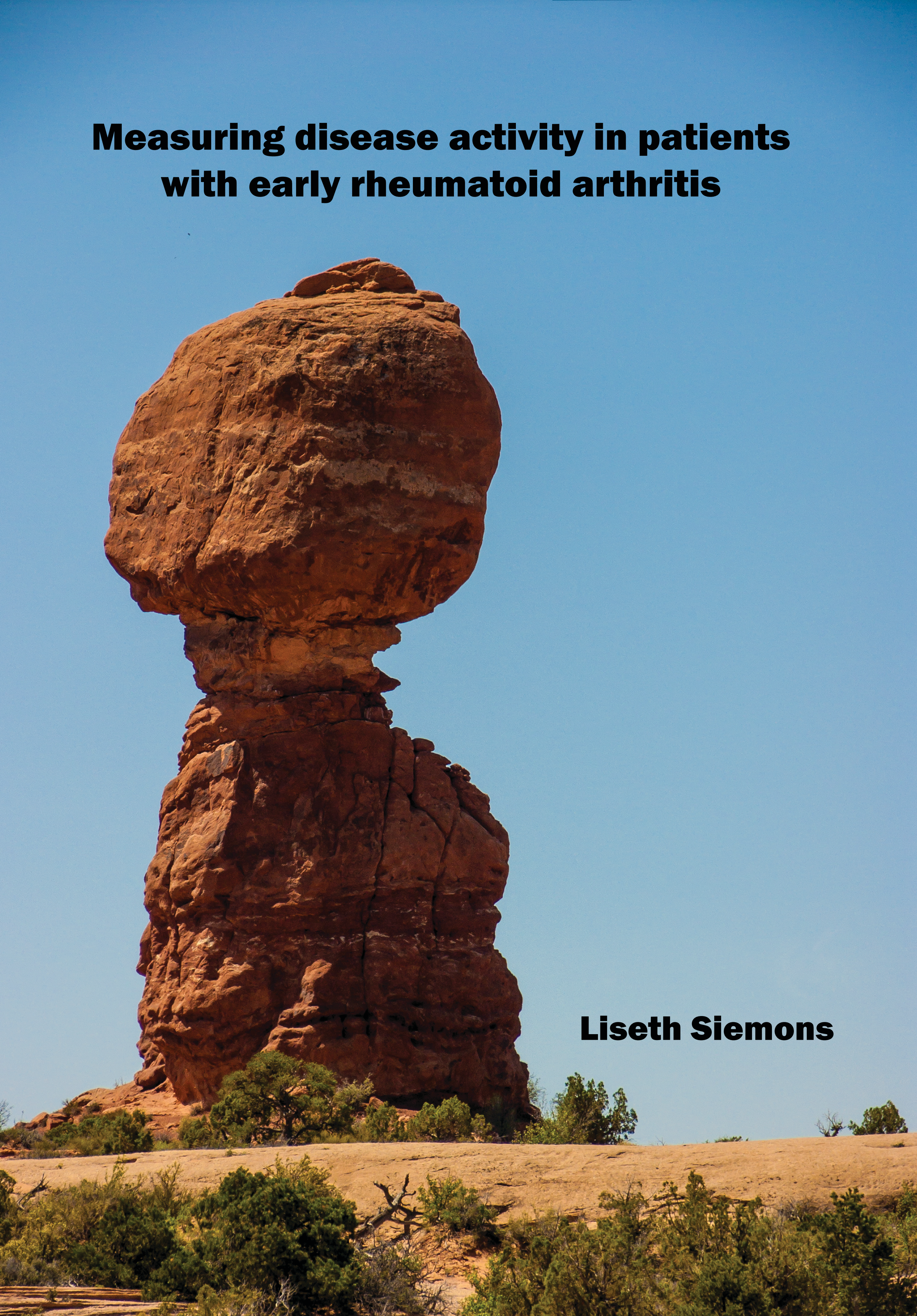
Measuring disease activity in patients with early rheumatoid arthritis

Liseth Siemons



Measuring disease activity in patients with early rheumatoid arthritis

Liseth Siemons

Committee members:

Supervisors:	Prof. dr. M.A.F.J. van de Laar	Universiteit Twente, Medisch Spectrum Twente
	Prof. dr. C.A.W. Glas	Universiteit Twente
Assistant-supervisor:	Dr. P.M. ten Klooster	Universiteit Twente
Members:	Prof. Dr. E. Krishnan	Stanford University (CA, USA)
	Prof. dr. P.L.C.M. van Riel	Radboud universitair medisch centrum
	Prof. dr. G. Zielhuis	Radboud universitair medisch centrum
	Dr. H. Vonkeman	Universiteit Twente, Medisch Spectrum Twente
	Prof. dr. J.A.M. van der Palen	Universiteit Twente, Medisch Spectrum Twente
	Prof. dr. R. Sanderman	Universiteit Twente, Rijksuniversiteit Groningen

Funding for this work was provided by the Dutch RhEumatoid Arthritis Monitoring (DREