19. Doeglas D, Krol B, et al. The assessment of functional status in rheumatoid arthritis: a cross cultural longitudinal comparison of the Health Assessment Questionnaire and the Groningen Activity Restriction Scale. *J Rheumatol* 1995;22;10:1834-1843.
20. Weiner E, Stewart B. *Assessing individuals*. 1984 Boston: Little Brown.
21. Streiner DL, Norman GR. *Health Status Measurement Scales. A practical guide to their development and use*. 2003 Oxford: Oxford University Press.
22. Neter J, Wasserman W, Whitmore GA. *Applied statistics*. 1982 Boston: Allyn and Bacon, Inc.
23. Nicassio PM, Wallston KA, Callahan LF, Herbert M, Pincus T. The measurement of helplessness in rheumatoid arthritis. The development of the arthritis helplessness index. *J Rheumatol* 1985;12:462-7.
24. Hagglund KJ, Roth DL, Haley WE, Alarcon GS. Discriminant and convergent validity of self-report measures of affective distress in patients with rheumatoid arthritis. *J.of Rheumatology* 1989;16:1428-32.

25. Meenan RF, Anderson JJ, Kazis LE. Outcome assessment in clinical trials. Evidence for the sensitivity of a health status measure. *Arthritis Rheum* 1984;27:1344-52.
26. Meenan RF, Gertman PM, Mason JH. Measuring health status in arthritis. The arthritis impact measurement scales. *Arthritis Rheum* 1980;23:146-52.
27. Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. *Med.Care* 1989;27:S178-S189.
28. Anderson JJ, Firschein HE, Meenan RF. Sensitivity of a health status measure to short-term clinical changes in arthritis. *Arthritis Rheum* 1989;32:844-50.
29. Hidding A, Santen Mv, Klerk Ed, Gielen X, Boers M. Discordance between self-report measures and clinical observations of functional disability in ankylosing spondylitis, rheumatoid arthritis and fibromyalgia. Group physical therapy in ankylosing spondylitis (thesis). 1993 Meppel: Krips Repro 94-104.
30. 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.
31. Anderson JJ, Chernoff MC. Sensitivity to change of rheumatoid arthritis clinical trial outcome measures. *J Rheumatol* 1993;20:535-7.
32. Potts MK, Brandt KD. Evidence of the validity of the Arthritis Impact Measurement Scales. *Arthritis Rheum* 1987;30:93-6.
33. 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.
34. Fitzpatrick R, Ziebland S, Jenkinson C, Mowat A, Mowat A. A comparison of the sensitivity to change of several health status instruments in rheumatoid arthritis. *J Rheumatol* 1993;20:429-36.
35. Liang MH, Larson MG, Cullen KE, Schwartz JA. Comparative measurement efficiency and sensitivity of five health status instruments for arthritis research. *Arthritis Rheum* 1985;28:542-7.
36. Meenan RF, Gertman PM, Mason JH, Dunaif R. The arthritis impact measurement scales. Further investigations of a health status measure. *Arthritis Rheum* 1982;25:1048-53.
37. Keefe FJ, Caldwell DS, Queen KT. Pain coping strategies in osteoarthritis patients. *J Consult Clin Psychol.* 1987;55:208-12.
38. Liang MH, Fossel AH, Larson MG. Comparisons of five health status instruments for orthopedic evaluation. *Med Care* 1990;28:632-42.
39. Goepfing J, Doyle MA, Charlton SL, Lorig K. A nursing perspective on the assessment of function in persons with arthritis. *Res Nurs Health* 1988;11:321-31.
40. Jacobs JW, Oosterveld FG, Deurbouts N. Opinions of patients with rheumatoid arthritis about their own functional capacity: how valid is it? *Ann Rheum Dis.* 1992;51:765-8.
41. Wallston KA, Brown GK, Stein MJ, Dobbins CJ. Comparing the short and long versions of the Arthritis Impact Measurement Scales. *J Rheumatol* 1989;16:1105-9.
42. Duffy CM, Watanabe Duffy KN, Gladman DD. The utility of the arthritis impact measurement scales for patients with psoriatic arthritis. *J Rheumatol* 1992;19:1727-32.
43. Hawley DJ, Wolfe F. Sensitivity to change of the health assessment questionnaire (HAQ) and other clinical and health status measures in rheumatoid arthritis: results of short-term clinical trials and observational studies versus long-term observational studies. *Arthritis Care Res* 1992;5:130-6.
44. Vliet Vlieland TP, van der Wijk TP, Jolie IM, Zwinderman AH, Hazes JM. Determinants of hand function in patients with rheumatoid arthritis. *J Rheumatol* 1996;23:835-40.
45. Weinberger M, Samsa GP, Tierney WM, Belyea MJ, Hiner SL. Generic versus disease specific health status measures: comparing the sickness impact profile and the arthritis impact measurement scales. *J Rheumatol* 1992;19:543-6.

46. Husted J, Gladman DD, Farewell VT, Long JA. Validation of the revised and expanded version of the Arthritis Impact Measurement Scales for patients with psoriatic Arthritis. *J Rheumatol* 1996;23:1015-9.
47. Callahan LF, Kaplan MR, Pincus T. The Beck Depression Inventory, Center for Epidemiological Studies Depression Scale (CES-D), and General Well-Being Schedule depression subscale in rheumatoid arthritis. Criterion contamination of responses. *Arthritis Care Res.* 1991;4:3-11.
48. Martin DP, Engelberg R, Agel J, Swiontkowski MF. Comparison of the Musculoskeletal Function Assessment Questionnaire with the Short Form-36, the Western Ontario and McMaster University Osteoarthritis Index, and the Sickness Impact Profile Health-status measures. *J Bone Joint Surg.* 1997;79A:1323-1335.
49. Wright JG, Young NL. A comparison of different indices of responsiveness. *J Clin Epidemiol* 1997;50:239-46.
50. Husted JA, Gladman DD, Farewell VT, Long JA, Cook RJ. Validating the SF-36 Health Survey Questionnaire in patients with psoriatic arthritis. *J Rheumatol* 1997;24:511-7.
51. Kosinski M, Keller SD, Hatoum HT, Kong SX, Ware JE, Jr. The SF-36 Health Survey as a generic outcome measure in clinical trials of patients with osteoarthritis and rheumatoid arthritis: tests of data quality, scaling assumptions and score reliability. *Med Care* 1999;37:MS10-MS22.
52. Soderman P, Malchau H. Validity and reliability of Swedish WOMAC osteoarthritis index: a self-administered disease-specific questionnaire (WOMAC) versus generic instruments (SF-36 and NHP). *Acta Orthop Scand.* 2000;71:39-46.
53. Garrett S, Jenkinson T, Kennedy LG, Whitelock H, Gaisford P, Calin A. A new approach to defining disease status in ankylosing spondylitis: the Bath Ankylosing Spondylitis Disease Activity Index. *J Rheumatol* 1994;21:2286-91.
54. Van Tubergen A, Coenen J, Landewe R. Assessment of fatigue in patients with ankylosing spondylitis: a psychometric analysis. *Arthritis Rheum* 2002;47:8-16.
55. Van Tubergen A, Debats I, Ryser L. Use of a numerical rating scale as an answer modality in ankylosing spondylitis-specific questionnaires. *Arthritis Rheum* 2002;47:242-8.
56. Rhind VM, Bird HA, Wright V. A comparison of clinical assessments of disease activity in rheumatoid arthritis. *Ann Rheum Dis.* 1980;39:135-7.
57. Mallya RK, Mace BE. The assessment of disease activity in rheumatoid arthritis using a multivariate analysis. *Rheumatol Rehabil.* 1981;20:14-7.
58. Vliet-Vlieland TPM, Zwiderman AH, Breedveld FC, Hazes JMW. Measurement of morning stiffness in rheumatoid arthritis clinical trials. *J Clin Epidemiol* 1997;50:6.
59. Oosterhof, J. Research to the reproducibility of the VAS for stiffness in patients with rheumatoid arthritis. 1992, Unpublished.
60. Callahan LF, Brooks RH, Summey JA, Pincus T. Quantitative pain assessment for routine care of rheumatoid arthritis patients, using a pain scale based on activities of daily living and a visual analog pain scale. *Arthritis Rheum* 1987;30:630-6.
61. Ritchie DM, Boyle JA, McInnes JM. Clinical studies with an articular index for the assessment of joint tenderness in patients with rheumatoid arthritis. *Q.J Med.* 1968;37:393-406.
62. Camp AV. An articular index for the assessment of rheumatoid arthritis. *Orthopaedics* 1971;4:39-45.
63. Dougados M, Gueguen A, Nakache JP, Nguyen M, Mery C, Amor B. Evaluation of a functional index and an articular index in ankylosing spondylitis. *J Rheumatol* 1988;15:302-7.
64. Prevoo ML, Riel PLCM van, Hof Mavan 't. Validity and reliability of joint counts. A longitudinal study in patients with recent onset rheumatoid arthritis. In: Prevoo ML, ed. *Clinical assessment in rheumatoid arthritis: aspects of validity and reproducibility.* Nijmegen: Catholic University Nijmegen 1995:19-31.
65. Hart LE, Tugwell P, Buchanan WW, Norman GR, Grace EM, Southwell D. Grading of tenderness as a source of interrater error in the Ritchie articular index. *J Rheumatol* 1985;12:716-7.

66. Doyle DV, Dieppe PA, Scott J, Huskisson EC. An articular index for the assessment of osteoarthritis. *Ann Rheum Dis*. 1981;40:75-8.
67. Gaston-Johansson F, Gustafsson M. Rheumatoid arthritis: determination of pain characteristics and comparison of RAI and VAS in its measurement. *Pain* 1990;41:35-40.
68. Kalla AA, Kotze TJ, Meyers OL, Parkyn ND. Clinical assessment of disease activity in rheumatoid arthritis: evaluation of a functional test. *Ann Rheum Dis*. 1988;47:773-9.
69. Fuchs HA, Brooks RH, Callahan LF, Pincus T. A simplified twenty-eight-joint quantitative articular index in rheumatoid arthritis. *Arthritis Rheum* 1989;32:531-7.
70. Prevo ML, Kuper IH, Hof MA van't, Leeuwen MA van, Putte LB van de, Riel PL van. Validity and reproducibility of self-administered joint counts. A prospective longitudinal follow-up study in patients with rheumatoid arthritis. *J Rheumatol* 1996;23:841-5.
71. Smolen JS, Breedveld FC, Eberl G. Validity and reliability of the twenty-eight-joint count for the assessment of rheumatoid arthritis activity. *Arthritis Rheum* 1995;38:38-43.
72. Orces CH, Del R, I, Abel MP, Escalante A. The number of deformed joints as a surrogate measure of damage in rheumatoid arthritis. *Arthritis Rheum* 2002;47:67-72.
73. Riemsma RP, Taal E, Rasker JJ, Houtman PM, Paassen HC van, Wiegman O. Evaluation of a Dutch version of the AIMS2 for patients with rheumatoid arthritis. *Br J Rheumatol* 1996;35:755-60.
74. Lorig K, Chastain RL, Ung E, Shoor S, Holman HR. Development and evaluation of a scale to measure perceived self-efficacy in people with arthritis. *Arthritis Rheum* 1989;32:37-44.
75. Bellamy N, Campbell J, Haraoui B. Clinimetric properties of the AUSCAN Osteoarthritis Hand Index: an evaluation of reliability, validity and responsiveness. *Osteoarthritis Cartilage* 2002;10:863-9.
76. Wigers SH, Skrandal A, Finset A, Gøtestam KG. Measuring change in fibromyalgic pain: the relevance of pain distribution. *J Musculoskeletal Pain* 1997;5:29-41.
77. McCarthy DJ, Jr., Gatter RA, Phelps P. A dolorimeter for quantification of articular tenderness. *Arthritis Rheum* 1965;44:551-9.
78. Wessel J. The reliability and validity of pain threshold measurements in osteoarthritis of the knee. *Scand.J Rheumatol* 1995;24:238-42.
79. Puttick M, Schulzer M, Klinkhoff A, Koehler B, Rangno K, Chalmers A. Reliability and reproducibility of fibromyalgic tenderness, measurement by electronic and mechanical dolorimeters. *J Musculoskeletal Pain* 1995;3:3-14.
80. Atkins CJ, Zielinski A, Klinkhoff AV. An electronic method for measuring joint tenderness in rheumatoid arthritis. *Arthritis Rheum* 1992;35:407-10.
81. Guthrie NG, Edworthy SM, Martin L, Fritzier MJ, Kinsella TD, Ryan JP. Comparing three techniques for the evaluation of fibromyalgia. *Physiotherapy Canada* 1994;46:94-7.
82. Daltroy LH, Larson MG, Roberts NW, Liang MH. A modification of the Health Assessment Questionnaire for the spondyloarthropathies. *J Rheumatol* 1990;17:946-50.
83. Smythe HA, Buskila D, Urowitz S, Langevitz P. Control and "fibrositic" tenderness: comparison of two dolorimeters. *J Rheumatol* 1992;19:768-71.
84. Tunks E, Crook J, Norman G, Kalaher S. Tender points in fibromyalgia. *Pain* 1988;34:11-9.
85. Langley GB, Fowles M, Sheppard H, Wigley RD. A simple pressure dolorimeter for the quantification of joint tenderness in inflammatory arthritis. *Rheumatol Int* 1983;3:109-12.
86. Mander M, Simpson JM, McLellan A, Walker D, Goodacre JA, Dick WC. Studies with an entheses index as a method of clinical assessment in ankylosing spondylitis. *Ann Rheum Dis* 1987;46:197-202.
87. Saag KG, Saltzman CL, Brown CK, Budiman-Mak E. The Foot Function Index for measuring rheumatoid arthritis pain: evaluating side-to-side reliability. *Foot Ankle Int* 1996;17:506-10.

88. O'Malley KJ, Suarez-Almazor M, Aniol J. Joint-specific multidimensional assessment of pain (J-MAP): factor structure, reliability, validity, and responsiveness in patients with knee osteoarthritis. *J Rheumatol* 2003;30:534-43.
89. Lankveld W van, Pad-Bosch P van 't, Putte LBA van de, Staak C van de, Näring G. Pijn bij reumatoïde artritis gemeten met de visuele analoge schaal en de Nederlandse versie van de 'McGill Pain Questionnaire'. *Ned Tijdschr Geneeskd.* 1992;136:1166-70.
90. Papageorgiou AC, Badley EM. (1989) The quality of pain in arthritis: the words patients use to describe overall pain and pain in individual joints at rest and on movement. *J Rheumatol* 16:106-12.
91. Burckhardt CS, Bjelle A. A Swedish version of the short-form McGill Pain Questionnaire. *Scand.J Rheumatol* 1994;23:77-81.
92. Melzack R. The McGill Pain Questionnaire: major properties and scoring methods. *Pain* 1:1975;277-99.
93. Anderson KO, Bradley LA, McDaniel LK. The assessment of pain in rheumatoid arthritis: disease differentiation and temporal stability of a behavioral observation method. *J Rheumatol* 1987;14:700-4.
94. Molendijk M, Colard I, Jacobs JWG, Oosterveld FGJ. Drukpunten bij het fibromyalgie syndroom; de betrouwbaarheid in de dagelijkse praktijk. *Ned Tijdschr Fysioth* 1991;101:31-6.
95. Downie WW, Leatham PA, Rhind VM, Wright V, Branco JA, Anderson JA. Studies with pain rating scales. *Annals of the Rheumatic Diseases* 1978;37:378-81.
96. Langley GB, Sheppaerd H. Problems associated with pain measurement in arthritis: comparison of the visual analogue and verbal rating scales. *Clin Exp Rheumatol* 1984;2:231-4.
97. Anderson DL. Development of an instrument to measure pain in rheumatoid arthritis: Rheumatoid Arthritis Pain Scale (RAPS). *Arthritis Rheum* 2001;45:317-23.
98. Wolfe F. Pain extent and diagnosis: development and validation of the regional pain scale in 12,799 patients with rheumatic disease. *J Rheumatol* 2003;30:369-78.
99. Hurst NP, Kind P, Ruta D, Hunter M, Stubbings A. Measuring health-related quality of life in rheumatoid arthritis: validity, responsiveness and reliability of EuroQol (EQ-5D). *Br J Rheumatol* 1997;36:551-9.
100. Tamiya N, Araki S, Ohi G. Assessment of pain, depression, and anxiety by visual analogue scale in Japanese women with rheumatoid arthritis. *Scand J Caring. Sci.* 2002;16:137-41.
101. Jenkinson TR, Mallorie PA, Whitelock HC, Kennedy LG, Garrett SL, Calin A. Defining spinal mobility in ankylosing spondylitis (AS). The Bath AS Metrology Index. *J Rheumatol* 1994;21:1694-8.
102. Moll JM, Wright V. The pattern of chest and spinal mobility in ankylosing spondylitis. An objective clinical study of 106 patients. *Rheumatol Rehabil* 1973;12:115-34.
103. Heemskerk MAMB, Aarst Mv, Windt DAWM. De reproduceerbaarheid van het meten van de passieve beweeglijkheid van de schouder met de EDI-320 digitale hoekmeter. *Ned Tijdschr Fysioth* 1997;107:146-9.
104. Ende CHM van den, Vliet-Vlieland TPM, Breedveld FC, Hazes JM. De EPM-ROM schaal: een waardevol meetinstrument voor de beweeglijkheid van de gewrichten van patiënten met reumatoïde artritis. *Ned Tijdschr Fysioth* 1994;104:86-91.
105. Vlieland TP, Ende CH van den, Breedveld FC, Hazes JM. Evaluation of joint mobility in rheumatoid arthritis trials: the value of the EPM-range of motion scale. *J Rheumatol* 1993;20:2010-4.
106. Ferraz MB, Oliveira LM, Araujo PM, Atra E, Walter SD. EPM-ROM Scale: an evaluative instrument to be used in rheumatoid arthritis trials. *Clin Exp.Rheumatol* 1990;8:491-4.
107. Reynolds PM. Measurement of spinal mobility: a comparison of three methods. *Rheumatol Rehabil* 1975;14:180-5.
108. Lefevre-Colau MM, Poiraudau S, Oberlin C. Reliability, validity, and responsiveness of the modified Kapandji index for assessment of functional mobility of the rheumatoid hand. *Arch.Phys Med.Rehabil* 2003;84:1032-8.

109. Viitanen JV, Kautiainen H, Suni J, Kokko ML, Lehtinen K. The relative value of spinal and thoracic mobility measurements in ankylosing spondylitis. *Scand J Rheumatol* 1995;24:94-7.
110. Viitanen JV. Thoracolumbar rotation in ankylosing spondylitis. A new noninvasive measurement method. *Spine* 1993;18:880-3.
111. Sturrock RD, Wojtulewski JA, Hart FD. Spondylometry in a normal population and in ankylosing spondylitis. *Rheumatol Rehabil* 1973;12:135-42.
112. Roberts WN, Liang MH, Pallozzi LM, Daltroy LH. Effects of warming up on reliability of anthropometric techniques in ankylosing spondylitis. *Arthritis Rheum* 1988;31:549-52.
113. Newcomer KL, Krug HE, Mahowald ML. Validity and reliability of the timed-stands test for patients with rheumatoid arthritis and other chronic diseases. *J Rheumatol* 1993;20:21-7.
114. Nordenskiöld UM, Grimby G. Grip force in patients with rheumatoid arthritis and fibromyalgia and in healthy subjects. A study with the Grippit instrument. *Scand J Rheumatol* 1993;22:14-9.
115. Dellhag B, Bjelle A. A Grip Ability Test for use in rheumatology practice. *J Rheumatol* 1995;22:1559-65.
116. Dellhag B, Hosseini N, Bremell T, Ingvarsson PE. Disturbed grip function in women with rheumatoid arthritis. *J Rheumatol* 2001;28:2624-33.
117. Stucki G, Schonbachler J, Bruhlmann P, Mariacher S, Stoll T, Michel BA. Does a muscle strength index provide complementary information to traditional disease activity variables in patients with rheumatoid arthritis? *J Rheumatol* 1994;21:2200-5.
118. Hamilton GF, McDonald C, Chenier TC. Measurement of grip strength: validity and reliability of the sphygmomanometer and Jamar Grip Dynamometer. *J Orthop Sports Phys Ther* 1992;16:215-9.
119. Helewa A, Goldsmith CH, Smythe HA. The modified sphygmomanometer-an instrument to measure muscle strength: a validation study. *J Chronic Dis* 1981;34:353-61.
120. Helewa A, Goldsmith CH, Smythe HA. Patient, observer and instrument variation in the measurement of strength of shoulder abductor muscles in patients with rheumatoid arthritis using a modified sphygmomanometer. *J Rheumatol* 1986;13:1044-9.
121. Spiegel JS, Paulus HE, Ward NB, Spiegel TM, Leake B, Kane RL. What are we measuring? An examination of walk time and grip strength. *J Rheumatol* 1987;14:80-6.
122. Lee P, Baxter A, Dick WC, Webb J. An assessment of grip strength measurement in rheumatoid arthritis. *Scand J Rheumatol* 1974;3:17-23.
123. Pincus T, Brooks RH, Callahan LF. Reliability of grip strength, walking time and button test performed according to a standard protocol. *J Rheumatol* 1991;18:997-1000.
124. Ghiselle EE. The measurement of occupational aptitude. 1966 Los Angeles: Berkely.
125. Huiskes CJAE, Kraaimaat FW, Bijlsma JWJ. Development of a self-report questionnaire to assess the impact of rheumatic diseases on health and lifestyle. *J Rehabil Sciences* 1990;3:65-70.
126. Stewart, AL. and Ware, JE. Measuring functioning and well-being: the Medical Outcomes Study Approach. 1992 Durham, North Carolina, Duke University Press.
127. Tugwell P, Boers M. OMERACT conference on outcome measures in rheumatoid arthritis clinical trials: conclusion. *J Rheumatol* 1993;20:590.
128. Wolfe F, Lassere M, Heijde D van der. Preliminary core set of domains and reporting requirements for longitudinal observational studies in rheumatology. *J Rheumatol* 1999;26:484-9.

Chapter 5

Which are the best instruments for measuring gait and gait-related activities in patients with rheumatic disorders? A systematic review.

Published as: Swinkels RAHM, Oostendorp RAB, Bouter LM. Which are the best instruments for measuring disabilities in gait and gait related activities in patients with rheumatic disorders? Clin Exp Rheum 2004;1:25-33

ABSTRACT*Objectives*

The first objective was to make an inventory of available instruments for the assessment of disabilities in gait and gait-related activities in patients with rheumatic disorders. Our second aim was to investigate which of these instruments have acceptable methodological quality with regard to reliability and validity. Our third aim was to investigate the assumption that the evaluation of convergent construct validity results in stronger correlations when validated against a more similar construct.

Methods and materials

A computer-aided literature search (1982-2001) in several databases was performed to identify studies focusing on the clinimetric properties of instruments to assess impairments in function in patients with rheumatic disorders. Data on intra-rater reliability, inter-rater reliability and convergent construct validity were extracted in a standardized way and compared to a priori defined criteria.

Results

In totally 78 instruments were eligible. Intra-observer reliability is investigated for 28 instruments. Only seven instruments demonstrated good reliability as well as good validity. Surprisingly, the convergent construct validation against a similar construct resulted often in lower correlations than validation against a less similar construct.

Conclusion

Based on the available information, the Rheumatoid Arthritis Quality of Life Scale and the Health Assessment Questionnaire seem to be the best instruments.

Introduction

In rehabilitation and allied health care for patients with rheumatic disorders, attention has shifted from disease severity to impairments, disabilities and problems in participation (1). In particular, the development of the *International Classification of Impairments, Disabilities and Handicaps* (ICIDH) (2), after its revision renamed as the *International Classification of Functioning, Disabilities and Health Problems* (ICF) (3) has encouraged this shift.

The main reason of this is that in rheumatology and in allied health care suitable treatment goals cannot be directly derived from diseases, but should be mainly based on the consequences of diseases (4,5). The ICF-classification distinguishes the disease itself from disease consequences and influencing contextual factors (4). If the treatment goals are expressed in terms of impairment, disability or problems in participation (as in the ICF), these must be demonstrated as reliable and valid (6,7).

From this point of view it is essential to choose adequate measurement instruments, based on evidence on their methodological quality (6,8-13).

This article focuses on the clinimetric properties of instruments to assess disabilities in gait and gait-related activities (running, moving around on different types of surfaces, climbing stairs and handling special aids for transportation) in patients with rheumatic disorders. The reason to focus particularly on disabilities in gait and gait-related activities is that this represents an essential part of health-related quality of life for these patients in the sense of their being able to manage activities independently. Therefore, improvement of the independency of the patient and the quality of gait and gait-related activities is frequently included as treatment goal for patients with rheumatic disorders.

The clinical variables that are examined need to be relevant for the treatment goals and need to be coherent to the prognosis of the disease (8-10, 14-21). From this point of view it is essential to use the correct and the most optimally measurement instrument(s) for assessment of aspects of gait and gait-related activities. We recently demonstrated that 37% of all 209 available instruments to assess people with rheumatic disorders focus on disabilities, partly on disabilities in gait and gait-related activities. A majority of 57% of all available instruments to assess people with rheumatic disorders focus on impairments (4).

In the ideal situation the choice for use of a measurement instrument should be based on the methodological quality of the available instruments (6, 8-13). However, an earlier study has demonstrated a lack of information about the methodological quality of instruments in use for the assessment of musculoskeletal disorders (11).

Reliability is a basic characteristic of an instrument that is required to be adequate: if an instrument is not reliable it will not be useful. If the reliability insufficient, it is also

relevant to assess the validity. In the literature, several aspects of validity are described, but because of frequent lack of a gold standard the construct validity is the most commonly assessed aspect. An instrument can be validated by studying its correlation with a similar (optimally comparable) construct (for instance the same disability) or by correlation with a less similar construct, such as age, gender, etc. In 73% of validity studies concerning construct validity, the validity is based on correlation with constructs that measure a totally different entity, and in 11% the validity values are based on correlation with a similar construct (22).

In this study we distinguished various levels of construct validity, because we expected to find stronger correlations if measurement instruments were validated against a similar construct than if validated against a less similar construct (22). If this were found to be so, this could be relevant for the interpretation of validity studies.

The aim of this article was threefold. The first objective was to make an inventory of available instruments for the assessment of disabilities in gait and gait-related activities in patients with rheumatic disorders. Our second aim was to investigate which of these instruments have acceptable methodological quality with regard to reliability and validity. Our third aim was to investigate the assumption that evaluation of convergent construct validity results in stronger correlations when validated against a dissimilar construct.

Method and materials

We performed a systematic review of the literature on the clinimetric properties of measurement instruments for all disabilities in gait and gait-related activities that are relevant in allied health care for patients with rheumatic disorders.

Inclusion procedure

The following criteria were applied:

- All studies had to focus on patients suffering from rheumatoid arthritis, seronegative polyarthritis (including psoriatic arthritis), osteoarthritis, ankylosing spondylitis, polymyositis or fibromyalgia.
- The studies had to contain information about the clinimetric properties of instruments to assess disabilities in gait and mobility.
- Different versions of an instrument were considered as separate measurement instruments.
- The questionnaire should contain at least one question or a separate subscale regarding gait or gait-related activities (running, moving around on different types of surfaces, climbing stairs and handling special aids for transportation).

- Only instruments for the measurement of adult patients were included.

Literature search

First the Medline database was searched for the period January 1982 - April 2001, using search terms for the relevant rheumatic disorders and various search terms for clinimetric properties (The detailed search strategy is available on request from the corresponding author). The database of the Centre for Documentation of the Dutch Institute of Allied Health Care was also searched for the period January 1988 – April 2001, using the same keywords. Furthermore, the search in both data-bases was repeated with the names of the identified measurement instruments. English, French, German and Dutch literature was included. The search was subsequently augmented with a manual search based on the references of the relevant publications, and therefore the search also yielded also in some publications from before 1982.

Data-extraction

All identified publications were assessed independently on the basis of their title and abstract by two reviewers (RS and YK). In case of disagreement (3%) the article was also assessed by a third reviewer. The assessment was based on a standardized data-collection form (23). This form consists of four sections: general description (name, first author, etc.), assessment domain (according to the ICDH-classification), methodological aspects (concerning reliability, validity, responsiveness) and aspects of feasibility. An explanation of all the abbreviations used in this article is given in the Appendix 3.

Methodological criteria for clinimetric properties

We investigated the following clinimetric properties: intra-rater reliability, construct validity and responsiveness. To interpret the data on reliability and validity we used criteria based on De Jong et al. (24), Eliasziw et al. (25) and Doeglas et al. (26) (Table 5.1).

Table 5.1. Cut-off-points used for intra-rater reliability, inter-rater reliability and construct validity.

	Intra-rater reliability	validation against similar construct	Validation against dissimilar construct
good	$x \geq 0.85$	$x \geq 0.65$	$x \geq 0.50$
moderate	$0.65 < x < 0.85$	$0.50 < x < 0.65$	$0.40 < x < 0.50$
poor	$x < 0.65$	$x < 0.50$	$x < 0.40$

Legend: x = Pearson's r, Spearman's rho or IntraClass Correlation Coefficient (ICC)

To investigate the influence of validating against different constructs (varying from optimally comparable to dissimilar) the construct validity was divided into four clusters, in which the constructs against which a measurement instrument is validated are defined according to their anticipated degree of similarity to the instrument at issue (Table 5.2).

Table 5.2 Categorization of comparators utilized in assessing construct validity.

Construct level	Definition
Construct 1	Validation against instruments that measure the same disability
Construct 2	Validation against instruments that measure the same disability as well as other disabilities
Construct 3	Validation against instruments that measure other disabilities than the instrument to be validated
Construct 4	Validation against generic instruments that measure disabilities as well as impairments and participation problems

'Construct 1' is the most convergent construct, which means that the validity is measured against a variable which is very similar to the variable to be validated (for example: the HAQ-Mob (variable mobility) that is validated against the AIMS-Mob). Validity is defined as 'construct 2' if the instrument is validated against instruments that measure the same construct as well as other disabilities: for example, if walking time is validated against another instrument that measures disabilities in walking, but also disabilities in personal care. 'Construct 3' indicates that the construct relates to other disabilities than the disability to be validated: for example, if walking time is validated against disabilities in interpersonal relations. 'Construct 4' is the least convergent validity, which means that the construct that is used to validate a variable relates to other domains than the variable that is to be validated (for example: a questionnaire for disabilities in walking (variable mobility) that is validated against an instrument that is intended for the assessment of muscle force).

We clustered 'construct 1' and 'construct 2' as 'validation against similar construct' and 'construct 3', 'construct 4' as 'validation against dissimilar construct'. This distinction is also reflected in the criteria for validity in the two last columns of Table 5.1. The argument for this distinction is the fact that 'similar construct validity' comes closest to the gold standard, and is therefore expected to result in stronger correlations than the dissimilar construct validity'(22).

Data-analysis

The data were analysed in the Statistical Package for Social Sciences (SPSS) 8.0. A classification of instruments was first made according to the type of disabilities, based on the ICF-classification (27).

The values of different studies of the same instrument were pooled for each clinimetric property that was assessed. When the relevant information was available, statistical pooling of the data was performed if the measurement instrument was validated against a similar construct. The values were pooled per construct. The pooled index is composed of measurements, appropriately weighted:

$$X = (\sum[n_i * x_i])/N$$

where X = pooled index, n_1^* = number of persons included in the study, x_1 = value of methodological aspect (Pearson's r, Spearman's rho or Intra Class Correlation Coefficient (ICC)) in the study, N = total number of persons in all studies included in the pooling). The pooled index was computed for Pearson's-/Spearman's correlation coefficients and the ICC values separately. Values for the convergent construct validity of multidimensional instruments can be strongly influenced by values of one or more sub-scales. Therefore, whenever possible, the data were also pooled for the separate sub-scales.

Results

For the assessment of disabilities in gait and gait-related activities, 78 instruments with a total of 36 subscales were found. Two adapted versions of the AIMS, a multidimensional questionnaire, were included in this analysis: AIMS2 and AIMSS (for all abbreviations see Appendix 3). Furthermore, beside the HAQ also a modified version of the HAQ (MHAQ) is included. The AIMS2-WaBe, AIMS2D-WaBe, FAS-Mob, HAQ-Walk, MHAQ-Walk, SW and the WT are intended for assessment of only walking or related activities (in this case climbing stairs); all other instruments or subscales are multidimensional and also measure other disabilities and/or impairments. For sixteen of these 78 instruments there are no data at all available regarding the methodological quality of these instruments (neither about reliability nor concerning validity based on the described constructs). The results of the remaining 61 instruments are presented in Table 5.3.

Intra-observer reliability was investigated for 28 instruments or subscales. 16 out of these 28 reach the criterion of $r/ICC \geq 0.85$, and 37 were found to have good validity. For 32 instruments and/or subscales construct validity is investigated in studies in which they were validated against the most similar constructs (construct 1 or

construct 2). Seventeen out of these 32 meet the criterion of $r/ICC \geq 0.65$. For 35 instruments and/or subscales construct validity is investigated in studies in which they were validated against a dissimilar construct (construct 3 or construct 4). If both reliability and validity are required to be 'good', 7 instruments meet the criteria (shown in grey in Table 5.3).

In this study we did not report on the responsiveness of the identified instruments. So far, research into the responsiveness of measurement instruments is hampered by the lack of consensus regarding the preferred method. Of the seven instruments that were found to be valid and reliable in this study, two (the DFI, and the RAQoL) are responsive according to the conclusions of the investigators (57,87,101). For HAQ, SIP and WT there are conflicting results regarding responsiveness and responsiveness of ASES-FSE and KFT are not investigated up to now.

Data on the validity of 49 instruments and/or sub-scales are available. For eighteen out of these 49 instruments, there are data on the optimally comparable construct validity as well as the imperfect construct validity. In 7 of those 18 instruments the correlation for the similar constructs proved to be stronger than the values for validation against the dissimilar constructs. For one instrument the validity was the same for the optimally comparable constructs as for validation against the imperfect constructs. In 10 out of 18 instruments the correlation values for the optimally comparable constructs proved to be lower than the values for validation against the imperfect constructs.

Table 5.3. Results for reliability and validity concerning instruments for the assessment of disabilities in gait or gait-related activities in patients with rheumatic disorders.

Measurement instrument / subscale@	Intra rater reliability	Construct validity if validated against instruments measuring			
		The same disability	The same disability as well as other disabilities	Other disabilities	Impairments as well as disabilities and participation problems or general aspects like gender, age
AIMS ²⁸⁻⁴⁹	0.86*				0.48*
AIMS-Disab ⁴²				0.88	
AIMS-GlobH ³⁸			0.79*		
AIMS-Mob ³⁸		0.74	0.74*		
AIMS-PhysA ^{29,39,44}	0.62	0.49	0.45	0.62	0.55
AIMS-PhysH ⁴⁸					0.76*
AIMS-PhysM ³⁷		0.79			
AIMSD ⁴⁷				0.43	0.65
AIMSD-PhysA ^{50,51}			0.71*		0.54
AIMSS-PhysA ⁴⁴	0.58		0.40	0.53	
AIMS2-WaBe ³⁶	0.92#				
AIMS2D-PhysA ⁵²					0.86
AIMS2D-WaBe ⁵²					0.67

ASES ⁵³	0.78*				-0.66
ASES-F; IE ^{53,54}	0.85				0.61
BIM ⁵⁵					0.43
DFI ^{56,96}	0.94				0.60
E-QoL ^{57,58,97,102}	0.74##*		0.49*		-0.78
FAS ⁵⁹	0.85##*				
FAS-Aid ⁵⁹	0.60##*				
FAS-Trans ⁵⁹	0.68##*				
FIQ ^{60,98}		0.47			
FIQ-PhysF ⁶¹	0.95				
FSAI ⁶²	0.82*				
FSI-GlobH ³⁸			0.66*		0.84
FSI-Mob ^{38,63}	0.85##*		0.61*		
GARS ^{64,65}		0.80			0.83*
GARS-ADL ^{64,65}		0.80			0.80
HAQ ^{42,58,34,66,97}	0.93##*		0.80	0.82*	0.67*
	0.94*				
HAQ-Disab ⁴²				0.88	
HAQ-GlobH ³⁸			0.71*		0.84
HAQ-Mob ³⁸		0.58	0.54*		
HAQ-ODI ⁶⁷		0.34			
HAQ-OthAc ⁶⁷		0.19			
HAQ-Walk ⁶⁷		0.23			
IRGL-Mob ^{38,68,69}		0.66*		0.12*	-0.71
IWB-GlobH ³⁸			0.56*		0.62
IWB-Mob ³⁸		0.50	0.70		
KFT ⁷⁰⁻⁷³	0.96			0.71*	0.72
LDQ ⁷⁴	0.94##*				
MDR ⁷⁵	0.93#				
MFAQ ^{76,77,101}	0.89				
	0.92#				
MHAQ ⁷⁸	0.91				
MHAQ-ODI ⁶⁷		0.34			
MHAQ-OthAc ⁶⁷		0.33			
MHAQ-Walk ⁶⁷		0.25			
MPPDQ ⁷⁹			0.66*		0.40*
MUMQ ⁸⁰⁻⁸²	0.73##*		0.45*	0.26*	0.46*
NHP-PhysM ^{83,84,105}				0.79	0.82
OSRA ⁸⁵					0.52
RAQoL ⁸⁶	0.90	0.87		0.62	
SF-36 ^{76,79,95,97,102,103,104}	0.78##*				
SF-36-PhysA ⁷⁹	0.77#		0.73*	0.56*	0.47*
SIP ^{48,87-8,99,100}	0.92*0.97#			0.55	0.71*
SIP-GlobH ³⁸			0.70*		0.87
SIP-Mob ³⁸		0.60	0.72*		
SIP-PhysH ⁸⁴					0.76*
SRQ ⁹⁰		0.44		0.58	
SW ⁹¹	0.81##*				
WOMAC-Disab ^{79,97,102,103,104,105}	0.78#		0.73*	0.60*	0.54*
WT ^{91,93,9,98,100}	0.89##*				0.66
	0.91				

Legend:

@ for explanation of abbreviations see Appendix 3

All values expressed in Pearson's-r or Spearman's-r

Intraclass Correlation Coefficient

* pooled value

In grey: instruments that meet the criteria for reliability as well as for validity

Discussion

For the assessment of disabilities in gait or gait-related activities in patients with rheumatic disorders, 78 instruments have been identified. Sixteen out of these 78 were found to have good reliability, and 37 were found to have good validity. Only 7 out of the 78 measurement instruments had good reliability as well as good validity.

The chosen methodological criteria in this study were possibly too severe. Though several authors in literature advocate a correlation of 0.4 for (construct) validity as 'reasonably high' (106-108) we only accepted a value of 0.50-0.65 as 'moderately strong' and a value ≥ 0.65 as 'strong' for construct validity. The reason to deviate from the level of ≥ 0.40 is the fact that it might be expected that a more optimal comparable construct validity generates higher values than validation against an imperfect construct (22). However, our study did not confirm this hypothesis. This could possibly be explained by the fact that the large majority of the instruments for assessment of walking and related activities are multi-dimensional. As demonstrated by Table 5.3 there are several other instruments that could be considered useful if the methodological criteria were moderated slightly. However, the combination of acceptable reliability as well as validity is sometimes lacking. For example in the literature it has been demonstrated that the performance of the SF-36 in terms of its psychometric and clinical validity, may be affected by the clinical context of the patient group in which it is applied, and that satisfactory performance cannot be guaranteed for all the groups of conditions. The same point is made for the E-QoL, which has been validated in population surveys and appears to perform well there; however, its value in rheumatic disease populations remains uncertain (Table 5.3). The problems of this scale (distribution of scores, disconcordance with VAS at lower levels of the E-QoL, and compression of item range) suggest a limited role for this instrument in rheumatologic conditions at this time.

In this study, several levels of construct validity were distinguished, based on the hypothesis that validation against an optimally comparable construct should result in stronger correlations than validation against a dissimilar construct. However, the result of this study do not confirm this hypothesis: for the validity of 10 out of 18 instruments, the value for validation against an optimally comparable construct proved to be lower than in case of validation against an imperfect construct. This is opposite to the results of a comparable study concerning validity of measurement instruments for the assessment of impairments in patients with rheumatic disorders (109). The reason for this discrepancy is unclear. A possible explanation could be the fact that the large majority of instruments for the assessment of walking and related

activities are multi-dimensional and for the most part focus on disabilities other than walking, whereas on the other hand instruments for the assessment of impairments are more frequently only focus on the impairment to be measured, and not on other domains.

The group of what is defined as 'rheumatic disorders' (see criteria in inclusion procedure) includes a variety of diseases. We also focused on differences for subgroups in the rheumatic disorders, because some measurement instruments focus in particular on specific characteristics of those patient populations and are developed just for only those patients. For example there have been several instruments developed for ankylosing spondylitis that are not relevant for patients with rheumatoid arthritis. The majority of the mentioned instruments are multi-dimensional. At this time, for the assessment of intervention in osteoarthritis of the lower extremities, the WOMAC is generally recommended as the most sensitive condition-specific instrument. However, the responsiveness varies by both subscale and intervention. Generally, the physical function subscale has large effect sizes, particularly in patients with hip and knee arthroplasty.

Of the great number of available instruments for the assessment of walking and related activities, 'WT' and 'ST' are the only instruments that measure only walking or stair walking respectively. In all other cases it concerns a subscale of a more extensive and mostly multi-dimensional instrument that measures various disabilities, not only walking. And even in case of WT there are different interpretations for patients with rheumatoid arthritis than for patients with osteoarthritis, total hip replacement etc.

The fact that majority of the instruments are questionnaires in which walking and related activities are only a small part is the consequence of our inclusion procedure. In other parts of the study we only included 1) instruments which focus mainly (50% or more of the items) on the impairment to be measured; 2) questionnaires with a sub-scale for the impairment in question that can be interpreted separately as a single entity. This requirement was not possible here, because applying this rule should result in only two instruments (WT and SW). In fact both instruments have moderate or good reliability, but there is lack of information concerning their validity. Combining reliability with validity leaves 7 instruments that reach the criteria. However, this does not mean that all others are neither reliable nor valid, because reliability is calculated only for 28 instruments. In other words: for 64% of the available instruments the reliability is unknown. Another restriction concerning the low number of data relating reliability is the fact that this review was restricted to studies focusing on populations with rheumatic disorders. The reliability and validity

of the instruments might just as well have been investigated in other patient populations.

Based on the available information the RAQoL and HAQ seem to be the best available instruments; however, both are multi-dimensional and the influence of certain subscales is not clear. For example, it is worth noting that the subscale HAQ-Walk has a very low validity score ($r=0.23$, construct 1). Both instruments could possibly result in a lot of unnecessary data for clinical use if the user is only interested in walking and gait-related activities. For that reason the usefulness of RAQoL and HAQ is under discussion if the intention is to measure only gait and gait-related activities. More attractive for clinical use could be the WT and the SW; however, there is lack of data concerning validity of those two.

In conclusion, for clinical applications the user needs to balance the importance of the available reliability, validity and amount of ballast of some instruments. Table 5.3 is intended as a guide in this process of balancing.

Acknowledgement

We wish to thank Mrs. Y. Kappe (YK) for reviewing the majority of the articles, together with the first author.

References

1. Rimmer JH. Health promotion for people with disabilities: the emerging paradigm shift from disability prevention to prevention of secondary conditions. *Phys Ther* 1999; 5:495-502.
2. World Health Organization. *International Classification of Impairments, Disabilities and Handicaps: a manual of classification relating to the consequences of disease*. Geneva: World Health Organization; 1980.
3. World Health Organization. *International Classification of Functioning, Disabilities and Health Problems*. Geneva: World Health Organization; 2001.
4. Swinkels RAHM. The ICF-classification as a system for structuring outcome measurement. *Physiotherapy Singapore* 2004;7(3):7-13.
5. Streiner WA, Ryser L, Huber E, Uebelhart D, Aeschlimann A, Stucki G. Use of the ICF model as a clinical problem solving tool in physical therapy and rehabilitation medicine. *Phys Ther* 2002; 82:1098-1107.
6. Feinstein AR. *Clinimetrics*. New haven: Yale University Press, 1987.
7. Hendriks HJM, Brandsma JW, Heerkens YF, Oostendorp RAB, Nelson RM. Intraobserver and interobserver reliability of assessments of impairments and disabilities. *Phys Ther* 1997;77:1097-1106.
8. Kirschner B, Guyatt G. A methodological framework for assessing health indices. *J Chron Dis* 1985;38:27-36.
9. Bombardier C, Tugwell P. Methodological considerations in functional assessment. *J Rheumatol* 1987;14(suppl 15): 6-10.
10. Bombardier C, Tugwell P. A methodological framework to develop and select indices for clinical trials: statistical and judgment approaches. *J Rheumatol* 1982;9:753-757.
11. Beckerman H, Bouter LM. Efficacy of physiotherapy for musculoskeletal disorders: what can we learn from research? *Br J Gen Pract* 1993;43:483-91.
12. McDowell I, Newell C. *Measuring health: a guide to rating scales and questionnaires*. New York: Oxford University Press, 1987.
13. Streiner DC, Norman GR. *Health Measurement Scales. A practical guide to their development and use*. Oxford: Oxford Medical Publications, 1989.
14. Bouter LM, Linden S van der, Koes BW. Effectmeting in de fysiotherapie. Een kritische beschouwing van de methodologie met speciale aandacht voor het kiezen van effectparameters. *Ned Tijdschr Fysiother* 1991;101:46-48.
15. Turner RR. Rehabilitation. In: Spilker B. ed. *Quality of life assessments in clinical trials*. New York: Raven Press, 1990:247-266.
16. Spector WD. Functional disability scales. In: Spilker B. ed. *Quality of life assessments in clinical trials*. New York: Raven Press, 1990:115-129.
17. Miettinen OS, Ellison RC, Peckham GJ. Overall prognosis as the primary criterion of outcome in a clinical trial. *Control Clin Trials* 1983;4:227-237.
18. Guyatt GH, Deyo RA, Charlson M. Responsiveness and validity in health status measurement: a clarification. *J Clin Epidemiol* 1989;42:403-408.
19. Deyo RA, Inui TS. Towards clinical applications of health status measures: sensitivity of scales to clinically important changes. *Health Services Research* 1984;19:275-289.
20. Deyo RA, Centor RM. Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance. *J Chron Dis* 1986;39:897-906.
21. Guyatt G, Waiter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. *J Chron Dis* 1987; 40:171-178.
22. Swinkels RAHM, Vet HCW de, Oostendorp RAB, Swinkels-Meewisse EJCM, Dijkstra PU, Bouter LM. Construct validity of impairment measures in rheumatic disorders: which constructs are selected for validation? *Clin Exp Rheum* 2005; in press.
23. Oostendorp RAB, Elvers JWH, Swinkels RAHM. Judgement list for reading articles (in Dutch). Dutch National Institute of Allied Health Professions, Amersfoort, The Netherlands, 1997.

24. Jong A de, Molenaar IW. An application of Mokken's model for stochastic, cumulative scaling in psychiatric research. *J. Psychiatr Res* 1987;21:137-49.
25. Eliasziw M, Young SL, Woodbury MG. Statistical methodology for the concurrent assessment of interrater and intrarater reliability: using goniometric measurements as example. *Phys Ther* 1994;74:777-788.
26. Doeglas D, Krol B, Guillemin F, Suurmeijer T, Sanderman R, Smedstad LM, et al. The assessment of functional status in rheumatoid arthritis: a cross cultural longitudinal comparison of the health assessment questionnaire and the Groningen Activity Restriction Scale. *J Rheumatol* 1995;22:1834-1843.
27. World Health Organization. ICDH-2, International Classification of Functioning and Disability. (Beta-2 draft, short version). Geneva: World Health Organization; 1999.
28. Meenan RF, Anderson JJ, Kazis LW, et al. Outcome assessment in clinical trials : evidence for the sensitivity of a health status measure. *Arthritis Rheum* 1984; 27:1344-1352.
29. Meenan RF, Gertman PM, Mason JH. Measuring health status in arthritis. *Arthritis Rheum* 1980;23:146-152.
30. Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. *Med Care* 1989;27:S178-S189.
31. Anderson JJ, Firschein HE, Meenan RF. Sensitivity of a health status measure to short-term clinical changes in arthritis. *Arthritis Rheum* 1989;32:844-850.
32. Hidding A, Santen M van, Klerk E de et al. Discordance between self-report measures and clinical observations of functional disability in ankylosing spondylitis, rheumatoid arthritis and fibromyalgia. Chapter 9 in: Hidding A. (Ed) Group physical therapy in ankylosing spondylitis. Meppel, Krips repro. 1993:94-104.
33. Buchbinder R, Bombardier C, Yeung M, Tugwell P. Which outcome measurements 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;38:1568-1580.
34. Anderson JJ, Chernoff MC. Sensitivity to change of rheumatoid arthritis clinical outcome measures. *J Rheumatol* 1993;20:535-537.
35. Potts MK, Brandt KD. Evidence of the validity of the arthritis impact measurement scales. *Arthritis Rheum* 1987;30:93-96.
36. 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.
37. Fitzpatrick R, Ziebland S, Jenkinson C, Mowat A. A comparison of the sensitivity to change of several health status instruments in rheumatoid arthritis. *J Rheumatol* 1993;20:429-36.
38. Liang MH, Larson MG, Cullen KE, Schwartz JA. Comparative measurement efficiency and sensitivity of five health status instruments for arthritis research. *Arthritis Rheum* 1985;28:542-547.
39. Meenan RF, Gertman PM, Mason JH, Dunaif R. The Arthritis Impact Measurement Scales: further investigations of a health status measure. *Arthritis Rheum* 1982;25:1048-1053.
40. Keefe FJ, Caldwell DS, Queen KT, et al. Pain coping strategies in osteoarthritic patients. *J Consult Clin Psychol* 1987;55:208-212.
41. Liang MH, Fossel AH, Larson MG. Comparisons of five health status instruments for orthopedic evaluation. *Med Care* 1990;28:632-642.
42. Goepfing J, Doyle MA, Charlton SL, Lorig K. A nursing perspective on the assessment of function in persons with arthritis. *Res Nurs Health* 1988;11:321-331.
43. Jacobs JWG, Oosterveld FGJ, Deurbouts N, et al. Opinions of patients with rheumatoid arthritis about their own functional capacity; how valid is it? *Ann Rheum Dis* 1992;51:765-768.
44. Wallston KA, Brown GK, Stein MJ, Dobbins CJ. Comparing the short and long versions of the arthritis impact measurement scales. *J Rheumatol* 1989;16:1105-1109.

45. Duffy CM, Watanabe KN, Gladman DD, et al. The utility of the Arthritis Impact Measurement Scales for patients with psoriatic arthritis. *J Rheumatol* 1992;19:1727-1732.
46. Hawley DJ, Wolfe F. Sensitivity to change of the health assessment questionnaire (HAQ) and other clinical and health status measures in rheumatoid arthritis: results of short-term clinical trials and observational studies versus long-term observational studies. *Arthritis Care and Research* 1992;5:130-136.
47. Vliet-Vlieland TPM, Wijk TP van der, Zwiderman JAH, Hazes JMW. Determinants of hand function in patients with rheumatoid arthritis. *J Rheumatol* 1996;23:835-840.
48. Weinberger M, Sasma GP, Tierney WM, Beleya MJ, Hiner SL. Generic versus disease specific health status measures: comparing the Sickness Impact Measurement Scales. *J Rheumatol* 1985;19:543-546.
49. Husted JA, Gladman DD, Farewell VT, Long JA. Validation of the revised and expanded version of the Arthritis Impact Measurement Scales for patients with psoriatic arthritis. *J Rheumatol* 1996;23:1015-1019.
50. Taal E, Jacobs JW, Seydal ER, Wiegman O, Rasker JJ. Evaluation of the Dutch Arthritis Impact Measurement Scales (Dutch-AIMS) in patients with rheumatic arthritis. *Br J Rheumatol* 1989;28:487-491.
51. Taal E, Seydal ER, Jacobs JW, Wiegman O, Rasker JJ. De Nederlandse Arthritis Impact Measurement Scales (Dutch-AIMS) – bepaling van fysieke en psychosociale gezondheidsaspecten van reumatoïde artritis. *Gedrag en Gezondheid* 1989;17:69-74.
52. Riemsma RP, Taal E, Rasker JJ, Houtman PM, Paassen HC van, Wiegman O. Evaluation of a Dutch version of the AIMS2 for patients with rheumatoid arthritis. *Br J Rheumatol* 1996;35(8):755-760.
53. Shoor SM, Holman HR. Development of an instrument to explore psychological mediators of outcome in chronic arthritis. *Trans Assoc Am Physicians* 1985;97:325-331.
54. Lorig K, Chastain RL, Ung E, Schoor S. Development and evaluation of a scale to measure perceived self-efficacy in people with arthritis. *Arthritis Rheum* 1989;32:37-44.
55. Bakheit A, Harries S, Hull R. Validity of a self-administered version of the Barthel index in patients with rheumatoid arthritis. *Clin Rehab* 1995;9:234-237.
56. Creemers MCW, Hof MA van 't, Franssen MJAM. A Dutch version of the functional index for ankylosing spondylitis : developemnt and validation in a long-term study. *Br J Rheumatol* 1994;33:842-846.
57. Hurst NP, Jobanputra P, Hunter M. Validity of Euroqol: a generic health status instrument in patients with rheumatoid arthritis. *Br J Rheumatol* 1994;33:655-662.
58. Hurst NP, Kind P, Ruta D. Measuring health-related quality of life in rheumatoid arthritis: validity, responsiveness and reliability of Euroqol (EQ-5D). *British Society for Rheumatology* 1997;36-5:551-559.
59. Marks R. Reliability and validity of self-paced walking time measures for knee osteoarthritis. *Arthritis Care and Research* 1994;7:51-53.
60. Lee P, Mukundrai K, Jasani W, Carson D. Evaluation of a functional index in rheumatoid arthritis. *Scand J Rheumatol* 1973;2:71-77.
61. Burckhardt CS, Clark SR, Bennett RM. The fibromyalgia impact questionnaire : development and validation. *J Rheumatol* 1991;18-5:728-733.
62. Jette AM, Deniston OL. Interobserver reliability of a functional status assessment instrument. *J Chron Dis* 1978;31:573-580.
63. Jette AM. The functional status index: reliability and validity of a self report functional disability measure. *J Rheumatol* 1987;14:S15-S19.
64. Doeglas D, Krol B, Guillemin F. The assessment of functional status in rheumatoid arthritis : a cross cultural, longitudinal comparison of the health assessment questionnaire and the Groningen activity restriction scale. *J Rheumatol*. 1995;22-10:1834-1843.
65. Callahan LF, McCoy A, Smith W. Comparison and sensitivity to change of self-report scales to assess difficulty, dissatisfaction, and pain in performing activities of daily living

- over one and five years in rheumatoid arthritis. *Arthritis Care and Research* 1992;5-3:137-145.
66. Ende C H M van den, Breedveld F C, Dijkmans B A C. The limited value of the Health Assessment Questionnaire as an outcome measure in short term exercise trials. *J Rheumatol.* 1997;24-10:1972-1977.
 67. Blalock SJ, Sauter SVH, DeVellis RF. The modified Health Assessment Questionnaire difficulty scale: a health status measure revisited. *Arthritis Care Research* 1990;3:182-188.
 68. Borstlap M, Laar M van der, Zant J. Components of health : an analysis in rheumatoid arthritis using quality of life questionnaires and clinical and laboratory variables. *Ann Rheum Dis* 1993;52-9:650-655.
 69. Bijlsma JWJ, Huiskes CJAE, Kraaijaat FW. Relation between patients' own health assessment with clinical and laboratory findings in rheumatoid arthritis. *J Rheumatol* 1991;18-5:650-653.
 70. Hakala M, Nieminen P. Functional status assessment of physical impairment in a community based population with rheumatoid arthritis: severely incapacitated patients are rare. *J Rheumatol* 1996;23:617- 623.
 71. Kalla AA, Kotze TJ, Meyers OL. Clinical assessment of disease activity in rheumatoid arthritis; evaluation of a functional test. *Ann Rheum Dis* 1988;47:773-779.
 72. Eberl DR, Fasching V, Rahlfs V. Repeatability and objectivity of various measurements in rheumatoid arthritis : a comparative study. *Arthritis Rheum* 1976;19:1278-1286.
 73. Kalla AA, Smith PR, Brown GMM. Responsiveness of Keitel Functional Index compared with laboratory measures of disease activity in rheumatoid arthritis. *Br J Rheumatol* 1995;34-2:141-149.
 74. Abbott CA, Helliwell S, Chamberlain MA. Functional assessment in ankylosing spondylitis: evaluation of a new self-administered questionnaire and correlation with anthropometric variables. *Br J Rheumatol* 1994;33:1060-1066.
 75. Durham J, MacNeish A, Rooney PJ. The MDR index of function in rheumatoid arthritis. *Clin Exp Rheumatol* 1985;3:297-302.
 76. Martin DP. Comparison of the Musculoskeletal Function Assessment Questionnaire with the Short Form-36, the Western Ontario and McMaster Universities Osteoarthritis Index, and the Sickness Impact Profile Health-status measures. *J Bone Joint Surg* 1997;79-A:1323-1335.
 77. Martin DP, Engelberg R, Agel J. Development of a musculoskeletal extremity health status instrument: the musculoskeletal function assessment instrument. *J Orthop Res* 1996;14-2:173-181.
 78. Pincus T, Summay JA, Soraci SA. Assessment of patient satisfaction in activities of daily living using a modified Stanford health assessment questionnaire. *Arthritis Rheum* 1983;26:1346-1353.
 79. Wright JG, Young NL. A comparison of different indices of responsiveness. *J Clin Epidemiol* 1997;50-3:239-246.
 80. Bakker C, Rutten M, Doorslaer E van, Bennet K, Linden S van der. Feasibility of utility assessment by rating scale and standard gamble in patients with ankylosing spondylitis or fibromyalgia. *J Rheumatol* 1994;21:269-274.
 81. Bakker C. Patient utilities in fibromyalgia and the association with other outcome measures. Maastricht - University Press, 1995.
 82. Bakker C, Rutten M, Hidding A, Doorslaer E van, Bennett K, Linden S van der. Patient utilities in ankylosing spondylitis and the association with other outcome measures. *J Rheumatol* 1994;21:1298-1304.
 83. Fitzpatrick R, Ziebland S, Jenkinson C, Mowat A, Mowat A. A general health status instrument in the assessment of rheumatoid arthritis. *Br J Rheumatol* 1992;31:87-90.
 84. Houssien DA, McHenna SP, Scott DL. The Nottingham Health Profile as a measure of disease activity and outcome in rheumatoid arthritis. *Br J Rheumatol* 1997;36:69-73.

85. Katz PP, Alfieri WS. Satisfaction with abilities and well-being: development and validation of a questionnaire for use among persons with rheumatoid arthritis. *Arthritis Care and Research* 1997;10:89-98.
86. Jong Z de, van der Heijde D, McKenna SP. The reliability and construct validity of the RAQoL : a rheumatoid arthritis- specific quality of life instrument. *Br J of Rheumatol* 1997;36:878-883.
87. Deyo RA, Inui TS, Leininger JD. Measuring functional outcomes in chronic disease: a comparison of traditional scales and a self-administered health questionnaire in patients with rheumatoid arthritis. *Med Care* 1983;21:180-192.
88. Bergner M, Bobbitt RA, Carter WB. The Sickness Impact Profile : development en final revision of a health status measure. *Med Care* 1981;19:787-805.
89. Papageorgiou AC, Badley EM. The quality of pain in arthritis: the words patients use to describe overall pain and pain in individual joints at rest and on movement. *J Rheumatol* 1989;16:106-112.
90. Pincus T, Callahan LF, Brooks RH. Self-report questionnaire scores in rheumatoid arthritis compared with traditional physical, radiographic and laboratory measures. *Ann Intern Med* 1989;110:259-266.
91. Marks R. Walking time measures for evaluating OA of the knee: reliability and sensitivity of self-paced walking time for assessing functional capacity in persons with osteoarthritis of the knee. *S Afr J Physiother* 1994;50:1:5-8.
92. Webb J, Wilson DW, Carson DW. Evaluation of digital joint circumference measurements in rheumatoid arthritis. *Scand J Rheumatol* 1973;2:127-131.
93. Borstlap M, Zant JL, Soesbergen SM van, Korst JK van der. Quality of life assessment: a comparison of four questionnaires for measuring improvement after total hip replacement. *Clin Rheumatol* 1995;14:15-20.
94. Grace EM, Cerecz E, Kassam Y, Buchanan WW. Fifty foot walking time: an inappropriate outcome measure. *Arthritis Rheum* 1986;29:516.
95. Husted JA, Gladman DD, Farewell VT, Long JA, Cook RJ. Validating the SF-36 Health Survey Questionnaire in patients with psoriatic arthritis. *J Rheumatol* 1997;24:511-517.
96. Hidding A, Santen M van, Klerk E de, Gielen X, Boers M, Geenen R, Vlaeyen J, Kester A, Linden S van der. Comparison between self-report measures and clinical observations of functional disability in ankylosing spondylitis, rheumatoid arthritis and fibromyalgia. *J Rheumatol* 1994;21:818-823.
97. Brazier JE, Harper R, Munro J, Walters SJ, Snaith ML. Generic and condition-specific outcome measures for people with osteoarthritis of the knee. *Br Society for Rheumatol* 1999;38:870-877.
98. King S, Wessel J, Bhambhani Y, Maikala R, Sholter D, Maksymowych W. Validity and reliability of the 6 minute walk in persons with fibromyalgia. *J Rheumatol* 1999;26:2233-2237.
99. Kosinski M, Keller SD, Hatoum HT, Kong SX, Ware JE. The SF-36 Health Survey as a generic outcome measure in clinical trials of patients with osteoarthritis and rheumatoid arthritis. *Med Care* 1999;37:MS10-MS22.
100. Kosinski M, Keller SD, Ware JE, Hatoum HT, Kong SX. The SF-36 Health Survey as a generic outcome measure in clinical trials of patients with osteoarthritis and rheumatoid arthritis. *Med Care* 1999;37:MS23-MS39.
101. Swiontkowski MF, Engelberg R, Martin DP, Agel J. Short musculoskeletal function assessment questionnaire: validity, reliability and responsiveness. *J Bone Joint Surg* 1999;81-A:1245-1260.
102. Fransen M, Edmonds J. Reliability and validity of the EuroQol in patients with osteoarthritis of the knee. *Br Society for Rheumatol* 1999;38:807-813.
103. Domsic RT, Saltzman CL. Ankle Osteoarthritis Scale. *Foot & Ankle* 1998;19:466-471.
104. Roos EM, Klässbo M, Lohmandeör LS. WOMAC Osteoarthritis Index. *Scand J Rheumatol* 1999;28:210-215.
105. Söderman P, Malchau H. Validity and reliability of Swedish WOMAC Osteoarthritis Index. *Acta Orthop Scand* 2000;71:39-46.

106. Ghiselle EE. The measurement of occupational aptitude. Berkely, Los Angeles, 1966.
107. Huiskes CJAE, Kraaimaat FW, Bijlsma JJJ. Development of a self-report questionnaire to assess the impact of rheumatic diseases on health and lifestyle. *J Rehab Sciences* 1990;3-2:65-70.
108. Stewart AL, Ware JE. Measuring functioning and well-being: the Medical Outcomes Study Approach. Durham, North Carolina, Duke University Press, 1992.
109. Swinkels RAHM, Oostendorp RAB, Bouter LM, Ende CHM van den. Measuring impairments in rheumatic disorders: what are the best instruments for allied health care? *Rheum Int* 2005;25(7):501-512.

Chapter 6

Reliability, validity and responsiveness of instruments to assess disabilities in personal care in patients with rheumatic disorders: a systematic review

Published as: Swinkels RAHM, Dijkstra PU, Bouter LM. Reliability, validity and responsiveness of instruments to assess disabilities in personal care in patients with rheumatic disorders: a systematic review. Clin Exp Rheum 2005;23:71-79

Abstract*Objectives*

The first aim was to make an inventory of available instruments and questionnaires for the assessment of disabilities in personal care in patients with rheumatic disorders. The second aim was to investigate which of these instruments have acceptable methodological quality with regard to reliability, validity and responsiveness. The third aim was to investigate the assumption that convergent validity results in stronger correlations when validated against a more similar construct.

Methods and materials

A computer-aided literature search (1982-2001) in several databases was performed to identify studies focusing on the clinimetric properties of instruments to assess impairments in function in patients with rheumatic disorders. Data were extracted in a standardised way and compared to a priori defined criteria.

Results

In total, 19 measurement instruments were included. Five out of these 19 were found to have acceptable reliability, while 12 had acceptable validity. Only three questionnaires met both criteria. Results concerning the responsiveness of these three questionnaires were conflicting. No difference was found in the strength of correlation between validation against the most similar construct versus validation against the least similar construct.

Conclusion

It is concluded that the Arthritis Impact Measurement Scale (AIMS) is the most suitable instrument for the assessment of disabilities in personal care.

Introduction

Personal care is an essential aspect of health-related quality of life for patients with rheumatic disorders. Due to the progressive nature of rheumatic disorders, it is important that patients are able to manage their daily life independently for as long as possible. This literature review focuses on the clinimetric properties of instruments that assess personal care disabilities (washing, dressing, eating and drinking) in patients with rheumatic disorders. To evaluate treatments aimed at reducing personal care disabilities or slowing their progress over time, measurement instruments should be used to objectify this. In a search for the optimal measurement instrument, knowledge about the domain to be measured is the first requirement. The domain to be measured must be relevant for treatment goals. In addition, the measurement instruments need to be reliable and valid, and for the evaluation of treatment they should also be responsive (1-5). There is no 'gold standard' for the assessment of (construct-) validity of instruments measuring different levels of personal care disability; therefore, construct validity is evaluated (6). However, there is no consensus on how construct validity should be evaluated (7;8). Such an evaluation can be based on comparison with a very similar construct (optimal construct comparison) or a different construct (imperfect construct comparison), such as age, gender, or an impairment, for instance range of motion, instead of a disability (9). It can be hypothesised that the outcome of construct validity evaluation may be influenced by the similarity or dissimilarity of the construct against which the measurement instrument is validated. Comparison with a similar construct might result in a stronger correlation than comparison with a dissimilar construct. In the latter case the validity of the instrument is under-valued.

Rehabilitation medicine, physiotherapy and occupational therapy focus on the consequences of diseases for a patient, and aim at reducing the burden of these consequences. Hence, in rehabilitation medicine there is a need to describe the functioning (consequences of disease) of the patient in daily life. The consequences of a disease for a specific patient and the patient's complaints are not only related to the impairments due to the disease (biomedical model), but also to coping, illness perception, etc. (bio-psychosocial model). A classification system which describes these consequences, should include diseases and symptoms, as well as the complaints and the ability of the patient to function. However, the most frequently used system for the classification of diseases, the *International Classification of Diseases and Related Health Problems* (ICD) (10), is not appropriate for the classification of health problems according to the bio-psychosocial model. The ICD

focuses only on impairments and/or diseases, and does not focus on the different domains of the bio-psychosocial model, which are a much broader context. This shortcoming of the ICD led to the development of the *International Classification of Impairments, Disabilities and Handicaps* (ICIDH) (11), later modified as the *International Classification of Functioning, Disability and Health* (ICF) (12).

This shift of focus from disease to the consequences of disease (which took place mainly in the late eighties and the nineties of the previous century), implied that the focus of measurement instruments also had to change. Not only impairments should be measured, but also disabilities to perform activities and problems in participation. 'Impairment' is defined as loss or abnormality of psychological, physiological or anatomical structure or function (for instance reduction in range of motion or loss of muscle strength). 'Disability' is defined as any restriction or lack of ability to perform an activity (for instance washing or dressing) in the manner or within the range considered normal for a human being, as a result of an impairment. 'Handicap' is defined as a disadvantage for a given individual that limits the fulfilment of a role that is normal for that person, related to age, gender, and social and cultural factors, as a result of an impairment or disability. However, the majority of instruments were developed before the above-mentioned shift in focus. As a consequence for patients with rheumatic disorders, 57% of the available measurement instruments (N=209) focus on impairments in function, approximately 37% measure disabilities to perform activities, and only 5% of all instruments deal with problems in participation (13).

As Stucki and Sigl emphasize, from the rehabilitation perspective, the measurement of functioning and health is relevant not only for evaluating the impact of the disease or the outcome of intervention. It is also used for the diagnosis and interventional management of limitations in functioning and health. Thus, from a rehabilitation perspective, measures of functioning, disability and health are examined much more closely at both the level of individual problems and the level of instrument scales. In rheumatology, both perspectives are important in the management of most patients.

Health status instruments usually include the dimensions of physical functioning, social function, emotional function, pain and the perception of well-being. However, there is unfortunately no clear relationship with the ICF framework (14)

The aim of this literature review was threefold. The first aim was to make an inventory of available instruments and questionnaires for the assessment of personal care disabilities in patients with rheumatic disorders. The second aim was to investigate which of these instruments have acceptable methodological quality with regard to reliability, validity and responsiveness. The third aim was to investigate the assumption that convergent construct validity results in stronger correlations when validated against a more similar construct.

Method and materials

Literature search

The Medline database was searched for the period January 1982 - April 2001, using specific search terms for the relevant rheumatic disorders and various search terms for clinimetric properties⁴. The database of the Centre for Documentation of the Dutch National Institute of Allied Health Professions was also searched for the period January 1988 – April 2001, using the same keywords. The names of the measurement instruments that were identified in the first searches were used as free text words in additional searches in the databases. Papers written in English, French, German and Dutch were included. The search was subsequently augmented with a manual search based on the reference list of the identified papers. This final search yielded a number of papers written before 1982.

Inclusion criteria

The following inclusion criteria were applied:

Papers had to focus on patients suffering from rheumatoid arthritis, seronegative polyarthritis (including psoriatic arthritis), osteoarthritis, ankylosing spondylitis, polymyositis or fibromyalgia.

Papers had to contain information about the clinimetric properties of instruments to assess personal care disabilities.

Different versions of an instrument were considered as separate measurement instruments.

Because many questionnaires focus on more than one domain of the ICIDH or ICF classification, or measure more than one variable, papers were included if the instruments focused mainly (50% or more of the items) on the disability to be measured, or if questionnaires had a sub-scale for personal care disability so that it can be interpreted separately as a single entity.

Only instruments for the measurement of adult patients were included.

Data-extraction

All papers that were identified were selected independently on the basis of title and abstract by two reviewers (RS and YK). A standardised scoring form (15) was used to assess reliability, validity and responsiveness. The domains according to the ICIDH classification were also assessed. In case of disagreements, which occurred in 3% of the assessments, the paper was also assessed by a third reviewer. An

⁴ The detailed search strategy is available on request from the corresponding author.

explanation of all the abbreviations of instruments and sub-scales mentioned in this article is given in Appendix 3.

Methodological criteria for clinimetric properties

Analysis and interpretation of the intra-observer reliability, construct validity and responsiveness were performed according to the criteria presented in Table 6.1 (16-18).

Table 6.1. Cut-off points for intra-rater reliability and construct validity

	Intra-rater reliability	Optimal convergent construct validity	Least convergent construct validity
good	$x \geq 0.85$	$x \geq 0.65$	$x \geq 0.50$
moderate	$0.65 \# x < 0.85$	$0.50 \# x < 0.65$	$0.40 \# x < 0.50$
poor	$x < 0.65$	$x < 0.50$	$x < 0.40$

Legend: x = Pearson's r , Spearman's ρ or Intra-Class Correlation Coefficient (ICC)

Construct validity was divided into five levels, in which the constructs against which the instruments were validated were ranked according to their degree of similarity to the instrument at issue, a method previously used by Swinkels et al (19) (Table 6.2). Thus the five levels indicate the degree of convergence.

'Level 1 construct' is the most convergent construct, which means that the instrument is validated against an instrument which is very similar to the instrument to be validated. For example, the HAQ-Mob (variable mobility) is validated against the AIMS-Mob (variable mobility). 'Level 2 construct' indicates that the instrument is validated against an instrument that measures the same construct, as well as other types of disabilities. For example, the HAQ-eat is validated against the total score of the AIMS (including more disabilities other than eating).

'Level 3 construct' indicates that the instrument is validated against an instrument that measures other disabilities than the disability to be validated. For example, the HAQ-eat is validated against the SIP-mobility. 'Level 4 construct' indicates that the instrument is validated against instruments that measure impairments instead of disabilities. For example, the HAQ-mobility is validated against the Larsen score.

Table 6.2. Levels of construct convergence used to assess construct validity ('Instrument' can mean a 'measurement instrument' (the SODA or a questionnaire) as well as a 'sub-scale' of a questionnaire (see Appendix 3))

Level of convergence of constructs	Definition	
Level 1 construct	Validation against instruments that measure the same disability	Optimal convergent construct
Level 2 construct	Validation against instruments that measure the same disability as well as other disabilities	Optimal convergent construct
Level 3 construct	Validation against instruments that measure other disabilities than the instrument to be validated	Less optimal convergent construct
Level 4 construct	Validation against instruments that measure impairments instead of the disability at issue	Less optimal convergent construct
Level 5 construct	Validation against generic instruments that measure impairments as well as disabilities and participation problems	Less optimal convergent construct

'Level 5 construct' is the least convergent validity, indicating that the instrument is validated against generic instruments that measure impairments as well as disabilities and problems in participation, or instruments that measure global aspects of the patients such as gender, age, etc.

In order to be able to interpret the results and to investigate the assumption that construct validity results in stronger correlations when validated against a more similar construct, the levels 1 to 5 construct were dichotomised. Level 1 and level 2 construct were clustered into 'optimal convergent construct validity' and levels 3 through 5 construct were clustered into 'least convergent construct validity' (Table 6.1). This dichotomy was made because an optimal convergent construct validity resembles the gold standard, and is therefore expected to result in stronger correlations than the least convergent construct validity (9;13).

There is no consensus about the most appropriate method to evaluate responsiveness (20). Therefore, the results are merely described as reported by the authors.

Data-analysis

The instruments were classified according to the type of disabilities, based on the ICF classification (12).

When the relevant information was available, statistical pooling of the data was performed if the measurement instrument was validated against the same construct. The values were pooled per construct. A pooled index of the reliability and validity of the instruments was compiled, weighted according to the formula:

$$X = \frac{(\sum[n_i \cdot x_i])}{N}$$

where X = pooled estimate, n_1 = number of persons included in the study, x_1 = correlation (Pearson's r , Spearman's ρ or Intra Class Correlation Coefficient (ICC)) reported in the study, N = total number of persons in all studies included in the pooling). The pooled estimate was separately computed for Pearson's r , Spearman's ρ and the ICCs. Values for the convergent construct validity of multidimensional instruments can be strongly influenced by values of one or more sub-scales (21-23). Therefore, whenever possible, the data were also pooled for the separate sub-scales.

The data were analysed in the Statistical Package for Social Sciences (SPSS) 8.0.

Results

In total, 19 different measurement instruments or sub-scales for the assessment of personal care disabilities were identified: five integral instruments (AIMS, AIMS-D, HAQ, MHAQ and SODA) and 14 sub-scales of multi-dimensional instruments. All integral instruments were, in fact, multi-dimensional questionnaires, in which only some of the questions focused on personal care.

Five out of these 19 instruments had good reliability (AIMS, HAQ, MHAQ, SIP and SODA), and eleven questionnaires or sub-scales were found to have good validity. (Table 6.3).

The construct validity of eight instruments and/or sub-scales was investigated in studies in which they were validated against the most optimally convergent constructs (level 1 or 2 construct). Three out of these eight met the criterion of $r/ICC \geq 0.65$. The construct validity of 16 instruments and/or sub-scales had been investigated in studies in which they were validated against a less convergent construct (level 3, 4 or 5 construct). Ten of these 16 instruments met the criterion for good validity.

Good reliability as well as good validity was found for three instruments (AIMS, HAQ and SIP, marked grey in Table 6.3). With regard to the responsiveness of the three instruments that were found to be both reliable and valid, the conclusions drawn in the various studies were conflicting.

Data on validity were available for all 19 instruments or sub-scales. For eight instruments there were data on comparisons with an optimal convergent construct and with a less optimal convergent construct. In four of the instruments the strength of the correlations with an optimal convergent construct was similar or higher than the strength of the correlations with a less convergent construct, and in four of the instruments the strength of the correlations with an optimal convergent construct was less than the strength of the correlations with a less convergent construct.

Table 6.3. Results for intra-rater reliability, validity and responsiveness of instruments for the assessment of disabilities in personal care in patients with rheumatic disorders

Measurement Instrument / sub-scale@	Intra-rater reliability	Construct validity if validated against instruments measuring					Responsiveness
		the same disability	the same disability as well as other disabilities	impairments	impairments as well as disabilities and participation problems	General aspects, such as gender, age etc.	
		Level 1	Level 2	Level 3	Level 4	Level 5	
AIMS ^(21;22;29-49)	0.86*				0.71*		1,2
AIMS-ADL ^(21;22;29-49)	0.59		0.30	0.34*	0.39*	0.29*	
AIMS2D-Selfcare ^(37;48-50)					0.77	0.22*	1
AIMSD ^(43;46;51)		0.65		0.43	0.65	-0.56	1
AIMSD-ADL ^(43;46;51;52)			0.69*		0.36	0.37*	
AIMSS-ADL ⁽²²⁾	0.56		0.33	0.40	0.29	0.29*	
GARS-ADL ^(18;53;54)			0.64	0.78*	0.63*		
HAQ ^(18;21;23;34;38;41;42;45;54-71)	0.94* 0.93##		0.80	0.82*	0.78	0.34*	1,2
HAQ-hygie ^(18;21;23;34;38;41;42;45;54-71)				0.10	0.42	0.15*	1,2
HAQ-eat ^(18;21;23;34;38;41;42;45;54-71)			0.43	0.65		-0.09*	1,2
HAQ-dresG ^(18;21;23;34;38;41;42;45;54-71)			0.21	0.73		0.19*	1,2
IRGL-Selfcare ⁽⁷²⁻⁷⁵⁾				0.28*	0.77	-0.18*	
MHAQ ^(23;59;60;63;64;67-69)	0.91			0.05		0.24*	1,2
MHAQ-hygie ^{49,50,53,54,57-60}				0.49	0.42		1,2
MHAQ-eat ^(23;59;60;63;64;67-69)					0.65	-0.05*	1,2
MHAQ-dresG ^(23;59;60;63;64;67-69)				0.26	0.73	-0.06*	1,2
SIP-BC&M ^(38;41;47;76-84)	0.92*					0.59*	1,2
SIP-eat ^(38;41;47;76-84)	0.92*					0.22	1,2
SODA ⁽⁸⁵⁾	0.93					0.26*	

Legend:

@ for explanation of abbreviations see Appendix 3

All values expressed in Pearson's-r or Spearman's-rho

Intra-class Correlation Coefficient

* pooled value

Responsiveness 1: studies that indicate that the instrument is responsive

Responsiveness 2: studies that indicate that the instrument is not responsive

In grey: instruments that meet the criteria for reliability as well as for validity

Discussion

Novel in our systematic review is firstly its focus on reliability, (construct) validity and responsiveness of all available measurement instruments for one single domain (disabilities in personal care in patients with rheumatic disorders). Secondly, we grouped the data on content validity optimally by classifying all constructs, that are used to validate a measurement instrument, according the ICF-classification. In this way, we divided optimal convergent constructs from less convergent constructs.

Based on the results of this study, it is concluded that the AIMS and the HAQ are the most suitable instruments for the assessment of personal care disabilities in patients with rheumatic disorders. The reliability and validity of both of these questionnaires was good to very good, based on the results of the evidence retrieved in the literature review. Although the validity of the integral HAQ is very good, analysis of its validity at the level of the sub-scales demonstrates, in general, unacceptably low values (in the majority <0.50) for HAQ-eat and for HAQ-hygie. The AIMS was responsive, in particular for the self-care sub-scale, whereas for the HAQ more conflicting conclusions were found for responsiveness. Therefore, the AIMS is more appropriate than the HAQ for measuring personal care disabilities, because the AIMS has a stronger validity for almost all of the sub-scales than the HAQ. The reliability of most instruments was good, with the exception of the ADL sub-scale of the AIMS and the AIMSS (Table 6.3).

Since only questionnaires which focus mainly (50% or more of the items) on disabilities in personal care were included, this reduced the total number of available questionnaires (N=69) to 19. In fact, in the review six different instruments were assessed; the other instruments were sub-scales or modifications of one of these six instruments (Table 6.3). The reliability of five of these instruments was good, but the reliability of the GARS was not investigated. There is a difference between the SODA and the other instruments, in that the SODA is a performance-based instrument that measures manual activities, whereas the other instruments are questionnaires based on the patient's perception of (dis)ability to perform activities.

Generic measures have some disadvantages, for example, items are often irrelevant to persons with certain kinds of disability. Another disadvantage of most generic measures is the fact that people with low levels of physical functioning are not asked about high-levels activities, and people with high levels of physical functioning are not asked about easy activities (24).

Therefore, it is recommended to supplement generic measures with a targeted measure whenever possible (Hays et al.) (24)

From the clinical point of view one of the contributions of this paper is, that we grouped data on the content validity optimally by classifying all constructs, that are

used to validate a measurement instrument, according to the ICF-classification. In this way, we divided optimal convergent constructs from less convergent constructs. Several levels of construct validity were distinguished in this review because it was hypothesised that validation against an instrument with an optimal convergent construct may result in stronger correlations than validation against a less convergent construct. However, the results do not support this hypothesis. In four out of eight validations, comparison with an optimal convergent construct resulted in stronger correlations than comparison with a less convergent construct. In the other 4 validations the results were just the opposite. Even when a more strict cut-off point for good validity (all ≥ 0.65) was applied, the results with regard to the assumed hypothesis did not change (data not shown). Similar results concerning validation against an optimal or a less optimal construct have been found for instruments to measure disabilities in gait and related activities in patients with rheumatic disorders (19). In that study it was found that in 7 out of 18 instruments the correlations with optimal constructs were stronger than the correlations with less optimal constructs. For one instrument the correlations were the same for both the optimal and the less optimal construct. In 10 out of 18 instruments the correlations with optimal constructs were weaker than the correlations with less optimal constructs.

In contrast to the results reported here are the findings of an earlier study that in 11 out of 18 instruments the correlations with optimal constructs were higher than the correlations with less optimal constructs (25). However, in that study the validity of instruments to measure impairments in patients with rheumatic disorders was reviewed. The explanation for these differences in outcome might be the following. The majority of instruments to assess impairments are uni-dimensional, while instruments to assess personal care disabilities are multi-dimensional, indicating that the total score of disabilities may be strongly influenced by one or more sub-scales (positive as well as negative) (21;23). In other words: the heterogeneity of questions and sub-scales in multi-dimensional questionnaires could explain these differences, because instruments to measure impairments are more homogeneous and uni-dimensional. Another explanation for the differences in review results could be that questionnaires to measure disabilities reflect the patient's perception of the disability at issue, whereas the majority of instruments to measure impairments are anthropometric, and reflect the assessment made by an observer.

The cut-off points for construct validity, as presented in Table 6.1, were chosen more or less arbitrarily, because in the literature a wide range of criteria are described, depending on the discipline underlying the study. In general, the highest cut-off criteria are applied in pharmacology (26;27), whereas lower cut-off criteria (0.21-0.54)

are applied in disciplines such as psychology and sociology, where there is often no gold standard available and the heterogeneity of constructs is usually greater (28). The responsiveness of the three instruments that were found to be both reliable and valid (AIMS, HAQ and SIP) is not clear. Conflicting outcomes are reported for responsiveness, probably because there is lack of consensus regarding the best method for the evaluation of responsiveness. Because AIMS, HAQ, and SIP are reliable and valid questionnaires that are widely used, the responsiveness of these instruments should be evaluated in people with rheumatic disorders. Therefore, there is need for additional and standardised research on the responsiveness of these instruments in homogeneous populations with rheumatic disorders, based on the same method for the measurement of responsiveness outcomes.

Based on the results of this review, it is now concluded that the AIMS is the most suitable instrument for the measurement of disabilities in personal care in patients with rheumatic disorders. We would like to emphasize that this conclusion is based solely on the clinimetric properties that we evaluated, and that other aspects, like availability, usefulness, difficulties to handle and interpret, time required to complete, etcetera are not assessed in this review.

Acknowledgment

We wish to thank Mrs. Y. Kappe for reviewing the articles, together with the first author.

References

1. Feinstein AR. Clinimetrics. New Haven, Yale University Press, 1987.
2. Kirschner B, Guyatt GH. A methodological framework for assessing health indices. *J Chron Dis* 1985; 38:27-36.
3. Bombardier C, Tugwell P. Methodological considerations in functional assessment. *J Rheumatol* 1987; 14 Suppl 15:6-10.
4. Beckerman H, Bouter LM. Efficacy of physiotherapy for musculoskeletal disorders: what can we learn from research? *Br J Gen Pract* 1993; 43:483-91.
5. McDowell I, Newell C. Measuring health: a guide to rating scales and questionnaires. 1987. New York, Oxford University Press.
6. Portney LG, Watkins MP. Foundations of clinical research. Applications to practice. Appleton and Lange, 2000:1-722.
7. Streiner DL, Norman GR. Health Status Measurement Scales. A practical guide to their development and use. 2nd ed., Oxford, Oxford University Press, 2003:1-231.
8. Jensen MP. Questionnaire validation: a brief guide for readers of the research literature. *Clin J Pain* 2003;19(6):345-52.
9. Swinkels RAHM, Bouter LM, Oostendorp RAB, Swinkels-Meewisse EJCM, Dijkstra PU, Vet HCW de. Construct validity of impairment measures in rheumatic disorders: which constructs are selected for validation? A systematic review. *Clin Exp Rheum* 2005; in press.
10. World Health Organization. International Statistical Classification of Diseases and Related Health Problems (ICD-10), 10th revision. 1993. Geneva, World Health Organisation.
11. World Health Organization. International Classification of Impairments, Disabilities and Handicaps: a manual of classification relating to the consequences of disease. 1980. Geneva, World Health Organisation.
12. World Health Organization. International Classification of Functioning, Disabilities and Health Problems. 2001. Geneva, World Health Organization.
13. Swinkels RAHM. The ICF-classification as a system for structuring outcome assessment. *Physiotherapy Singapore* 2004; 7(3):7-13.
14. Stucki G, Sigl T. Assessment of the impact of disease on the individual. *Best Pract Res Clin Rheumatol* 2003; 17(3):451-73.
15. Oostendorp RAB, Elvers JWH, Swinkels RAHM. Judgement list for reading articles. Amersfoort 1997, unpublished.
16. DeJong A, Molenaar IW. An application of Mokken's model for stochastic cumulative scaling in psychiatric research. *J Psychiatr Res* 1987;21:137-49.
17. Eliasziw M, Young SL, Woodbury MG, Fryday-Field K. Statistical Methodology for the Concurrent Assessment of Interrater and Intrarater Reliability: Using Goniometric Measurements as an Example. *Physical Therapy* 1994;74(8):777-88.
18. Doeglas D, Krol B, Guillemin F et al. The assessment of functional status in rheumatoid arthritis: a cross cultural, longitudinal comparison of the Health Assessment Questionnaire and the Groningen Activity Restriction Scale. *J Rheumatol* 1995; 22(10):1834-43.
19. Swinkels RAHM, Oostendorp RAB, Bouter LM. Which are the best instruments for measuring disabilities in gait and gait-related activities in patients with rheumatic disorders. *Clin Exp Rheumatol* 2004;22(1):25-33.
20. Terwee CB, Dekker FW, Wiersinga WM, Prummel MF, Bossuyt PM. On assessing responsiveness of health-related quality of life instruments: guidelines for instrument evaluation. *Qual Life Res* 2003;12(4):349-62.
21. Fitzpatrick R, Ziebland S, Jenkinson C, Mowat A, Mowat A. A comparison of the sensitivity to change of several health status instruments in rheumatoid arthritis. *J Rheumatol* 1993;20(3):429-36.

22. Wallston KA, Brown GK, Stein MJ, Dobbins CJ. Comparing the short and long versions of the Arthritis Impact Measurement Scales. *J Rheumatol* 1989;16(8):1105-9.
23. Ziebland S, Fitzpatrick R, Jenkinson C, Mowat A, Mowat A. Comparison of two approaches to measuring change in health status in rheumatoid arthritis: the Health Assessment Questionnaire (HAQ) and modified HAQ. *Ann Rheum Dis* 1992;51(11):1202-5.
24. Hays RD, Hahn H, Marshall G. Use of the SF-36 and other health-related quality of life measures to assess persons with disabilities. *Arch Phys Med Rehabil* 2002;83(12 Suppl 2):S4-S9.
25. Swinkels RAHM, Oostendorp RAB, Bouter LM, Ende CHM van den. Measuring impairments in rheumatic disorders: what are the best instruments for allied health care? *Rheum Int* 2005, in press.
26. Cox-Reijven PL, van Kreel B, Soeters PB. Bioelectrical impedance measurements in patients with gastrointestinal disease: validation of the spectrum approach and a comparison of different methods for screening for nutritional depletion. *Am J Clin Nutr* 2003;78(6):1111-9.
27. Christensen AE, Tobiassen M, Jensen TK, Wielandt H, Bakketeig L, Host A. Repeated validation of parental self-reported smoking during pregnancy and infancy: a prospective cohort study of infants at high risk for allergy development. *Paediatr Perinat Epidemiol* 2004;18(1):73-9.
28. Vlaeyen JWS, Kole-Snijders AMJ, Rotteveel AM, Ruesink R, Heuts PHTG. The role of fear of movement/(re)injury in pain disability. *J Occup Rehabil* 1995; 5(4):235-52.
29. Meenan RF, Anderson JJ, Kazis LE et al. Outcome assessment in clinical trials. Evidence for the sensitivity of a health status measure. *Arthritis Rheum* 1984; 27(12):1344-52.
30. Meenan RF, Gertman PM, Mason JH. Measuring health status in arthritis. The arthritis impact measurement scales. *Arthritis Rheum* 1980;23(2):146-52.
31. Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. *Med Care* 1989;27(3 Suppl):S178-S189.
32. Anderson JJ, Firschein HE, Meenan RF. Sensitivity of a health status measure to short-term clinical changes in arthritis. *Arthritis Rheum* 1989; 32(7):844-50.
33. Hidding A, Santen M van, Klerk Ed, Gielen X, Boers M. Discordance between self-report measures and clinical observations of functional disability in ankylosing spondylitis, rheumatoid arthritis and fibromyalgia. Group physical therapy in ankylosing spondylitis (thesis). Meppel, Krips Repro, 1993:94-104.
34. 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(11):1568-80.
35. Anderson JJ, Chernoff MC. Sensitivity to change of rheumatoid arthritis clinical trial outcome measures. *J Rheumatol* 1993;20(3):535-7.
36. Potts MK, Brandt KD. Evidence of the validity of the Arthritis Impact Measurement Scales. *Arthritis Rheum* 1987;30(1):93-6.
37. 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):1-10.
38. Liang MH, Larson MG, Cullen KE, Schwartz JA. Comparative measurement efficiency and sensitivity of five health status instruments for arthritis research. *Arthritis Rheum* 1985;28(5):542-7.
39. Meenan RF, Gertman PM, Mason JH, Dunaif R. The arthritis impact measurement scales. Further investigations of a health status measure. *Arthritis Rheum* 1982;25(9):1048-53.

40. Keefe FJ, Caldwell DS, Queen KT, Gil KM, Martinez S, Crisson JE, Ogden W, Nunley J. Pain coping strategies in osteoarthritis patients. *J Consult Clin Psychol* 1987;55(2):208-12.
41. Liang MH, Fossel AH, Larson MG. Comparisons of five health status instruments for orthopedic evaluation. *Med Care* 1990;28(7):632-42.
42. Goepfinger J, Doyle MA, Charlton SL, Lorig K. A nursing perspective on the assessment of function in persons with arthritis. *Res Nurs Health* 1988;11(5):321-31.
43. Jacobs JW, Oosterveld FG, Deuxbouts N, Rasker JJ, Taal E, Dequeker J, Uytterhoeven R. Opinions of patients with rheumatoid arthritis about their own functional capacity: how valid is it? *Ann Rheum Dis* 1992;51(6):765-8.
44. Duffy CM, Watanabe Duffy KN, Gladman DD, Brubacher BB, Buskila D, Langevitz P, Farewell VT. The utility of the arthritis impact measurement scales for patients with psoriatic arthritis. *J Rheumatol* 1992;19(11):1727-32.
45. Hawley DJ, Wolfe F. Sensitivity to change of the health assessment questionnaire (HAQ) and other clinical and health status measures in rheumatoid arthritis: results of short-term clinical trials and observational studies versus long-term observational studies. *Arthritis Care Res* 1992;5(3):130-6.
46. Vliet-Vlieland TP, Wijk TP van der, Jolie IM, Zwiderman AH, Hazes JM. Determinants of hand function in patients with rheumatoid arthritis. *J Rheumatol* 1996;23(5):835-40.
47. Weinberger M, Samsa GP, Tierney WM, Belyea MJ, Hiner SL. Generic versus disease specific health status measures: comparing the sickness impact profile and the arthritis impact measurement scales. *J Rheumatol* 1992;19(4):543-6.
48. Husted J, Gladman DD, Farewell VT, Long JA. Validation of the revised and expanded version of the Arthritis Impact Measurement Scales for patients with psoriatic Arthritis. *J Rheumatol* 1996;23(6):1015-9.
49. Goossens PH, Heemskerk B, Tongeren J van, Zwiderman AH, Vlieland TP, Huizinga TW. Reliability and sensitivity to change of various measures of hand function in relation to treatment of synovitis of the metacarpophalangeal joint in rheumatoid arthritis. *Rheumatology (Oxford)* 2000;39(8):909-13.
50. 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(8):755-60.
51. Oosterveld FGJ, Jacobs JWG, Deuxbouts N, Rasker JJ, Dequeker J, Taal E. De DUTCH-AIMS bij RA: een goede maat voor fysiek functioneren. *Ned Tijdschr Fysioth* 1990;100(12):317-9.
52. Taal E, Jacobs JW, Seydel ER, Wiegman O, Rasker JJ. Evaluation of the Dutch arthritis Impact Measurement Scales (DUTCH-AIMS) in patients with rheumatoid arthritis. *British J Rheumatol* 1989;28:487-91.
53. Kempen GIJM, Doeglas DM, Suurmeijer ThPBM. Groningen Activiteiten Restrictie Schaal (GARS). Handleiding. 1-31. 1993. Groningen., Noordelijk Centrum voor Gezondheidsvraagstukken (Report).
54. Suurmeijer TP, Doeglas DM, Moum T, Briancon S, Krol B, Sanderman R, Guillemin F, Bjelle A, Heuvel WJ van der. The Groningen Activity Restriction Scale for measuring disability: its utility in international comparisons. *Am J Public Health* 1994;84(8):1270-3.
55. Leigh JP, Fries JF. Predictors of disability in a longitudinal sample of patients with rheumatoid arthritis. *Ann Rheum Dis* 1992;51(6):221-7.
56. Wolfe F, Cathey MA. The assessment and prediction of functional disability in rheumatoid arthritis. *J Rheumatol* 1991;18(9):1298-306.
57. Ende CHM van den, Breedveld FC, Dijkmans BA, Hazes JM. The limited value of the Health Assessment Questionnaire as an outcome measure in short term exercise trials. *J Rheumatol* 1997;24(10):1972-7.
58. Nordenskiold U, Grimby G. Assessments of disability in women with rheumatoid arthritis in relation to grip force and pain. *Disabil Rehabil* 1997;19(1):13-9.

- 59 Pincus T, Summey JA, Soraci SA, Wallston KA, Hummon NP. Assessment of patient satisfaction in activities of daily living using a modified Stanford Health Assessment Questionnaire. *Arthritis Rheum* 1983;26(11):1346-53.
- 60 Daltroy LH, Larson MG, Roberts NW, Liang MH. A modification of the Health Assessment Questionnaire for the spondyloarthropathies. *J Rheumatol* 1990;17(7):946-50.
- 61 Siegert CE, Vleming LJ, Vandenbroucke JP, Cats A. Measurement of disability in Dutch rheumatoid arthritis patients. *Clin Rheumatol* 1984; 3(3):305-9.
- 62 Wolfe F, Kleinheksel SM, Cathey MA, Hawley DJ, Spitz PW, Fries JF. The clinical value of the Stanford Health Assessment Questionnaire Functional Disability Index in patients with rheumatoid arthritis. *J Rheumatol* 1988;15(10):1480-8.
- 63 Husted JA, Gladman DD, Long JA, Farewell VT. A modified version of the Health Assessment Questionnaire (HAQ) for psoriatic arthritis. *Clin Exp Rheumatol* 1995;13(4):439-43.
- 64 Gardiner PV, Sykes HR, Hassey GA, Walker DJ. An evaluation of the Health Assessment Questionnaire in long-term longitudinal follow-up of disability in rheumatoid arthritis. *Br J Rheumatol* 1993;32(8):724-8.
- 65 Fitzpatrick R, Newman S, Lamb R, Shipley M. A comparison of measures of health status in rheumatoid arthritis. *Br J Rheumatol* 1989;28(3):201-6.
- 66 Hurst NP, Kind P, Ruta D, Hunter M, Stubbings A. Measuring health-related quality of life in rheumatoid arthritis: validity, responsiveness and reliability of EuroQol (EQ-5D). *Br J Rheumatol* 1997; 36(5):551-9.
- 67 Blalock SJ, Sauter Sv, DeVellis RF. The Modified Health Assessment Questionnaire Difficulty Scale. A health status measure revisited. *Arthritis Care Res* 1990;3(4):182-8.
- 68 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(6):461-5.
- 69 Blackmore MG, Gladman DD, Husted J, Long JA, Farewell VT. Measuring health status in psoriatic arthritis: the Health Assessment Questionnaire and its modification. *J Rheumatol* 1995;22(5):886-93.
- 70 Brazier JE, Harper R, Munro J, Walters SJ, Snaith ML. Generic and condition-specific outcome measures for people with osteoarthritis of the knee. *Rheumatology (Oxford)* 1999; 38(9):870-7.
- 71 Fries JF, Spitz P, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. *Arthritis Rheum* 1980;23(2):137-45.
- 72 Huiskes CJAE, Kraaijmaat FW, Bijlsma JW. De ontwikkeling van de IRGL. *Gedrag & Gezondheid* 1990;18(2):78-89.
- 73 Huiskes CJAE, Kraaijmaat FW, Bijlsma JWJ. Development of a self-report questionnaire to assess the impact of rheumatic diseases on health and lifestyle. *J Rehabil Sciences* 1990;3(2):65-70.
- 74 Borstlap M, Zant JL, Soesbergen RM van, Korst JK van der. Quality of life assessment: a comparison of four questionnaires: for measuring improvements after total hip replacement. *Clin Rheumatol* 1995; 14(1):15-20.
- 75 Bijlsma JW, Huiskes CJ, Kraaijmaat FW, Veen MJ Van der, Huber-Bruning O. Relation between patients' own health assessment and clinical and laboratory findings in rheumatoid arthritis. *J Rheumatol* 1991;18(5):650-3.
- 76 MacKenzie CR, Charlson ME, DiGioia D, Kelly K. Can the Sickness Impact Profile measure change? An example of scale assessment. *J Chron Dis* 1986;39(6):429-38.
- 77 Mattsson E, Brostrom LA. The physical and psychosocial effect of moderate osteoarthrosis of the knee. *Scand J Rehabil Med* 1991;23(4):215-8.
- 78 Deyo RA, Inui TS, Leininger JD, Overman SS. Measuring functional outcomes in chronic disease: a comparison of traditional scales and a self-administered health status questionnaire in patients with rheumatoid arthritis. *Med Care* 1983;21(2):180-92.

- 79 Bergner M, Bobbitt RA, Carter WB, Gilson BS. The Sickness Impact Profile: development and final revision of a health status measure. *Medical Care* 1981;19(8):787-804.
- 80 Bruin AF de, Buys M, Witte LP de, Diederiks JP. The sickness impact profile: SIP68, a short generic version. First evaluation of the reliability and reproducibility. *J Clin Epidemiol* 1994;47(8):863-71.
- 81 Deyo RA, Inui TS. Toward clinical applications of health status measures: sensitivity of scales to clinically important changes. *Health Services Research* 1984;19(3):275-89.
- 82 Martin DP, Engelberg R, Agel J, Swiontkowski MF. Comparison of the Musculoskeletal Function Assessment Questionnaire with the Short Form-36, the Western Ontario and McMaster University Osteoarthritis Index, and the Sickness Impact Profile Health-status measures. *J Bone Joint Surg* 1997;79.A(9, sept.).
- 83 Witte L de, Jacobs H, Horst F van der, Lutik A, Joosten J, Philipsen H. De waarde van de Sickness Impact Profile als maat voor het functioneren van patiënten. *Gezondheid & Samenleving* 1987;8(2):120-7.
- 84 Sullivan M, Ahlmen M, Bjelle A. Health status assessment in rheumatoid arthritis. Further work on the validity of the sickness impact profile. *J Rheumatol* 1990;17(4):439-47.
- 85 Lankveld W van, Pad BP van't, Bakker J, Terwindt S, Franssen M, Riel PCLM van. Sequential occupational dexterity assessment (SODA): a new test to measure hand disability. *J Hand Ther* 1996;9(1):27-32.

Chapter 7

General discussion

General discussion

Our systematic review of the literature until the end of April 2001 generated 3,110 documents. Based on the abstracts, 890 out of these 3,110 documents were considered to be relevant for this study. This generated a total of 209 measurement instruments, 104 (57%) of which assessed impairments in body structure and function, and 78 (43%) assessed disabilities in activities and problems in participation. After data reduction, 149 measurement instruments remained (67 actual measurement instruments and 82 sub-scales that can be scored separately). Finally, 239 documents were included in the data-analysis (see Appendix 1).

In the literature, as well as in this thesis, the use of different terminology for 'clinimetric properties' and 'psychometric properties' might be confusing. In order to avoid misinterpretation this should be elucidated. The term 'clinimetrics' was introduced by Feinstein in 1987 (1), while for the same phenomena the term 'psychometric properties' was used in psychological disciplines. In more recent literature it is proposed that the term 'psychometric' should be used for unidimensional scales and 'clinimetric' for multidimensional indexes (2). In that way, the two terms are not contradictory. De Vet et al. (3) emphasize that these two terms have different meanings: 'clinimetric' refers to instruments that measure multiple constructs with a single index (e.g. the AIMS), while 'psychometric' refers to instruments that measure a single construct with multiple items (e.g. the Beck Depression Index). However, none of these distinctions, based either on the number of dimensions or the principle of item selection, are fully exclusive, thus provoking the (undesired) interchangeability of the terms 'psychometric' and 'clinimetric'.

An important goal in the use of measurement instruments in clinical practice, and also in scientific research, is to objectify and quantify symptoms and signs in a standardized way. In order to do this correctly, the instruments must be reliable, valid, responsive, feasible and useful. Reliability refers to the degree to which individuals can be distinguished from each other, despite measurement errors;. validity refers to the degree to which an instrument measures what it is intended to measure; responsiveness is the sensitivity of an instrument to detect changes over time. Feasibility concerns aspects such as the time required for the assessment, availability in the required language, costs, and a clear format for the interpretation of the measurement data, and usefulness implies that the measurement instrument generates useful information in addition to the data obtained from anamnesis and physical examination.

In Chapter 1 it was stated that the term 'reproducibility' includes two concepts: 'agreement' (lack of measurement error) and 'reliability' (the extent to which individuals can be distinguished from each other, despite measurement errors). In fact, for instruments that are intended to measure changes in health status over time, agreement is much more relevant than reliability (4;5). This distinction could not be included in our data-analyses, because in the majority of studies the agreement coefficients were not mentioned. In theory, if the research question requires a distinction between persons, reliability parameters are the most appropriate, but if the aim is to measure change in health status, information about the absolute measurement error is preferable (6). Therefore, in clinical practice agreement parameters will be more informative than reliability parameters, because clinicians are usually quite familiar with the variability between persons and the changes that are clinically important.

In Chapter 2 it is concluded that for use in clinical practice the majority of questionnaires are intended to measure the status of the patient (diagnostic: 59.9% (7)) and that the evaluative capacity of the instruments is less often at issue. In fact, although virtually all the instruments that are used in clinical physiotherapy settings are intended to be evaluative instruments, responsiveness has been investigated in only 36.3% (not all proved to be responsive), and agreement of the instruments is scarcely mentioned. This might be due to the fact that the above-mentioned theory on how to assess reproducibility has only very recently been published in the literature, and was certainly not common at the time of our literature review. Responsiveness is reported in the literature regularly, but assessed in many different ways, which made it impossible to draw conclusions.

This thesis addresses the question of which instruments are the most optimal, and should be recommended with regard to reliability and construct validity. It also evaluates the appropriateness of the International Classification of Functioning, Disability and Health (ICF) as a framework for analysis of the literature in the present study and discusses the relevance and (dis)advantages of the ICF in relation to outcome assessment and the method of data-analysis that was applied. We also discussed the hypothesis that validation against a similar construct results in higher correlation values than validation against a dissimilar construct, as well as the consequences with regard to the interpretation of the results of our study. Finally, in this General Discussion the results of our study are discussed in relation to the results reported by the OMERACT group. This chapter concludes with the implications of this study for clinical practice and research, and some recommendations for future research.

All the abbreviations of the measurement instruments are listed in Appendix 3.

ICF and outcome assessment

One of the aims of the present study was to give an overview of the development and content of the International Classification of Functioning, Disability and Health (ICF) (8) and to investigate the degree to which various measurement instruments measure the different domains of the ICF in patients with rheumatic disorders (Chapter 2). In addition, we wanted to provide a framework based on the ICF for the selection of measurement instruments in general, and questionnaires in particular (Chapter 2).

A brief overview of the development and content of the ICF is given in Chapter 2. Of the 209⁵ measurement instruments and questionnaires that were identified, 57.1% are inventories of (impairments in) body structures and function, 37.4% are inventories of (limitations in) activities, and 5.5% measure problems in participation. It is concluded that there are no available instruments or questionnaires which cover all domains of the ICF. Furthermore, we conclude that the ICF provides a useful conceptual framework as a starting point for the classification of available (anthropometrical) measurement instruments and questionnaires, in particular for the assessment of patients with chronic diseases, because the ICF classifies the *consequences* of these diseases. However, the ICF was not developed with the intention of creating measurement instruments or questionnaires; the aim was to classify the different dimensions of a health problem. That gives rise to the question of how useful the ICF is as a framework for analysis of the domains and properties of measurement instruments, such as the goals of the measurement and the validity of the psychometric properties. We assume that the ICF classification is also suitable as a basis on which to formulate the different constructs that are used in validity studies. However, many questionnaires do not conform with the ICF, possibly because the majority of questionnaires were developed in the 1970s and 1980s, before the ICIDH (which later became the ICF) was introduced. In Chapter 2, a 5-step model, based on the ICF, is presented for the selection of measurement instruments for use in clinical settings. Under Step 5 it is stated that the frequent use of questionnaires generates large quantities of data, which makes correct interpretation easier. To avoid misinterpretation, it should be emphasized that this statement is only correct under the condition that data are collected in the same way and with the same goals, in order to make it possible to compare all the data collected in different studies. Useful data can then be generated to provide reference values for the (correct)

⁵ The numbers mentioned in Chapter 2 do not correspond with the numbers mentioned in the other chapters, because in the results of the other chapters the instruments are filtered for the target groups in this study (physiotherapy and rehabilitation medicine). The results in Chapter 2 also contain outcome measures that are not only specifically appropriate for those groups, but also for rheumatologists.

interpretation of scores. In more recent literature it is strongly recommended that the ICF framework is used to review, identify and describe patient problems, and that condition-specific ICF Core Sets are used to check the problems that are typical for a given condition (9). The 2004 ICF Core Set for rheumatoid arthritis is presented in Table 7.1. However, this consists of a core set of domains to be covered by measurement instruments to assess patients with rheumatic disorders, but does not contain recommendations for the most optimal measurement instruments (10). The consequence of this will be discussed later in this Chapter under the heading 'OMERACT improvements'.

The results of our study, with regard to the most optimal measurement instruments, are very much in agreement with the ICF Core Set presented in Table 7.1. However, one of the most remarkable deviations is the category of 'Gait pattern functions': only 15% of the experts⁶ would include this in the ICF Core Set. Furthermore, we investigated instruments for the measurement of impairments in mental functions, but such impairments are not mentioned at all in the ICF Core Set. This is really amazing, because due to the multidimensional character of rheumatic disorders, which affect the whole person, and due to the chronic course of rheumatic disorders, they can have a major impact on the mental health of a patient.

⁶ The experts were 7 physicians, 7 rheumatologists, 1 nurse, 1 occupational therapist, and 1 physical therapist.

Table 7.1 ICF Core Set for rheumatoid arthritis and percentage of experts willing to include the category at issue in the Brief ICF Core Set. 50% represents a preliminary cut-off. >50% is shown in bold typeface (adapted from Stucki et al., 2004 (10)).

ICF domain	%	ICF category title
Body functions	100	Sensations of pain
	100	Mobility of joint functions
	85	Muscle power functions
	70	Exercise tolerance functions
	65	Sensations related to muscles and movement functions
	15	Gait pattern functions
	5	Sleep functions
Activities and participation	90	Walking
	90	Remunerative employment
	75	Fine hand use
	75	Changing basic body position
	65	Hand and arm use
	65	Carrying out daily routine
	45	Lifting and carrying objects
	40	Using transportation
	30	Dressing
	30	Washing oneself
	30	Recreation and leisure
	25	Intimate relationships
	10	Work employment
	5	Eating

Similar and dissimilar constructs

The ICF domains were also used as a basis for the classification of the constructs used in validation studies (see Chapter 3 and Table 7.1). To investigate the influence of validation against different constructs (varying from optimally comparable to imperfectly comparable) the construct validity was divided into five clusters, in which the constructs against which a measurement instrument is validated are defined according to their anticipated degree of similarity to the instrument at issue. The contents and background of these clusters are described in Chapter 4 under the heading '*Methodological criteria for psychometric properties*'. These clusters were established by consensus of a working group, consisting of two ICF experts, a rheumatologist, an epidemiologist, a physiotherapist and two research fellows. With regard to measurement instruments for the assessment of impairments in body

structures and function, only a small number of instruments (16%) were validated against level 1 (the most optimal comparable level), and there is a trend that validation against a similar construct yields higher correlation values than validation against a dissimilar construct. Our data imply that a majority of 80% of the investigated instruments have been validated against constructs that measure another impairment and/or limitation in activity, or other aspects than the domain to be validated. From the results of the present study it seems that the correlations for construct validity are, in general, weaker when validated against the most dissimilar construct (level 6 construct), compared to the correlations when validated against a level 1 or level 2 construct. However, in seven out of 22 measurement instruments the opposite was found. An explanation for this phenomenon might be that weak correlations of certain sub-scales are compensated by strong correlations of other sub-scales of the same measurement instrument that are more similar to the general construct against which they are validated. Furthermore, it could be that the chosen impairment construct of the measurement instrument to be validated does not cover the intended construct, or covers a broader construct, therefore possibly resulting in a different correlation than expected if validated against a dissimilar construct.

However, with regard to the results concerning measurement instruments for the assessment of disabilities in personal care, no difference was found in the strength of correlation between validation against the most similar construct versus validation against the least similar construct. For the assessment of disabilities in gait and gait-related activities, validation against a similar construct even resulted in lower correlations than validation against a less similar construct in 10 out of 18 instruments. A possible explanation could be the fact that the large majority of instruments for the assessment of gait and gait-related activities are multi-dimensional, and predominantly focus on limitations in activities (disabilities) other than walking, whereas instruments for the assessment of impairments more frequently only focus on the impairment to be measured, and not on other domains. The different results for instruments to assess of impairments in body structure and function, on the one hand, and instruments to assess of limitations in activities, on the other hand, can possibly be explained by the fact that weak correlations of certain sub-scales are compensated by strong correlations of other sub-scales that are more similar to the general construct against which they are validated. Based on this assumption, it could be concluded that multidimensional questionnaires need to be validated against multidimensional constructs, or separately for each sub-scale against similar constructs.

Another explanation for the puzzling discrepancy we found might be the correlation between impairments in body structures and functions, on the one hand, and limitations in activities, on the other hand. For example, the correlation between pain

and limitations in the activities of daily living is $r=0.39$ in patients with low back pain (11). Comparable moderate to poor correlations are demonstrated in patients with osteoarthritis (12;13). Probably more than one impairment in body structures or function is present. In addition to pain, high scores for disease activity, stiffness and swollen joints might result in considerable limitations in the activities of daily living. Finally, it could be that the chosen impairment construct of the measurement instrument to be validated does not cover the intended construct, or covers a broader construct, therefore possibly resulting in an unexpected correlation if validated against a dissimilar construct.

Already in the eighties, Deyo concluded that physical scales (related to impairments) show expected correlations with accepted clinical measures when these are available, but that social function scales are much more problematic (14). An explanation could be the fact that, in general, impairments are unidimensional, and therefore less dependent on other factors, whereas disabilities in activities and social function are multidimensional, and depend on many more variables.

As mentioned before, the ICF domains were used as a basis for the classification of constructs that were used for more similar or less similar validation. The reason for this was the fact that the measurement instruments and/or sub-scales to be validated were also classified according the ICF, and a measurement instrument was considered to have good construct validity if two criteria were met: 1) the correlation coefficient was 0.50 or higher, and 2) the measurement instrument in question was validated against similar constructs (level 1 and level 2). As far as we know, no other criteria to evaluate and compare the clinimetric properties of measurement instruments were available at the start of our study. More recently, however, more elaborate criteria have been developed for clinimetric studies (15). This recently developed checklist describes twelve properties: content validity, readability and comprehension, internal consistency, construct validity, floor and ceiling effects, reliability, agreement, responsiveness, interpretability, minimal clinically important difference, time to administer, and administration burden.

According to this checklist, construct validity should be assessed by testing predefined hypotheses (which need to be as specific as possible). Terwee et al. (15;16) suggest that construct validity should be scored positively if hypotheses are specified in advance, and if at least 75% of the results agree with these hypotheses in (sub-)groups of at least 50 patients. If this criterion was applied to the data of the present study, only very few measurement instruments would be classified as 'valid', because for a great majority less than 75% of the results agreed with the hypotheses (the construct against which the instrument was validated). This could be due to the fact that the construct validity really is too low. However, it could also be due to the fact that the ICF is not suitable, either to classify the measurement instruments to be

validated or to classify the constructs that are used for validation. Certainly, the ICF was never developed with the intention to create assessment instruments or questionnaires, but to classify the different domains of a health problem. Furthermore, measurement instruments, and questionnaires in particular, are rarely developed on the basis of the ICF framework. This could also justify our a posteriori conclusion that the ICF is probably not the most optimal framework for the analysis of measurement instruments.

Validity of questionnaires: perception versus performance

Another point of discussion that could be relevant is the fact that the great majority of measurement instruments are questionnaires, and most questionnaires are completed by the patient. This implies that the measurement reflects the patient's perception of his or her own (dis)abilities. This is not a problem if perception is what the instrument is intended to measure, but bias could occur if the intended measurement is of a more objective nature. This does not mean that the validity of the questionnaire is under debate, but probably that the patient's perception could be a source of bias in completing the questionnaire, as has been demonstrated in some studies. Reneman et al. (17) investigated the concurrent validity of two approaches (self-report questionnaires and performance tests) to measure disabilities in patients with chronic low back pain. Spearman's rank correlations indicated poor to moderate correlation, and it is concluded that the results of questionnaire and performance-based disability measurements differ substantially. In the majority of cases, patients under-estimate their capacities when completing a questionnaire. These findings are sustained in other studies measuring the agreement between self-evaluation and the observation of performance (18-21).

This could also explain the difference in the results of instruments to assess of impairments in structures and function and the results of instruments to assess of limitations in personal care activities and limitations in gait and gait-related activities, with regard to the hypothesis that construct validity results in higher correlations when validated against a more suitable construct.

Because of the poor correlations between perception-based scores (questionnaires) and performance-based scores (performance tests), it is recommended that both a performance test and a questionnaire are used to obtain a more comprehensive assessment of the disability of patients with chronic disorders. However, in order to extrapolate the conclusions concerning patients with chronic low back pain (22), it is recommended that such validity studies should be repeated for patients with rheumatic disorders.

Responsiveness of instruments

'Responsiveness' has become a rising topic in the literature during the past decade. In particular, in the past 5-7 years there have been significant innovations concerning the theory about responsiveness. Nowadays, 'responsiveness' is defined as the ability of an evaluative measurement instrument to detect clinically relevant changes in health status over time (22). In fact, responsiveness can be considered as longitudinal validity. The most frequently used techniques to determine the responsiveness are the Receiver Operator Curve (ROC), where the area under the curve is regarded as an indicator for responsiveness, and the Effect Size, being a statistical expression of the magnitude of the difference between two treatments or the magnitude of a relationship between two variables. Furthermore, in the literature the term 'sensitivity' is sometimes considered to be synonymous with 'responsiveness' (14). Recently, however, it is becoming more clear that the different terms are confusing, and need to be redefined. In fact, the effect size is more an indicator for the effectiveness of a treatment than for the responsiveness of a measurement instrument. It would be more relevant from this point of view to determine the 'minimally detectable important change' that can be detected by an instrument. The 'minimally detectable important change' is a stronger indicator of the ability of an instrument to detect changes over time than for the effectiveness of a treatment. In this respect it is relevant to emphasize the difference between the 'Smallest Detectable Difference' (SDD), defined as the smallest difference that can be measured without bias, and the 'Minimally Important Change (MIC), which is the smallest change that is clinically relevant. When the SDD is smaller than the MIC, the measurement instrument is useful for evaluation of the course in a clinical setting. If the SDD is larger than the MIC, the measurement instrument at issue could be useful for scientific research, because the SDD-group will improve with \sqrt{N} . Crosby et al. recently published an extensive overview of methods to determine MIC (23). In the literature we retrieved (until 2001), the use of variation as a kind of longitudinal construct validity is rare. The majority of statistical parameters that are used to express changes and variation in the course, like effect size, standardized response mean, or Guyatt responsiveness statistic, are not always interchangeable and comparable. The area under the curve and the effect size are the most frequently used techniques, (the terms 'minimally important change' and 'Smallest Detectable Difference' are mainly used in the more recent literature), and therefore the conclusions in the reviewed studies with regard to the responsiveness of an instrument do not add anything to the discussion concerning changes in the health status. Summarizing, many different responsiveness statistics are used, implying that it was not possible to draw final conclusions with regard to responsiveness in our

reviews. We therefore decided not to include data on responsiveness in the majority of the results of our reviews. However, we realise (certainly a posteriori) that due to this advance in knowledge over time we came to the conclusion that an essential element in the clinimetric properties of measurement instruments is actually lacking, especially taking into account the increasing body of knowledge and the number of responsiveness studies that have been performed during the past 5-10 years.

The best measurement instruments

The primary aim of this thesis was to generate an overview of available instruments to assess patients with rheumatic disorders and to qualify these instruments on the basis of their clinimetric properties. The results with regard to reliability and (construct) validity are summarized in Table 7.2. If available, a maximum of three instruments for each relevant domain of the ICF are included, depending on the available evidence with regard to reliability and validity. For example, for the measurement of impairments in mental functions the AHI proved to have the lowest construct validity, but for the STAI and the BDI no data are available with regard to their reliability among patients with rheumatic disorders. The reliability of the STAI and the BDI was mainly tested in other patient populations.

With regard to the measurement of disabilities in gait and gait-related activities, the Top3 optimal instruments are the WT, the RAQoL and the HAQ, and the RAQoL proved to be the most reliable and most valid instrument. However, this

Table 7.2 The most optimal measurement instruments for impairments in body function and limitations in gait and gait-related activities and in personal care in patients with rheumatic disorders.

Measurement instrument / sub-scale	Intra-rater reliability	Construct validity	Remarks
Measurement of impairments in mental functions			
STAI		0.68*	Reliability investigated in other populations than patients with rheumatic disorders
BDI		0.67*	Reliability investigated in other populations than patients with rheumatic disorders
AHI	0.53	0.46*	
Measurement of impairments in sensation (stiffness)			
BASDAI	0.74	0.70	Mainly intended for patients with ankylosing spondylitis
VAS-S	0.93	0.51-0.91*	0.91* if validated against disability measures
Measurement of impairment in sensory functions (pain)			
AUSCAN-OHI	0.84	0.65	
RAPS	0.92	0.52-0.68	
VAS-P	0.80	0.82*	
Measurement of impairments in joint mobility			
Gonio	0.89	0.92	Applicable for every joint
Shob	0.95	0.66*-0.88*	Validity value depending on construct used for validation
Chest	0.95	0.60	Mainly relevant for patients with ankylosing spondylitis
Measurement of impairments in muscle force			
Sphy	0.93	0.87*	
Measurement of joint swelling			
AI	0.80*-0.83*	0.88*	
Measurement of disabilities in gait or gait-related activities			
WT	0.89*-0.91	0.66	Validated against other domains (impairments, and general aspects)
RAQoL	0.90	0.87	Questionnaire for health-related quality of life, of which disabilities in gait are only a part
Measurement of disabilities in personal care			
AIMSD	0.86*	0.65	
HAQ	0.94*	0.80	Validated against instruments that measure the same disability as well as other disabilities

* pooled value.

is an instrument to assess health-related quality of life, of which gait and gait-related activities are only a part. It also measures many other dimensions, and not only gait and gait-related activities. Consequently, the WT seems to be more appropriate, although the WT has been validated against others domains (impairments, and general aspects), and it is not fully clear what implications this has for its validity. The HAQ is reliable, and seems to have good construct validity, but the reported values concern the total HAQ score. Data-analyses for each HAQ sub-scale demonstrated that only the sub-scale 'walk' had a poor correlation with an optimally suitable

construct ($r=0.23$). A possible explanation for this was discussed earlier in this chapter (under 'Similar and dissimilar constructs').

OMERACT improvements and relationship with the results of the present study

OMERACT (Outcome Measures in Rheumatoid Arthritis Clinical Trials) is endorsed by the International League of Associations for Rheumatology (ILAR), and aims to establish a core set of outcome measures for future clinical trials concerning rheumatic disorders (24;25). The preliminary core set of outcome measures, as defined by the OMERACT and the American College of Rheumatology (ACR) for rheumatoid arthritis clinical trials, was as follows (26):

- Acute-phase reactants
- Disability
- Joint pain/tenderness
- Joint swelling
- Pain
- Patient global assessment
- Physician global assessment
- Radiographs for studies of 1 year or longer

It should be noted that these criteria should not be used as diagnostic criteria but for the classification of patients in studies (26). The OMERACT group has carried out a lot of work, classified domains, established a core set of outcome measures, and created a paradigm to capture the essential elements of an outcome measure: the OMERACT filter. The OMERACT filter has been developed to assess the applicability of measures in a certain setting. The word "applicable" is intended to include all aspects necessary for the appropriate selection of a measure. This filter has three components: truth, discrimination and feasibility (25), and is, in particular, intended for use in clinical research, and not primarily for use in clinical practice. The component 'truth' applies to validity, which is discussed extensively in our study. The component 'discrimination' includes (among other things) responsiveness. The restrictions with regard to responsiveness in our study have been discussed earlier in this chapter, under 'Responsiveness of instruments'. The most important difference between the OMERACT filter on the one hand, and the present study on the other hand, concerns the third factor: 'feasibility'. Feasibility refers to how easily the measure can be applied, given the constraints of time, money, availability and interpretability (are reference data and interpretation key available and easy to use?). These aspects are not mentioned in the present study, although they may be decisive in determining the success of a measure. Later, based on a consensus process, an ICF Core Set was developed for rheumatoid arthritis (28). This ICF Core Set can certainly serve as a framework for comparison of the results of the present

study with the Core Set: both are based on the ICF, and the ICF Core Set is very similar to the OMERACT outcome measures (see Table 7.1). However, this Core Set also describes the *domains* of the ICF that are considered to be most relevant, and does not focus on specific measurement instruments. When we compare the results of the present study in Table 7.2, the most optimal instruments for the assessment of impairments in body functions and structures correspond with the five impairments in body structures and functions that are considered to be the most important in Table 7.1. The exception is measurement of impairments in mental functions, which is not mentioned at all in the ICF Core Set. There are health status measurement instruments included in the recommendations made by OMERACT, ASAS (Assessment in Ankylosing Spondylitis) and the WHO. 'Health status' or 'quality of life' embodies the dimensions of physical, social and emotional function. If these concepts are attributed to health, the term 'health-related quality of life' is used. However, there is considerable variation in the terminology and interpretation of this concept. Health status measurement instruments usually include the dimensions of physical function, social function, emotional function, pain and the perception of well-being. However, to complicate matters, there is unfortunately no clear relationship with the ICF framework (29). On the contrary, in the publication of the ICF Core Set for rheumatoid arthritis there is hardly any reference to the, recently increasing, number of OMERACT outcome measures.

With regard to the results concerning limitations in activities and participation, walking was found to be the most important (90% consensus, Table 7.2), but for disabilities in personal care there was only 30% consensus for dressing and washing oneself. However, the domains that were found to be the most relevant concerned fine hand use, changing body position, hand and arm use, and carrying out daily routine. Those aspects are all essential for the activities of personal care. So, in fact there is a substantial overlap between the ICF Core Set and the domains investigated in the present study. Unfortunately, neither the OMERACT nor the ICF Core Set focuses on specific measurement instruments in detail, as we did in our reviews. The results of our study could possibly be useful for further development of the OMERACT and the ICF Core Set with regard to the level of measurement instruments to be selected. It must be borne in mind, however, that the goals of OMERACT and the goals of the present studies are not identical: the intention of the OMERACT group is to describe outcome measures to be used in future clinical trials, while our review focuses in particular on use in clinical situations. This difference could also have consequences with regard to the criteria that are applied for the choice of the measurement instruments and the criteria for methodological quality: in research it is easier to deal with random measurement errors by calculating the averages for a group of patients, while in clinical practice there is typically only one measurement per patient, so it is

not possible to average the measurements. Double and triple measurements would be an option, but this also has disadvantages, such as the time involved, and lack of reproducibility of the measurements. Therefore, the criteria for measurements in clinical practice should be even more severe than for measurements in clinical research (29).

Implications for clinical practice

For a correct interpretation of outcome assessment it is also important to distinguish between measurement instruments to be used in a clinical setting (level of individual patient) and those to be used for scientific research (level of groups). As mentioned above, the criteria for measurements in clinical practice should be set higher than for those in clinical research. This has actually already been incorporated in our review, because the chosen cut-off points are reasonably high, compared with the cut-off points used in the majority of clinimetric studies. This leads to the conclusion that, from this point of view, the selection of the most optimal instruments, as reported in Table 7.2 of this chapter, will probably be appropriate for application in clinical practice. However, this selection does have some limitations if the measurement instruments are to be used to evaluate treatment effect, because in general there is a lack of information concerning the responsiveness of the instruments. Conflicting outcomes are reported for responsiveness, probably because there is lack of consensus regarding the best method for the evaluation of responsiveness (15).

In conclusion, the most optimal instruments listed in Table 7.2 are of good methodologically quality and may be useful in clinical practice. However, there is insufficient evidence with regard to the responsiveness of these instruments.

In clinical situations it is recommended that condition-specific instruments are used as much as possible in order to check for problems that are typically encountered in patients with a given condition. From this point of view it is important to determine in advance what the instrument is intended to measure. Also relevant for the use of measurement instruments in clinical settings are the aspects of feasibility and usefulness: they should preferably not be too expensive, must be available (in case of a questionnaire) in the required language, should take only a reasonably short time to complete (maximum 10-15 minutes) and, if possible, there should be reference values for a correct interpretation.

Another limitation of our study is the fact that we focus mainly on assessments in rehabilitation medicine and physiotherapy, which implies that acute-phase reactants, blood tests and radiography are not incorporated. This is a disadvantage in comparison with the OMERACT Core Set of outcome measures.

The results of a large-scale study investigating the use of outcome measures in rehabilitation medicine within Europe demonstrate that a large number of measures

are being used in only a small number of locations and for relatively few patients (30). This leads to the conclusion that the majority of clinicians do not use measurement instruments. It is therefore recommended that the use of instruments in clinical situations is promoted, in order to objectify and quantify clinical findings and the results of treatment, and to achieve more standardization in data-collection. Furthermore, it is recommended that standardized measurement instruments should be used to monitor the course of (certain aspects of) a disorder.

Implications for clinical research

Based on the results of the present study, it is recommended that both self-assessment questionnaires and performance-based tests should be used, if available and of good quality.

Taking into account the enormous amount of self-assessment questionnaires that are available, it would be helpful if there was a checklist for rating clinimetric quality, in order to separate the wheat from the chaff. At the time of the present study there was no such checklist available, so we had to develop one ourselves. Meanwhile, Bot et al. have developed a checklist to rate the clinimetric quality of questionnaires (16), and we therefore recommend that their checklist is used for the selection of questionnaires to evaluate health status in clinical trials. The main difference between this checklist and the list we used is that our list also indicated the goals of the questionnaire and the ICF domains that it covers, and described the feasibility in more detail (time required to complete the questionnaire, availability in several languages, availability of an interpretation key, etc.). With regard to the reliability and validity, we also distinguished between the correlation measures that were used. On the other hand, the checklist developed by Bot et al. focuses in more detail on methodological aspects, such as floor and ceiling effects, and the minimal clinically important difference. The Bot checklist was further improved in 2005, resulting in a more definite list of quality criteria for the design, methods and outcome of clinimetric studies (16). It is recommended that these criteria are applied in future research so that the clinimetric properties of questionnaires can be evaluated and compared in a standardized way.

Finally, we would like to emphasize that, although we carried out an extensive systematic review and analysis of the literature, we performed no quality assessment of the sources. This implies, that weak quality studies could also be included. It is open for discussion what impact this potential source of bias could have had on the final results of this study.

Recommendations for future research

In research on the 'reproducibility' of measurement instruments, it is recommended that a distinction is made between 'reliability' and 'agreement', and that the data on agreement are also published.

In the literature there is a substantial lack of information regarding the responsiveness of measurement instruments. It is essential to be able to measure minimally important clinical changes in order to evaluate an improvement or deterioration in health status (31). Furthermore, changes in scores on health status questionnaires and performance tests are difficult to interpret. Firstly, because of the difference in the methods used to calculate responsiveness, and secondly, due to the many conflicting results of studies with regard to the responsiveness of measurement instruments.

Therefore, there is a need for additional and standardized research on the responsiveness of these instruments in homogeneous populations with rheumatic disorders, based on the same method for the measurement of responsiveness outcomes and the same criteria for the interpretation of changes in scores. Crosby et al. (23) published an extensive overview of the different methods to determine minimally important change, distinguishing between anchor-based and distribution-based approaches. The anchor-based approaches described in that review lacked a measure of precision. More recently, De Vet et al. introduced an integrated method that combines the advantages of an anchor with measures of precision to establish cut-off points based on a desired confidence level (31). It is recommended that in future research this method is used to assess the responsiveness of measurement instruments.

Furthermore, it is recommended that the criteria proposed by Terwee et al. (15) are applied in future (validity) research in order to evaluate and compare the clinimetric properties of questionnaires in a more standardized way. This could be supplemented by more studies on construct validity, distinguishing between the different degrees of similarity of the constructs, in order to confirm the trend observed in the present study. This hypothesis should be tested statistically in order to reinforce our conclusions with regard to this hypothesis.

Finally, in this thesis we emphasize that it is desirable that the OMERACT and the ICF Core Sets are further developed for measuring populations with rheumatic disorders, not only with regard to the different domains to be included, but also the level of the measurement instruments to be selected.

Reference List

- 1 Feinstein AR: Clinimetrics. New Haven, Yale University Press, 1987.
- 2 Fayers PMHDJ: Causal variables, indicator variables and measurement scales: an example from quality of life. *J R Stat Soc* 2002;(165):233-61.
- 3 Vet HCW de, Terwee CB, Bouter LM: Variance and dissent. *Clinimetrics and psychometrics: two sides of the same coin. J Clin Epidemiol* 2003; 56:1146-7.
- 4 Guyatt GH, Kirshner B, Jaeschke R: Measuring health status: what are the necessary measurement properties? *J Clin Epidemiol* 1992; 45(12):1341-5.
- 5 Vet HCW de, Bouter LM, Bezemer PD, Beurskens AJ: Reproducibility and responsiveness of evaluative outcome measures. Theoretical considerations illustrated by an empirical example. *Int J Technol Assess Health Care* 2001; 17(4):479-87.
- 6 Vet HCW de, Terwee CB, Knol D, Bouter LM: How to assess reproducibility: parameters of reliability and agreement. Submitted 2005.
- 7 Swinkels RAHM: The ICF-classification as a system for structuring outcome assessment. *Physiotherapy Singapore* 2004;7(3):7-13.
- 8 World Health Organization. International Classification of Functioning, Disabilities and Health Problems. 2001. Geneva, World Health Organization.
- 9 Stucki G, Sigl T: Assessment of the impact of disease on the individual. *Best Pract Res Clin Rheumatol* 2003;17(3):451-73.
- 10 Stucki G, Cieza A, Geyh S et al.: ICF Core Sets for rheumatoid arthritis. *J Rehabil Med* 2004;(44 Suppl):87-93.
- 11 Waddell G: The back pain revolution. 1st ed., Edinburgh, Churchill Livingstone, 1998.
- 12 Heuts PH, Vlaeyen JW, Roelofs J et al.: Pain-related fear and daily functioning in patients with osteoarthritis. *Pain* 2004;110(1-2):228-35.
- 13 Baar ME van Dekker J, Lemmens JA, Oostendorp RAB, Bijlsma JW: Pain and disability in patients with osteoarthritis of hip or knee: the relationship with articular, kinesiological, and psychological characteristics. *J Rheumatol* 1998; 25(1):125-33.
- 14 Deyo RA, Inui TS: Toward clinical applications of health status measures: sensitivity of scales to clinically important changes. *Health Services Research* 1984;19(3):275-89.
- 15 Terwee CB, Bot S, Boer MR de, Windt DAWM van der, Knol DL, Dekker J, Bouter LM, Vet HCW de: Quality criteria for clinimetric studies of health status questionnaires. Submitted 2005.
- 16 Bot SD, Terwee CB, Windt DAWM van der, Bouter LM, Dekker J, Vet HCW de: Clinimetric evaluation of shoulder disability questionnaires: a systematic review of the literature. *Ann Rheum Dis* 2004;63(4):335-41.
- 17 Reneman MF, Jorritsma W, Schellekens JM, Goeken LN: Concurrent validity of questionnaire and performance-based disability measurements in patients with chronic nonspecific low back pain. *J Occup Rehabil* 2002;12(3):119-29.
- 18 Gronblad M, Jarvinen E, Hurri H, Hupli M, Karaharju EO: Relationship of the Pain Disability Index (PDI) and the Oswestry Disability Questionnaire (ODQ) with three dynamic physical tests in a group of patients with chronic low-back and leg pain. *Clin J Pain* 1994; 10(3):197-203.
- 19 Gronblad M, Hurri H, Kouri JP: Relationships between spinal mobility, physical performance tests, pain intensity and disability assessments in chronic low back pain patients. *Scand J Rehabil Med* 1997; 29(1):17-24.
- 20 Novy DM, Simmonds MJ, Lee CE: Physical performance tasks: what are the underlying constructs? *Arch Phys Med Rehabil* 2002; 83(1):44-7.
- 21 Wijnhuizen GJ, Ooijendijk W: Measuring disability, the agreement between self evaluation and observation of performance. *Disabil Rehabil* 1999;21(2):61-7.
- 22 Vet HCW de, Terwee CB, Bouter LM: Current challenges in clinimetrics. *J Clin Epidemiol* 2003; 56(12):1137-41.
- 23 Crosby RD, Kolotkin RL, Williams GR: Defining clinically meaningful change in health-related quality of life. *J Clin Epidemiol* 2003; 56(5):395-407.

- 24 Bellamy N, Kirwan J, Boers M, Brooks P, Strand V, Tugwell P, Altman R, Brandt K, Dougados M, Lequesne M: 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(4):799-802.
- 25 Boers M, Brooks PV, Vibeke Strand C, Tugwell P: The OMERACT filter for outcome measures in rheumatology. *J Rheumatol* 1998;25(2):198-9.
- 26 Brooks P, Hochberg M: Outcome measures and classification criteria for the rheumatic diseases. A compilation of data from OMERACT (Outcome Measures for Arthritis Clinical Trials), ILAR (International League of Associations for Rheumatology), regional leagues and other groups. *Rheumatology (Oxford)* 2001; 40(8):896-906.
- 27 Bellamy N: Clinimetric concepts in outcome assessment: the OMERACT filter. *J Rheumatol* 1999;26(4):948-50.
- 28 Stucki G, Cieza A, Geyh S et al.: ICF Core Sets for rheumatoid arthritis. *J Rehabil Med* 2004;(44 Suppl):87-93.
- 29 Stucki G, Sigl T: Assessment of the impact of disease on the individual. *Best Pract Res Clin Rheumatol* 2003; 17(3):451-73.
- 30 Haigh R, Tennant A, Biering-Sorensen F et al.: The use of outcome measures in physical medicine and rehabilitation within Europe. *J Rehabil Med* 2001; 33(6):273-8.
- 31 Vet HCW de, Ostelo RWJG, Terwee CB, Roer N van der, Knol DL, Beckerman H, Boers M, Bouter LM: Minimally important change determined by a visual method integrating an anchor-based and distribution-based approach. Submitted 2005.

Appendix 1

Summary of the literature search

Appendix 1: Summary of the literature search

A literature search was performed to identify publications concerning clinimetrics in general (not specifically focussing on rheumatic disorders) in the databases of MEDLINE and the Documentation Centre of the Dutch Institute of Allied Health Care over the periods 1982-1997 and 1987-1997, respectively. The search was updated in May 2001.

Table 1 presents the results of the literature search in headlines.

Search:		Hits:	Identified:
I	DOCU, general	197	26
II	MEDLINE, general	11	7
III	DOCU, specific	280	159
IV	Additional search DOCU	251	20
V	Additional search MEDLINE	765	216
VI	DOCU, on names of instruments	398	58
VII	MEDLINE, on names of instruments	423	3
VIII	Tracking references		319
IX	Update May 2001	785	82
TOTAL		3110	890

Table 1 Summary of the results of the systematic literature searches
DOCU = Documentation Centre of the Dutch Institute of Allied Health Care

The following search commands were used (in Dutch and English in the DOCU database and in English in MEDLINE):

I.**Search in DOCU database**

- 1: klinim\$ or clinim\$
- 2: assessm\$ and method\$
- 3: (C0100 or C0110) and meetinstrumenten
- 4: 1 or 2 or 3

Catch: 197 documents

Retrieved: 26 documents

II.**Search in MEDLINE**

CLINIMETR* and review in PT (Publication Type)/method*

Catch: 11 documents**Retrieved: 7 documents****III****Specific search for literature concerning measurement instruments for rheumatic disorders****Measurement instruments for the assessment of impairments**

According to the pre-determined inclusion criteria (relevant impairments in body function and structure, relevant disabilities in activities, relevant rheumatic disorders), a literature search was performed to identify measurement instruments for the assessment of patients with the following rheumatic disorders: rheumatoid arthritis, seronegative polyarthritis (including psoriatic arthritis), osteoarthritis, ankylosing spondylitis, polymyositis or fibromyalgia.

Specific search in DOCU database

- 1 : RA
- 2 : meetinstrumenten or assessm\$ or klinim\$ or clinim\$ or vragenlijst\$ or meetschalen
- 3 : 1 and 2
- 4 : bewust\$ and 3
- 5 : (slaap\$ or slapen\$) and 3
- 6 : (denk\$ or think\$) and 3
- 7 : (drift\$ or wilsleven\$) and 3
- 8 : intelligen\$ and 3
- 9 : (aandacht\$ or attentie\$) and 3
- 10: geheugen\$ and 3
- 11: (waarneming\$ or realiteit\$) and 3
- 12: lichaamsschema and 3
- 13: (gevoel\$ or stemming\$ or emotie\$) and 3
- 14: RA or rheuma\$ or reuma\$ or bechterew\$ or JCA or fibromyal\$ or \$artros\$ or arthralg\$ or jicht or polymyalg\$ or sle
- 15: meetinstrum\$ or assessm\$ or klinim\$ or clinim\$ or vragenlijst\$ or meetscha\$
- 16: beweeglijkheid\$ or \$mobiliteit\$ or bewegingsuitslag
- 17: 14 and 15 and 16
- 18: RA or rheuma\$ or reuma\$ or bechterew\$ or JCA or fibromyal\$ or \$artros\$ or arthralg\$ or jicht or polymyalg\$ or sle
- 19: meetinstrum\$ or assessm\$ or klinim\$ or clinim\$ or vragenlijst\$ or meetscha\$
- 20: psychomotor\$ or (motoriek\$ and psychol\$)
- 21: 18 and 19 and 20
- 22: propriocep\$ or mechanorecep\$
- 23: 18 and 19 and 21
- 24: \$tonus\$ or dystoni\$
- 25: 18 and 19 and 24
- 26: (motor adj control) or bewegingssturing
- 27: 18 and 19 and 26
- 28: spierkracht\$ or (spieren and kracht) or (muscle adj strength) or contractiekracht\$

- 29: 18 and 19 and 28
 30: huid\$ or dermatol\$
 31: 18 and 19 and 30
 32: pijn\$
 33: (18 and 19 and 32) or (pijnmeetinstru\$ and 19)
 34: RA or rheuma\$ or reuma\$ or bechterew\$ or JCA or fibromyal\$ or \$artros\$ or arthralg\$ or jicht or polymyalg\$ or sle
 35: meetinstrum\$ or assessm\$ or klinim\$ or clinim\$ or vragenlijst\$ or meetscha\$
 36: 34 and 35
 37: gaan or gangstoornis\$ or gangpatroon or looppatroon or lopen or loopstoornis\$
 38: 36 and 37
 39: communica\$ or spraak\$
 40: 35 and 39
 41: staan or houding\$ or (motorische adj stoornis\$) or motoriek or grijpen or reiken or schrij\$ or vasthouden or knijpen or buigen or buksten or dragen or hangen or hurken or klimmen or knielen or neerzetten or opstaan or tillen or werpen
 42: 36 and 41
 43: gaan or lopen or traplopen or springen or dansen or fietsen or hardlopen or zwemmen or schaatsen or verplaatsen
 44: 36 and 43
 45: ADL or IADL or (functionele adj toestand) or functiestoornis\$
 46: 36 and 45
 47: (persoonlijke adj hygiene) or kleden or verzorgen or wassen or toilethyg\$ or eten or drinken
 48: 36 and 47
 49: huishoud\$ or woonomgeving or wonen or maaltijd\$ or keuken\$
 50: 36 and 49
 51: interaction\$ or relatie\$ or sociale adj factoren or psychosociale adj factoren
 52: 36 and 51
 53: arbeid\$ or vrijetijd\$ or recreati\$
 54: 36 and 53
 55: \$hulpmiddel\$ or orthesen or prothe\$
 56: 36 and 55

Catch: 280 documents

Retrieved: 159 documents

IV

Additional search in DOCU databse for possibly missed documents

- 1 RA or r?euma\$ or \$artros\$ or fibromyal\$ or bechterew or spondylitis adj ankylo\$ or polyart?ri\$ or polymyos\$ or polymyalg\$
 2 meetinstrum\$ or assessem\$ or klinim\$ or clini\$ or vragenlijst or meetscha\$ or index
 3 1 and 2

Catch: 420 documents, 169 of which were already included, so the additional catch was 251 documents.

Retrieved: 20 new documents

V**Additional search in MEDLINE**

The MEDLINE search was performed with the thesaurus.

Search strategy:

- 1 : spondylitis-ankylosing / all sub-headings
- 2 : explode arthritis / all sub-headings
- 3 : fibromyalgia / all sub-headings
- 4 : arthralgia / all sub-headings
- 5 : polymyalgia rheumatica / all sub-headings
- 6 : explode rheumatic diseases / all sub-headings
- 7 : 1 or 2 or 3 or 4 or 5 or 6
- 8 : explode outcome assessment (health care) / all sub-headings
- 9 : reproducibility of results / all sub-headings
- 10: sensitivity and specificity / all sub-headings
- 11: 8 or 9 or 10
- 12: 7 and 11

Catch: 765 documents

Retrieved: 216 new documents

VI**Search in DOCU database on names of measurement instruments**

Search strategy:

- 1 articular adj index
- 2 fuchs and articular:
- 3 epm-rom\$
- 4 dutch-aims
- 5 doyle and articular
- 6 dougados and functional adj index
- 7 disease adj activity and score
- 8 daltroy and spondyla\$
- 9 convery and polyartic\$
- 10 barthel\$
- 11 eysenck and persona\$
- 12 fibromyal\$ and impact
- 13 edss
- 14 eular
- 15 edi-320
- 16 dolorimeter
- 17 disabili\$ and distress
- 18 cbsq or (curtin and back)
- 19 beck and depressi
- 20 gross adj motor
- 21 self-efficacy and scale
- 22 aims
- 23 shygmoma\$
- 24 fim
- 25 acr
- 26 health adj assessment adj question\$
- 27 roland and disability

28 stai or (trait and anxiety)
 29 sequent\$ and dexterity
 30 SF-36
 31 quebec and disability adj scale
 32 scale and quality and life
 33 pulses and profile
 34 problem and elicitation
 35 faces and pain
 36 oswestry
 37 observation and pain and method
 38 pain and numerical and rating
 39 martin and vigorimeter
 40 gars
 41 walking adj time
 42 wire and tracing
 43 waddell and meetinstrumen\$
 44 katz adj index
 45 haq
 46 hospital adj anxiety
 47 jamar
 48 karnofsky
 49 jebsen and test
 50 sip or sickness adj impact
 51 self-report and questionna\$
 52 pain and drawing
 53 aneroid
 54 helplessness adj index
 55 borstkas
 56 bradburn
 57 denver and develop\$
 58 disabil\$ and interview
 59 impact adj profile and disabil\$
 60 (donaldson and adl) or adl adj evaluation adj form
 61 dunham and spondy\$
 62 dutch adj functional
 63 dutch-aims
 64 egger and articular
 65 euro adj quality
 66 daily and activity and question\$
 67 functional adj assessment and inventory
 68 functional adj assessment and profile
 69 fcq or functional adj capacity adj questio\$
 70 functional adj index and ra
 71 functional adj life and scale
 72 functional adj limitations adj profile
 73 fsi or functional adj status adj index
 74 fsrs or functional adj status adj rating adj system
 75 general adj health adj status adj index
 76 health adj insurance adj exper\$ or hie
 77 health adj utilit\$
 78 living adj behavior

- 79 instrumen\$ adj adl-scale
- 80 preliminary and diagnostic
- 81 rapid and disabili\$
- 82 rap or (rehabil\$ and activit\$ and profile\$)
- 83 rehabilita\$ and indicators
- 84 rolian
- 85 rotterdam and symptom\$
- 86 symptom and checklist
- 87 scott and index
- 88 shober\$
- 89 simulated and adl
- 90 spitzer and quality
- 91 stoke and index
- 92 time and care
- 93 toronto and (functional or question\$)
- 94 torsion and dynamo\$
- 95 vinger-bodem\$
- 96 keitel adj functional or kft
- 97 kenny and self
- 98 (klein and bell) or (klein and adl)
- 99 knee and pain and questio\$
- 100 lambeth or lansburry
- 101 lawton
- 102 lee and index
- 103 lequesne
- 104 levels and rehabil\$ and scale
- 105 locomotion and score
- 106 longitu\$ and functio\$ and assessm\$
- 107 mactar
- 108 long-term and disabili\$ and questio\$
- 109 opaq or (osteop\$ and assessment and question\$)
- 110 osteop\$ and functional and disabil\$
- 111 osteopo\$ and question\$ and quality
- 112 paulus and criteria
- 113 pfeffer and activity
- 114 self and maintain\$
- 115 pinch and gauges
- 116 poly and articu\$ and index
- 117 vitaliona
- 118 who and performan\$
- 119 performance and status and scale

Cath: 398 new documents

Retrieved: 58 new documents

VII

Search in MEDLINE on names of measurement instruments

In particular measurement instruments with lacking information.

Search strategy:

- 1 affect balance scale
- 2 curtin adj back adj screening

3	denver developmental screening test
4	disability impairment interview
5	expanded disability status scale
6	evaluation of daily activity question*
7	(eysenck personality question* or EPQ) and rheuma*
8	health insurance experiment RAND
9	health utilities index
10	instrumental adl scale or iadl scale
11	karnofsky performance status scale
12	bradburn scale of well being
13	daltroy functional status or functional status spondylarth*
14	disability distress scale
15	donaldson adl evaluation
16	functional assessment profile
17	functional index ra
18	functional status rating system
19	general health status index
20	groningen activiteiten
21	hoskins squires test
22	hospitality anxiety depression scale and rheuma*
23	independent living behavior checklist
24	jebsen test
25	kenny self care evaluation
26	klein bell adl scale
27	lambeth disability screening
28	lawton instrumental
29	levels of rehabilitation scale
30	longitudinal functional assessment system
31	martin vigorimeter
32	oecd long-term disability question*
33	osteoporosis assessment question*
34	osteoporosis quality of life questionnaire or oqlq
35	oswestry and pain and spondylitis ankylopo*
36	pain faces scale not child*
37	pfeffer functional activities
38	preliminary diagnostic question*
39	physical self (maintenance) scale
40	pulses profile
41	quality of life scale and rheuma*
42	quebec back pain disability scale
43	rapid disability rating scale
44	rehabilitation indicators
45	roland disability questionnaire
46	rotterdam symptom checklist
47	scott index
48	vigorimeter
49	vitaliano ways of coping checklist
50	stai trait or state trait
51	time care profile
52	spitzer quality of life index
53	toronto functional capacity questionnaire

54 waddell disability index

Catch: 423 documents

Retrieved: 3 new documents

VII

Search update MEDLINE 2000

SilverPlatterASCII 3.0WINNMEDLINE (R) 2000

"Spondylitis-Ankylosing"/ all sub-headings

explode "Arthritis"/ all sub-headings

explode "Rheumatic-Diseases"/ all sub-headings

rheumatoid

arthritis

rheumatoid arthritis

bechterew

fibromyalg*

"rheumatic"

"disorders"

"rheumatic"

"disease*"

"rheumatic disorders" or "rheumatic disease*"

"ankylosing"

"spondylitis"

"ankylosing spondylitis"

#1 or #2 or #3 or #6 or #7 or #8 or #13 or #16

measure*

questionnai*

testing

assessment*

index

"rating"

"scale"

scale*

score

scoring

measure* or questionnai* or testing or assessment* or index or "rating scale" or

scale* or score or scoring

"reproducibility-of-results"

explode "sensitivity-and-specificity"/ all sub-headings

valid*

useful*

reliabil*

interrater

intrarater

specifici*

sensitiv*

valid* or useful* or reliabil* or interrater or intrarater or specifici* or sensitiv*

interobserver

intraobserver

interobserver or intraobserver

#38 or #41

#28 and #17
 #29 or #30 or #42
 #43 and #44

Search update DOCU database May 2001

1: Clinim\$ OR klinim\$ OR meetscha\$ OR assessmen\$ OR vragenlijst\$ OR meetinstrument\$	9475 docs
2: Arthralg\$ OR polymyalg\$ OR jicht OR sle OR coxarthro\$ OR gonarthro\$ OR jca OR Bechterew OR arthrose OR fibromyal\$ OR reuma\$ OR RA OR rheuma\$	3655 docs
3: JCA	155 docs
4: 2 NOT 3	3500 docs
5: 1 AND 4	546 docs
6: "2001" OR "2000" OR "1999" OR "1998"	16518 docs
7: 5 AND 6	46 docs
8: intrabeoord\$ OR linterbeoord\$ OR responsiviteit OR specificiteit OR psychometrie OR sensitiviteit OR validiteit\$ OR validi\$ OR betrouwbaar\$	3890 docs
9: 7 AND 8	15 docs

Catch update: 785 documents
Retrieved: 82 documents

Appendix 2

Criteria for methodological cut-off points

Appendix 2: Criteria for methodological cut-off points

Reproducibility

Intra-interobserver reliability

According to the literature, reliability coefficients for intra-observer reliability need to be higher than reliability coefficients for inter-observer reliability (1), and the required minimum reliability differs between scientific research and research in clinical settings. For correct interpretation of the cut-off points for 'good' reliability, it must be emphasized that these depend on the sample size, since measurement of a sample of 1,000 patients can tolerate a much less reliable instrument than a measurement of sample of 10 patients. Therefore, for individual patients the criteria should be more strict than for group measurements. Deyo and Inui proposed a minimal value of 0.90 as criterion for intra-observer reliability in individual patients, but for large groups the minimal value could be lower (2). However, Meenan et al. considered a cut-off point of ≥ 0.80 to be acceptable for individual measurements (3).

Landis and Koch suggest the following application of reliability coefficients as a rule of thumb (4):

0.00 – 0.20	= slight
0.21 – 0.40	= fair
0.41 – 0.60	= moderate
0.61 – 0.80	= substantial
0.81 – 1.00	= almost perfect

They made no differentiation between cut-off points for inter-observer reliability and intra-observer reliability.

Martin et al. used 0.70 as a criterion for 'adequate reliability' for questionnaires and group comparisons (5).

Based on the Landis & Koch classification (4), Eliasziw et al. (6) claimed that the inter-observer reliability must at least be 'substantial' ($r=0.61-0.80$), or 'almost perfect' ($r=0.81-1.00$) for intra-observer reliability.

Finally, the value reported in a study can depend on the parameter of correlation that is used for the assessment of reliability. The shortcoming of Pearson's r as a reliability index is that it does not take systematic observer bias into account (7). Haas recommended to use of the weighted Kappa as optimal reliability index for ordinal data, and the Intra Class Correlation Coefficient (ICC) (one-way ANOVA) as the most appropriate statistic for continuous data (7).

Using the Kappa index as a measure of agreement, a Kappa of 0.40 or lower is defined as 'poor correlation', and a Kappa of ≥ 0.75 indicates 'excellent correlation' (8).

Based on the above-mentioned sources we decided to use the following cut-off points for Pearson's r , Spearman's ρ or the ICC:

Intra-observer reliability:	good	$0.85 \leq 1.00$
	moderate	$0.65 \leq \alpha \leq 0.85$
	poor	$\alpha < 0.65$

Inter-observer reliability:	good	$0.80 \leq 1.00$
	moderate	$0.60 \leq \alpha \leq 0.80$
	poor	$\alpha < 0.60$

Construct validity

The literature provides very little data concerning construct validity: hardly any clearly defined cut-off points for acceptable validity, and in the available data the margins are wide and the cut-off points are relatively low.

Drenth (9) refers to an extensive study carried out by Ghiselle, concluding that the average correlations lie between 0.30 and 0.45 (10). This corresponds with the criteria applied by Huiskes, who defines a correlation of 0.40 as 'reasonably high' (11). Deyo (2), and more recently also Streiner (12), state that perfect correlation is clearly not desirable, because this would imply that the new instrument would only provide redundant information. Physical scales would show expected correlations with accepted clinical measurement instruments whenever these are available. However, the validation of social functioning scales is more problematic, and should probably not focus on physician assessments.

Martin et al. (5) apply the same criteria as Huiskes, based on the standard proposed by Stewart and Ware (13), in which $\rho > 0.40$ ($p < 0.01$) is defined as cut-off point for good validity.

Doeglas et al. (14) refer to De Jong & Molenaar (15), and take as a rule of thumb correlation criteria for construct validity in which $0.30 \leq \alpha < 0.40$ is 'weak', $0.40 \leq \alpha < 0.50$ is 'moderately strong' and $\alpha \geq 0.50$ is 'strong'.

Based on the above-mentioned sources we decided to use the following cut-off points for construct validity:

good	$0.65 \leq \alpha$
moderate	$0.50 \leq \alpha \leq 0.65$
poor	$\alpha < 0.50$

where α is the index of association.

In the pooling of data we separated ICC values from Pearson's r and Spearman's ρ .

There were three reasons for our choice of a relatively high level of cut-off points:

1. We consider it to be of great clinical importance to use reproducible and valid measurement instruments for the measurement of individual patients in clinical practice.
2. We preferred to be on the safe side, leaving little room for discussion.
3. We aimed to separate the wheat from the chaff, based on the philosophy that it is better to end up with a Top 3 of good instruments than a Top 20 of disputable instruments.

Reference List

- 1 Eliasziw M, Young SL, Woodbury MG, Fryday-Field K: Statistical methodology for the concurrent assessment of interrater and intrarater reliability: using goniometric measurements as an example. *Phys Ther* 1994;74(8):777-88.
- 2 Deyo RA, Inui TS: Toward clinical applications of health status measures: sensitivity of scales to clinically important changes. *Health Services Research* 1984;19(3):275-89.
- 3 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):1-10.
- 4 Landis JR, Koch GG: The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159-74.
- 5 Martin DP, Engelberg R, Agel J, Swiontkowski MF: Comparison of the Musculoskeletal Function Assessment questionnaire with the Short Form-36, the Western Ontario and McMaster Universities Osteoarthritis Index, and the Sickness Impact Profile health-status measures. *J Bone Joint Surg Am* 1997;79(9):1323-35.
- 6 Eliasziw M, Young SL, Woodbury MG, Fryday-Field K: Statistical Methodology for the Concurrent Assessment of Interrater and Intrarater Reliability: Using Goniometric Measurements as an Example. *Phys Ther* 1994;74(8):777-88.
- 7 Haas M: Statistical methodology for reliability studies. *J Manip Physiol Therap* 1991;14(2):119-32.
- 8 Bouter LM, Dongen MCJM van: *Epidemiological Research: Design and Interpretation.* (in Dutch: *Epidemiologisch onderzoek. Opzet en interpretatie*). 5th ed., Houten/Antwerpen, Bohn Stafleu Van Loghum, 2005:1-359.
- 9 Drenth P: *Introduction to the test theory* (in Dutch: *Inleiding in de testtheorie*). Deventer, Van Loghum Slaterus, 1985.
- 10 Ghiselle EE: *The measurement of occupational aptitude.* 1st ed., Los Angeles, Berkely, 1966.
- 11 Huiskes CJAE, Kraaimaat FW, Bijlsma JWJ: Development of a self-report questionnaire to assess the impact of rheumatic diseases on health and lifestyle. *J Rehabil Sciences* 1990;3(2):65-70.
- 12 Streiner DL, Norman GR: *Health Status Measurement Scales. A practical guide to their development and use.* 2nd ed., Oxford, Oxford University Press, 2003:1-231.
- 13 Stewart AL, Ware JE. *Measuring functioning and well-being: the Medical Outcomes Study Approach.* 1992. Durham, North Carolina, Duke University Press.
- 14 Doeglas D, Krol B, et al.: The assessment of functional status in rheumatoid arthritis: a cross cultural longitudinal comparison of the Health Assessment Questionnaire and the Groningen Activity Restriction Scale. *J Rheumatol* 1995;22(10):1834-43.
- 15 Jong A de, Molenaar IW: An application of Mokken's model for stochastic cumulative scaling in psychiatric research. *J Psychiatr Res* 1987;21:137-49.

Appendix 3

List of abbreviations of measurement instruments and sub-scales

Appendix 3: List of abbreviations of measurement instruments and sub-scales

ABBREVIATION	NAME OF MEASUREMENT INSTRUMENT:	ABBREVIATIONS OF SUB-SCALES:
AHI	Arthritis Helplessness Index	Anx anxiety Depri depression
AI	Articular Index	Pain pain
AIM2D	Arthritis Impact Measurement Scale2 Dutch	Anx anxiety Depri depression EmoF emotional function MenH mental health Mob mobility Pain pain PhysA Physical Activities WaBe Walking and Bending
AIMS	Arthritis Impact Measurement Scale	ADL Activities of Daily Living Anx anxiety Depri depression Disab disabilities EmoF emotional function GlobH Global Health MenH mental health Mob mobility Pain pain PhysA Physical Activities PhysH Physical Health PhysM Physical Mobility
AIMS2	Arthritis Impact Measurement Scale 2	Anx anxiety Depri depression EmoF emotional function MenH mental health Mob mobility Pain pain Selfcare Self-care
AIMSD	Arthritis Impact Measurement Scale - Dutch	WaBe Walking and Bending ADL Activities of Daily Living Anx anxiety Depri depression EmoF emotional function MenH mental health Mob mobility Pain pain PhysA Physical Activities
AIMSS	AIMS short version	ADL Activities of Daily Living Anx anxiety Depri depression EmoF emotional function MenH mental health Mob mobility Pain pain PhysA Physical Activities
ASES	Arthritis Self Efficacy Scale	FSE Functional Self Efficacy
AUSCAN-OHI	Australian/Canadian Osteoarthritis Hand Index	Pain pain Stiff stiffness PhysF physical function
BASDAI	Bath Ankylosing Spondylitis Disease Activity Index	
BASMI	Bath Ankylosing Spondylitis Metrology Index	Stiff-D Stiffness duration Stiff-S Stiffness severity
BDI	Beck Depression Inventory	
BIM	Barthel Index Modified	
Chest	Chest Expansion	
DFI	Functional Index (Dutch)	
Dol	Dolorimeter	
EDI	Electric Digital Inclinator-320	
EI	Enthesis Index	
EPM	Escola Paulista de Medicina Range of Motion Scale	
E-QoL	Euro Quality of Life Scale	
FAS	Functional Assessment Survey	Aids Aid-tools Trans Transport
FFD	Finger Floor Distance	

FFI	Functional Foot Index		
FIQ	Functional Index Questionnaire	PhysF	Physical Function
FSAI	Functional Status Assessment Instrument		
FSI	Functional Status Index	GlobH	Global Health
		Mob	Mobility
GARS	Groninger Activity Restriction Scale	ADL	Activities of Daily Living
Gonio	Goniometer		
Gripp	Grippit		
HAQ	Health Assessment Questionnaire	Disab	Disabilities
		DressG	Dressing & grooming
		Eat	Eating
		GlobH	Global Health
		Hygie	Hygiene
		Mob	Mobility
		ODI	Overall Disability Index
		OthAc	Other Activities
		Walk	Walking
IRGL	Invloed van Reuma op Gezondheid en Leefwijze	Mob	Mobility
IWB	Index of Well Being	Selfcare	Self-care
		GlobH	Global Health
		Mob	Mobility
J-MAP	Joint-Specific Multidimensional Assessment of Pain		
KFT	Keitel Functional Test		
LDQ	Leeds Disability Questionnaire		
MDR	MDR-Index		
MFAQ	Musculoskeletal Function Assessment Questionnaire		
MHAQ	Modified Health Assessment Questionnaire	DressG	Dressing & grooming
		Eat	Eating
		Hygie	Hygiene
		ODI	Overall Disability Index
		OthAc	Other Activities
		Walk	Walking
MKI	Modified Kapandji Index		
MobSpine	Mobility assessment spine in ankylosing spondylitis.	CCD	Chin to Chest Distance
		ChExp	Chest Expansion
		OWD	Occiput to Wall Distance
		VitCp	Vital Capacity
MPPDQ	MACTAR Patient Preference Disability Questionnaire		
MPQ	McGill Pain Questionnaire		
MS-D	Morning Stiffness, duration	Stiff-D	Stiffness duration
		Stiff-S	Stiffness severity
MS-S	Morning Stiffness, severity		
MUMQ	Maastricht Utility Measurement Questionnaire		
NHP	Nothingham Health Profile	PhysM	Physical Mobility
OSRA	Overall Status in Rheumatoid Arthritis		
P-NRS	Pain Numeric Rating Scale		
RAPS	Rheumatoid Arthritis Pain Scale		
RAQoL	Rheumatoid Arthritis Quality of Life Scale		
RPS	Regional Pain Scale		
SAJ	Self Assessment Joint Count		
SF-36	Short Form 36	PhysA	Physical Activities
Shob	Shober Test		
SIP	Sickness Impact Profile	BC&M	Body Care & Movement
		GlobH	Global Health
		Mob	Mobility
		PhysH	Physical Health
SODA	Sequential Occupational Dexterity Assessment		
Sphy	Sphygmomanometer (handheld dynamometer)		
Spond	Spondylometer (Dunham)		
SRQ	Self-Report Questionnaire		
SSAI	Spielberger State-Anxiety Inventory		
STAI	Spielberger Trait-Anxiety Inventory		
Stest	Stiffness test		
SW	Stair Walk		
VAS-P	Visual Analogue Scale Pain		
VAS-S	Visual Analogue Scale Stiffness		
WOMAC	WOMAC-Osteoarthritis Index	Disab	Disabilities
WT	Walking time-50 foot		

Summary

Measurement instruments for patients with rheumatic disorders: a clinimetric appraisal.**Summary**

People with rheumatic disorders form a substantial percentage of the population with chronic musculoskeletal disorders, leading to substantial impairments in body functions and structures and many limitations in the activities of daily living. For this reason, in recent decades many measurement instruments have been developed for the assessment of patients with rheumatic disorders. This thesis gives an overview of the available instruments and assesses the clinimetric properties of these instruments.

Chapter 1 is the General Introduction, which reports on the prevalence and clinical features of rheumatic disorders and the classification of rheumatic disorders according to the American College of Rheumatology (ACR) and the European League Against Rheumatism (EULAR). It also describes the importance of the multidimensional impact of the majority of rheumatic disorders for quality of life, and why this approach results in a multidimensional description of the health status of the patient with a rheumatic disorder. It is hypothesized that in the validation of instruments which measure the consequences of rheumatic disorders, validating against a dissimilar (less similar) construct will result in lower correlations than validating against a similar (optimal comparable) construct. This hypothesis was the reason why we decided to differentiate all available data regarding the construct validity of measurement instruments into validity based on correlation with more homogeneous constructs, and validity based on correlation with more heterogeneous constructs.

The framework underlying this thesis is the International Classification of Impairments, Disabilities and Handicaps (ICIDH). *Chapter 2* gives an overview of the development and content of the ICIDH and, in particular the revision of the beta draft 2-ICIDH, which resulted in the International Classification of Functioning, Disability and Health (ICF). In a systematic review of the literature the properties of all available measurement instruments for the assessment of patients with rheumatic disorders are classified as 'diagnostic' (60%), 'prognostic' (4%) or 'evaluative' (36%). A short general description and definition of the ICF domains is followed by a more detailed description of the ICF domains in measurement instruments and questionnaires that apply to patients with rheumatic disorders, most of which are intended for the

assessment of impairments in functions (57%). A further 37% measure disabilities and limitations in activities, and 6% focus on problems in participation. Finally, five steps are suggested for the selection of the most suitable instruments for clinical use, based on the most important factors in the selection of optimal instruments: the treatment goals (based on ICF domains), and methodological quality and feasibility of the instruments.

Chapter 3 focuses on the construct validity of instruments that measure impairments in body structures and function in patients with rheumatic disorders. An inventory was made of all constructs against which such instruments were validated. The ICF classification is used as basis for the classification of all constructs that have been used in validity studies. The constructs are divided into six levels, varying from the most similar construct to the least similar construct. Forty-two measurement instruments for the assessment of impairments in body structures and functions were identified and, in total, these have been validated 600 times. Sixteen percent of the instruments that measure impairments have been validated against sub-scales or instruments that measure the most similar construct. A minority (20%) of the instruments have been validated against more similar constructs, and 80% have been validated against more or less dissimilar constructs. Furthermore, we investigated whether validation against the same or a similar construct results in higher correlation values than validation against a less similar construct. The available data show that in 50% approximately of the studies the values for construct validity are much lower if validated against a less similar construct, compared to validation against the most similar construct or the same impairment combined with (an)other impairment(s) in body structures and function (Table 3.4). It is discussed whether different cut-off points should be used for the qualification of correlation coefficients, depending on the degree of similarity, to validate an instrument.

Chapter 4 presents the results of a systematic review of the literature on all available measurement instruments (49) for the assessment of impairments in body structures and function in patients with rheumatic disorders. Furthermore, we investigated the methodological quality of these instruments with regard to reliability and validity. The results are stratified up for the six categories of impairments that are most frequently assessed with measurement instruments (mental functions, stiffness, pain, mobility, muscle force and swelling) in patients with rheumatic disorders. Taking into account the predefined cut-off points for reliability and validity, 11 of the 49 instruments met the criteria for reliability as well as validity. Finally, this chapter describes the consequences of validation against an optimal comparable construct versus

validation against an incomparable construct. Summarizing the data on measurement instruments to assess the relevant impairments, data on the construct validity of 40 instruments and/or sub-scales are available. In 14 out of 21 instruments the correlation values for validation against the optimally comparable constructs proved to be higher than the values for validation against the incomparable constructs. Chapter 4 therefore gives advice with regard to which instruments are the most appropriate for use in a clinical setting, based on their clinimetric properties. These are: for impairments in mental functions the AIMS⁷ or (more specifically) the STAI; for stiffness the BASDAI for patients with ankylosing spondylitis and the VAS-S for patients with other rheumatic disorders; for pain the AIMS, the AUSCAN-OHI and the RAPS; joint mobility the Gonio, Chest and Shob; for muscle force the Sphy; and, finally, for joint swelling the AI.

Chapter 5 gives a systematic review of the literature regarding all available measurement instruments for the assessment of gait and gait-related activities in patients with rheumatic disorders and the clinimetric properties of these instruments with regard to reliability and construct validity. A total of 78 instruments, including 36 sub-scales were identified. The intra-observer reliability of 28 instruments or sub-scales was investigated. 16 of these 28 met the criterion of $r/ICC^8 \geq 0.85$, and 37 were found to have good validity. The construct validity of 32 instruments and/or subscales was investigated in studies in which they were validated against the most similar constructs. Seventeen of the 32 met the predefined criterion of $r/ICC \geq 0.65$. If both reliability and validity are required to be 'good', 7 instruments meet the criteria. In our data-analysis we also investigated the assumption that evaluation of convergent construct validity (validation against an optimal comparable construct) results in stronger correlations when validated against a dissimilar construct. Data on the validity of 49 instruments and/or sub-scales are available. For 18 of these 49 instruments, data are available on the optimally comparable construct validity as well as the imperfect construct validity. In 7 of those 18 instruments the correlation for the similar constructs proved to be stronger than the values for validation against the dissimilar constructs. For one instrument the validity was the same for the optimally comparable constructs as for validation against the incomparable constructs. In 10 of the 18 instruments the correlation values for the optimally comparable constructs proved to be lower than the values for the incomparable constructs. Based on the available information, it is concluded that the Rheumatoid Arthritis Quality of Life scale (RAQoL) and the Health Assessment Questionnaire (HAQ) seem to be the best

⁷ For an explanation of all abbreviations of the instruments and sub-scales, see the Appendix 3.

⁸ r = Pearson's product-moment coefficient of correlation or Spearman's rank correlation coefficient.
ICC = Intraclass Correlation Coefficient

available instruments for assessment of disabilities in gait and gait-related activities. However, both are multi-dimensional and the influence of certain sub-scales is not clear. These instruments also provide ballast information, that is not related to gait or gait-related activities. It is therefore concluded that more appropriate instruments for clinical use could be the Walking Time (WT) and the Stair Walking (SW); however, there is lack of data concerning the validity of these two instruments.

Chapter 6 gives an overview of the results of a systematic review of the literature regarding all available measurement instruments for the assessment of disabilities in personal care in patients with rheumatic disorders, as well as the clinimetric properties of these instruments (reliability, construct validity and responsiveness). In total, 19 measurement instruments were included, five of which were found to have good reliability, and 12 had acceptable validity. Only three questionnaires met both criteria (AIMS, HAQ and SIP), but the results concerning the responsiveness of these three questionnaires were conflicting. Like in Chapter 4 and 5, we also investigated the assumption that the evaluation of convergent construct validity (validation against an optimal comparable construct) results in stronger correlations when validated against a dissimilar construct, in this chapter with regard to disabilities in personal care. No difference was found in the strength of correlation between validation against the most similar construct versus validation against the least similar construct. It is concluded that the Arthritis Impact Measurement Scale (AIMS) is the most suitable instrument for the assessment of disabilities in personal care. For the three instruments that were found to be both reliable and valid (AIMS, HAQ and SIP) conflicting outcomes are reported for responsiveness. On basis of the results in this Chapter, it is recommended that generic measures are supplemented with a targeted measure whenever possible.

In *Chapter 7* the issues which are important for the interpretation and extrapolation of the results are discussed. First, the Top-3 best available instruments for each domain of assessment, as described in this thesis, are summarized in one Table. Also discussed is the appropriateness of the ICF as a framework for analysis of the literature, as performed in this thesis. The relevance and (dis)advantages of the ICF are discussed in relation to outcome assessment and the method of data-analysis used in this thesis. Furthermore, this final chapter discusses the hypothesis that validation against the same or a similar construct results in higher correlation values than validation against a dissimilar construct, as well as the consequences for the interpretation of the results of our study. Also discussed are the implications of recent innovations in clinimetrics with regard to responsiveness. Finally, the results of our

study are discussed in relation to the results reported by the OMERACT group (OMERACT is an abbreviation of Outcome Measures in Rheumatoid Arthritis Clinical Trials). Chapter 7 concludes with the implications of this study for clinical practice and clinical research, and some recommendations for future research.

Samenvatting

Meetinstrumenten voor patiënten met reumatische aandoeningen: een klinimetrische beoordeling.

Samenvatting

Van alle chronische klachten over het bewegingsapparaat vormen reumatische aandoeningen een substantiële groep. Reumatische aandoeningen kenmerken zich door stoornissen in functies en anatomische eigenschappen van structuren, beperkingen in activiteiten van het dagelijks functioneren en participatieproblemen. Daarom zijn de afgelopen decennia vele meetinstrumenten ontwikkeld om deze stoornissen, beperkingen en participatieproblemen bij patiënten met reumatische aandoeningen in kaart te brengen. Dit proefschrift geeft een overzicht van de beschikbare meetinstrumenten en de klinimetrische eigenschappen ervan.

Hoofdstuk 1 is de algemene introductie van de dissertatie. Hierin wordt een overzicht gegeven van de prevalentie, de klinische verschijnselen en de classificaties van reumatische aandoeningen volgens de *American College of Rheumatology* (ACR) en de *European League Against Rheumatism* (EULAR). Tevens wordt het belang beschreven van de invloed van reumatische aandoeningen op de kwaliteit van leven. De hypothese wordt toegelicht dat, bij bepaling van de constructvaliditeit, validering tegen een niet-vergelijkbaar construct zal leiden tot een lagere correlatie dan validering tegen een optimaal vergelijkbaar construct. In het kader van deze hypothese is besloten een onderscheid te maken tussen validiteit welke is gebaseerd op correlaties met homogene (optimaal vergelijkbare) constructen enerzijds, en correlaties met heterogene (niet-vergelijkbare) constructen anderzijds.

Het uitgangspunt voor dit proefschrift is de International Classification of Impairments, Disabilities and Handicaps (beta draft-2 ICDH). *Hoofdstuk 2* geeft een overzicht van de ontwikkeling en de inhoud van de ICDH, en in het bijzonder de revisie van de beta-draft 2-ICDH, welke heeft geleid tot de International Classification of Functioning, Disability and Health (ICF). Op basis van een systematisch literatuuronderzoek zijn de eigenschappen van de beschikbare meetinstrumenten voor patiënten met reumatische aandoeningen geklasseerd als 'diagnostisch' (60%), als 'prognostisch' (4%) of als 'evaluatief' (36%). Na een beschrijving en definitie van de verschillende domeinen van de ICF, wordt aangegeven welke ICF-domeinen zijn vertegenwoordigd in de meetinstrumenten voor patiënten met reumatische aandoeningen. De meeste meetinstrumenten meten stoornissen in functies en anatomische eigenschappen van structuren (57%), 37%

meet beperkingen in activiteiten en 6% richt zich op problemen in participatie. Tot slot wordt een 5-stappen schema beschreven voor de selectie van de meest optimale meetinstrumenten in klinische situaties, gebaseerd op de belangrijkste bepalende factoren: de behandeldoelen (gebaseerd op de ICF-domeinen), de methodologische kwaliteit en de hanteerbaarheid van de instrumenten.

Hoofdstuk 3 gaat over de constructvaliditeit van instrumenten voor het meten van stoornissen in anatomische eigenschappen van structuren en functies bij patiënten met reumatische aandoeningen. Er is een inventarisatie gemaakt van alle constructen waartegen zulke instrumenten zijn gevalideerd. De constructen zijn verdeeld in zes niveaus, variërend van het meest vergelijkbare construct tot het minst vergelijkbare construct. Voor het meten van stoornissen in functies en anatomische eigenschappen van structuren zijn 42 instrumenten gevonden, welke gezamenlijk 600 keer zijn gevalideerd. Zestien procent van deze instrumenten is gevalideerd tegen subschalen of tegen instrumenten die een optimaal vergelijkbaar construct meten. Een minderheid van 20% is gevalideerd tegen optimaal vergelijkbare of min of meer vergelijkbare constructen, en 80% van de instrumenten is gevalideerd tegen niet-vergelijkbare constructen. Tevens is geïnventariseerd of validering tegen een vergelijkbaar construct ook leidt tot een sterkere correlatie dan validering tegen niet-vergelijkbare constructen. Uit de beschikbare data blijkt dat in ongeveer 50% van alle studies de correlatie zwakker is bij validering tegen een niet-vergelijkbaar construct in vergelijking tot validering tegen een optimaal vergelijkbaar construct (Tabel 3.4). Ter discussie komt in hoeverre verschillende afkappunten zouden moeten worden gebruikt voor de beoordeling van de gevonden correlaties, afhankelijk van de mate van vergelijkbaarheid van de gebruikte constructen.

Hoofdstuk 4 geeft een systematisch overzicht van de literatuur van de beschikbare meetinstrumenten (49) voor het meten van stoornissen in anatomische eigenschappen van structuren en functies bij patiënten met reumatische aandoeningen. Daarnaast is de methodologische kwaliteit van deze meetinstrumenten met betrekking tot betrouwbaarheid en validiteit in kaart gebracht. De resultaten zijn geordend naar de zes meest voorkomende stoornissen bij patiënten met reumatische aandoeningen (stoornissen in mentale functies, stijfheid, pijn, mobiliteit, spierkracht en zwelling). Uitgaande van vooraf bepaalde afkappunten voor de betrouwbaarheid en de validiteit, zijn 11 van de 49 instrumenten betrouwbaar en valide. Tot slot worden in dit hoofdstuk de gevolgen beschreven van de validering tegen een optimaal vergelijkbaar construct versus validering tegen een niet-vergelijkbaar construct. De data over meetinstrumenten voor het meten van de

relevante anatomische eigenschappen van structuren en functies samenvattend, blijken gegevens over de constructvaliditeit beschikbaar van 40 instrumenten of subschalen. Bij 14 van de 21 meetinstrumenten blijken de correlatiecoëfficiënten bij validering tegen een optimaal vergelijkbaar construct hoger te zijn dan bij validering tegen een niet-vergelijkbaar construct. Op basis van de methodologische kwaliteit wordt in hoofdstuk 4 aangegeven welke meetinstrumenten het meest optimaal zijn voor klinisch gebruik. Dat zijn: voor stoornissen in mentale functies de AIMS⁹, of (meer specifiek) de STAI; voor stijfheid de BASDAI voor patiënten met spondylitis ankylopoëtica en de VAS-S voor patiënten met andere reumatische aandoeningen; voor pijn de AIMS, de AUSCAN-OHI en de RAPS; voor gewrichtsmobiliteit de Gonio en Shob; voor spierkracht de Sphy; en, tot slot, voor gewrichtszwelling de AI.

Hoofdstuk 5 beschrijft een systematisch literatuuroverzicht van de instrumenten voor het meten van (beperkingen in) het gaan en daaraan gerelateerde activiteiten bij patiënten met reumatische aandoeningen en de klinimetrische eigenschappen van deze instrumenten ten aanzien van betrouwbaarheid en validiteit. Totaal zijn 78 meetinstrumenten gevonden, met gezamenlijk 36 verschillende subschalen. Van 28 meetinstrumenten is de intra-beoordelaarsbetrouwbaarheid onderzocht. 16 van de 28 meetinstrumenten overstegen het criterium van $r/ICC^{10} \geq 0.85$, en voor 37 bleek ook de validiteit voldoende. Voor 32 meetinstrumenten/subschalen is de validiteit bepaald door validering tegen een optimaal vergelijkbaar construct. Zeventien van deze 32 meetinstrumenten overstegen het vooraf vastgestelde criterium van $r/ICC \geq 0.65$. Voor zeven meetinstrumenten blijkt zowel de betrouwbaarheid als ook de validiteit 'goed' te zijn. In de data-analyse is tevens de veronderstelling onderzocht dat validering tegen een vergelijkbaar construct resulteert in sterkere correlaties dan validering tegen een niet-vergelijkbaar construct. Voor 49 meetinstrumenten of subschalen zijn gegevens over de validiteit beschikbaar. Voor 18 van de 49 meetinstrumenten zijn gegevens beschikbaar over zowel validering tegen een vergelijkbaar construct als ook over validering tegen een niet-vergelijkbaar construct. Voor zeven van de 18 meetinstrumenten bleek dat validering tegen een vergelijkbaar construct resulteert in een sterkere correlatie in vergelijking met de validering tegen een niet-vergelijkbaar construct. Voor één meetinstrument bleek geen verschil tussen validering tegen een vergelijkbaar construct en validering tegen een niet-vergelijkbaar construct, en voor tien van de 18 meetinstrumenten bleken de correlatiecoëfficiënten bij validering tegen een optimaal vergelijkbaar construct lager dan bij validering tegen een niet-vergelijkbaar construct. Op basis van de

⁹ Voor een verklaring van alle afkortingen van de meetinstrumenten en subschalen zie Appendix 3.

¹⁰ r = Pearson's product-moment correlatie coëfficiënt of Spearman's rank correlation coefficient. ICC = Intraclass Correlation Coëfficiënt

beschikbare informatie werd geconcludeerd dat de RAQoL and the HAQ de best beschikbare instrumenten zijn voor het meten van beperkingen in het gaan en gaan-gerelateerde activiteiten. Dit zijn beide multidimensionale meetinstrumenten, en de invloed van bepaalde subschalen op de validiteit van deze meetinstrumenten is niet duidelijk. Bovendien geven deze instrumenten extra informatie over andere activiteiten dan het gaan. Daarom wordt geconcludeerd dat voor klinisch gebruik de WT en de SW wellicht meer bruikbaar zijn, hoewel geen gegevens beschikbaar zijn over de validiteit van deze twee meetinstrumenten.

Hoofdstuk 6 beschrijft de resultaten van een systematisch literatuuronderzoek naar instrumenten voor het meten van beperkingen in activiteiten in de persoonlijke verzorging bij patiënten met reumatische aandoeningen; alsmede de klinimetrische eigenschappen van deze instrumenten (betrouwbaarheid, constructvaliditeit en responsiviteit). Totaal zijn 19 meetinstrumenten gevonden, waarbij voor vijf een goede betrouwbaarheid bleek te zijn aangetoond en voor 12 een acceptabele validiteit. Drie van de 19 bleken zowel betrouwbaar als ook valide (AIMS, HAQ and SIP). De resultaten van de responsiviteit van deze drie meetinstrumenten bleken tegenstrijdig. Zoals ook in de hoofdstukken 4 en 5 is de hypothese onderzocht dat validering tegen een vergelijkbaar construct resulteert in sterkere correlaties dan validering tegen een niet-vergelijkbaar construct met betrekking tot beperkingen in de persoonlijke verzorging. Er bleek geen verschil in de correlatiecoëfficiënten tussen validering tegen een optimaal vergelijkbaar construct versus een niet-vergelijkbaar construct. Geconcludeerd wordt dat de AIMS het meest optimale meetinstrument is voor het meten van beperkingen in de persoonlijke verzorging. Voor de drie meetinstrumenten welke zowel betrouwbaar als valide blijken, zijn de gegevens over de responsiviteit tegenstrijdig. Op grond van de resultaten uit dit hoofdstuk werd geadviseerd om, indien mogelijk, naast een generieke uitkomstmaat tevens een (domein)specifieke uitkomstmaat te gebruiken.

In *hoofdstuk 7* worden de interpretatie en de extrapolatie van de resultaten van dit proefschrift bediscussieerd. De Top-3 van de beste beschikbare meetinstrumenten voor elk relevant domein wordt in één tabel samengevat. Verder wordt ter discussie gesteld in hoeverre de ICF geschikt is als uitgangspunt voor de literatuuranalyse zoals die is beschreven in dit proefschrift. De relevantie, voordelen en nadelen van het gebruik van de ICF worden besproken in relatie tot meetinstrumenten en de methode van data-analyse zoals die is toegepast in dit proefschrift. Tevens wordt in dit hoofdstuk ingegaan op de hypothese dat validering tegen een vergelijkbaar construct resulteert in sterkere correlaties dan validering tegen een niet-vergelijkbaar

construct, en op de consequenties ervan voor de interpretatie van de resultaten van dit literatuuronderzoek. Ook wordt ingegaan op de implicaties van recente ontwikkelingen in de klinimetrie met betrekking tot de responsiviteit. Tot slot worden de resultaten van dit proefschrift besproken in relatie tot die van de OMERACT-groep (OMERACT: Outcome Measures in Rheumatoid Arthritis Clinical Trials). Hoofdstuk 7 eindigt met de implicaties voor toepassing in klinische context en in klinisch wetenschappelijk onderzoek, alsmede met enkele aanbevelingen voor toekomstig onderzoek.

Dankwoord

Een dissertatie schrijven in je vrije tijd en als extraneus heeft, naast voordelen, ook veel nadelen. Ik heb ze geloof ik allemaal mogen proeven, en kan inmiddels ook dáár een boek over schrijven. Desalniettemin een leerzaam traject, zij het niet altijd in positieve zin. Een aantal mensen ben ik bijzondere dank verschuldigd vanwege hun inzet en/of bijdrage en ondersteuning.

Allereerst mijn eerste promotor Lex Bouter: nadat je merkte dat je jarenlange (overigens terechte) ontmoedigingsbeleid geen effect meer had heb je veel voor me betekend. Veel respect heb ik voor je snelheid, je deskundigheid en je zakelijk en persoonlijk bijzonder prettige communicatie. Met name ben ik je dank verschuldigd voor de wijze waarop je je schouders eronder hebt gezet om dit voor mij mogelijk te maken en me daarin te ondersteunen. Nadat je merkte dat langer ontmoedigen niet meer hielp, heb je gezorgd voor een enorme stroomversnelling. Zonder het wellicht te realiseren heb je ook persoonlijk een belangrijk stempel gedrukt op mijn relativiseringsvermogen en mijn visie in positieve zin sterk veranderd.

Tevens mijn tweede promotor, Rob Oostendorp: zonder jou was ik in 1997 nooit aan dit project begonnen en had ik waarschijnlijk een heel andere, onvoorspelbare route gelopen. Beiden kregen we 'er een staart van', maar ik waardeer je brede kennis en je persoonlijkheid en de vele malen dat onze paden elkaar de afgelopen 25 jaar kruisten, en ben daarom blij dat je tot het eind aan boord bent gebleven.

Verder ben ik dank verschuldigd aan Pieter Dijkstra: op het diepste en donkerste punt in de tunnel kwam jij me met een lamp tegemoet. Zonder jouw hulp en ondersteuning, zowel inhoudelijk als moreel, was de handdoek al lang in de ring gegooid. Dat heeft veel voor me betekend! Ik meen te mogen zeggen dat we beiden, zoals boze tongen wel eens zeggen, 'niet de makkelijkste karakters' zijn, maar in samenwerking bleken die twee eigenwijze mensen een perfecte combinatie.

De leescommissie, Prof. dr. M. Boers, Prof. dr. M. de Haan, Prof. dr. G.J. Lankhorst, Prof. dr. H.C.W. de Vet, Prof. dr. P. Vaes, en dr. A. de Weijer wil ik bedanken voor het lezen en beoordelen van mijn proefschrift. In het bijzonder Riekje de Vet en Maarten Boers voor de vele kritische kanttekeningen en nuttige suggesties die in sterke mate hebben bijgedragen tot de nodige kwaliteitsverbetering van hoofdstukken 1 en 7.

Veel mensen in directe omgeving hebben een grote rol gespeeld in de morele ondersteuning en reflectie: in het bijzonder Jan en Annelies, Wendy en Anton wil ik hartelijk danken; gelukkig hadden jullie er meer vertrouwen in dan ik, dank voor jullie luisterende oren en warmte!

Lieve Mam, dank voor levenslang morele steun, vrijheid en belangstelling. U was het vaak niet eens met de deadlines en werkdruk die ik mezelf uit gedrevenheid meende

te moeten opleggen, maar bleef wel altijd achter me staan, waarvoor veel dank en respect! Heel jammer dat papa dit niet meer heeft mogen meemaken.

Mendy en Koen: dank voor jullie eindeloze geduld en begrip. Er zijn niet veel kinderen die moeten aanzien hoe twee ouders tegelijkertijd trachten een proefschrift af te ronden. Hopelijk zijn ook jullie blij dat er tenminste één boekje vast af is. Dank omschrijven is soms moeilijk, maar zoals Koen adviseerde: "Zeg toch gewoon: BEDANKT!"

Tot slot Ilse: de enige die me onvoorwaardelijk en consequent trouw bleef; dank voor je vele hulp, alle eindeloze gesprekken die we er over hadden en alle ruimte en vrijheid die je me al zoveel jaren geeft: petje af en dank voor zoveel liefde en steun!

About the author

Raymond A.H.M. Swinkels was born in Tilburg, in the Netherlands, on December 20th, 1955. After completing secondary school in Eindhoven, he studied Physiotherapy at the Academy for Physiotherapy in Breda from 1974 to 1978, followed by Orthopaedic Medicine in London (Dr. J. Cyriax), and Manual Therapy in Eindhoven (graduated in 1984). After graduation he studied Movement Sciences at the Vrije Universiteit in Amsterdam from 1985 to 1991. Since 1979 he has been working as a physiotherapist in a multidisciplinary Medical Care Centre in a primary care setting. In addition to his work as a physiotherapist, he was involved in teaching 'Methodology of Manual Therapy' at the Vrije Universiteit in Brussels (Belgium) from 1992 to 2001, at the University of Genova (Italy) (1999-2005) and at the Academic Hospital of the Ruhr-Universität in Recklinghausen (Germany) for the *Verband für Physikalische Therapie* (1994-2005). Since 1995 he has been a member of the Editorial Board of *Manual Therapy Journal*, and since 2003 Editor-in-Chief of *Stimulus* journal. During the past decade he has been active in various professional and scientific organisations, including the Scientific Board of Dutch Organization of Quality Assurance (CBO), the Health Council of the Netherlands, and the Royal Dutch Society for Physiotherapy. From 1997 to 1999 he worked as a researcher at the Dutch Institute of Allied Health Care in Amersfoort, where he did the groundwork for his PhD thesis on clinimetrics in rheumatic disorders. During the period 1998-2001 he followed various courses on statistics, methodology and research in the Postgraduate Epidemiology Programme organized by the Institute for Research and Extramural Medicine (EMGO Institute) and the Nutrition and Toxicology Research Institute in Maastricht (NUTRIM).

Core list of publications

Swinkels RAHM, Oostendorp RAB. Upper cervical instability: fact or fiction? *J Manip and Physiol Ther* 1996;19(3):185-194.

Swinkels RAHM, Beeton K, Alltree J. Pathogenesis of upper cervical instability. *Manual Ther* 1996;1(3):127-132.

Cattrysse E, Swinkels RAHM, Oostendorp RAB, Duquet W. Upper cervical instability: are clinical tests reliable? *Manual Ther* 1997;2(2):97-105.

Bie RA de, Hendriks H, Lensen A, Moorsel S van, Opraus K, Remkes W, Swinkels RAHM. KNGF-guideline acute ankle sprain. (in Dutch: KNGF-Richtlijn Acuut Enkelletsel). *Ned T Fysioth*, 1998;108(1):Suppl.

Swinkels RAHM, Visser E. Klinimetrie bij reumatische aandoeningen. *Ned Tijdschr Ergother*. 1998;26(6):111.

Vaes PH, Swinkels RAHM, Oostendorp RAB. Analysis of study populations, interventions and outcome measures in patients with low back pain. CD-ROM Proceedings of the 7th Scientific Conference of the International Federation of Orthopedic Manipulative Therapists (IFOMT). Singer KP, (Ed). ISBN 1740520270, November 2000:446-450.

Berkel DM van, Oostendorp RAB, Ravensberg CD van, Scholten-Peeters GGM, Storm I. Visser AC de, Pool JJM, Swinkels RAHM. Manuele therapie en fysiotherapie: verschil in kenmerken van patiënten. *Ned T Fysiother* 2002;112(6):143-152.

Oostendorp, RAB, Berkel, DM van, Ravensberg, CD van, Scholten-Peeters, GGM, Storm, I. Visser, AC de, Pool, JJM, Swinkels, RAHM. Manuele therapie en fysiotherapie: verschil in kenmerken van patiënten met aspecifieke lage-rugpijn. *Ned T Fysiotherapie* 2002;112(6):153-159.

Swinkels RAHM, Swinkels-Meewisse EJCM. Clinical diagnostics of pregnancy related low back pain in physiotherapy practice. (in Dutch: Klinische diagnostiek van zwangerschapsgelateerde rugpijn in de fysiotherapiepraktijk). *Jaarboek Fysiotherapie/Kinesitherapie* 2003, Chapter 17:228-243. Bohn Stafleu van Loghum, Houten, 2002.

Swinkels-Meewisse EJCM, Swinkels RAHM, Verbeek ALM, Vlaeyen JWS, Oostendorp RAB. Psychometric properties of the Tampa Scale for kinesiophobia and the fear-avoidance beliefs questionnaire in acute low back pain. *Manual Ther* 2003;8(1):29-36.

Bie RA de, Heemskerk MAMB, Stomp DJ, Lensen AF, Moorsel S. van, Rondhuis G, Swinkels RAHM, Hendriks HJM. KNGF-richtlijn Chronisch enkelletsel. Ned T Fysioth 2003;2: suppl 1-38.

Swinkels RAHM. Which questionnaires for what patient? (in Dutch: Welke vragenlijsten bij welke patiënt?). Stimulus 2003;22(3):302-313.

Swinkels RAHM, Oostendorp RAB, Bouter LM. Which are the best instruments for measuring disabilities in gait and gait-related activities in patients with rheumatic disorders? Clin Exp Rheum 2004;1:25-33.

Swinkels-Meewisse EJCM, Swinkels RAHM, Verbeek ALM, Vlaeyen JWS, Oostendorp RAB. Statistics and pain-related fear measures in acute low back pain. Manual Ther 2004;9(1):47-48.

Oostendorp RAB, Scholten-Peeters GGM, Swinkels RAHM, Bekkering GE, Heijmans MWFGJ, Huijbregts PA, Hendriks HJM. Evidence-Based Practice in Physical and Manual Therapy: Development and Content of Dutch National Practice Guidelines for Patients with Non-Specific Low Back Pain. J Manual & Manipulative Ther. 2004;1:21-31.

Swinkels RAHM, Oostendorp RAB. Outcome assessment and measurement in spinal musculoskeletal disorders. In: Grieve's Modern Manual Therapy, The vertebral column, 3rd ed. By JD Boyling and GA Jull (Eds). Churchill Livingstone. 2004, Chapter 41: 591-601.

Swinkels RAHM. The ICF-classification as a system for structuring outcome measurement. Physiotherapy Singapore 2004; 7(3):7-13.

Swinkels RAHM, Dijkstra PU, Bouter LM. Reliability, validity and responsiveness of instruments to assess disabilities in personal care in patients with rheumatic disorders: a systematic review. Clin Exp Rheum 2005;23:71-79.

Swinkels RAHM, Bouter LM, Oostendorp RAB, Swinkels-Meewisse EJCM, Dijkstra PU, Vet HCW de. Construct validity of instruments measuring impairments in body structures and function in rheumatic disorders: which constructs are selected for validation? A systematic review. Clin Exp Rheum 2005; in press.

Ravensberg CD van, Oostendorp RAB, Berkel LM van, Scholten-Peeters GGM, Pool JJM, Swinkels RAHM, Huijbregts PA. Physical Therapy and Manual Physical Therapy: Differences in Patient Characteristics. J Manual & Manip Ther 2005,13(2):113-124.

Swinkels RAHM, Bouter LM, Oostendorp RAB, Ende CHM van den. Impairment measures in rheumatic disorders for rehabilitation medicine and allied health care. A Systematic review. *Rheum Int* 2005;25(7):501-512.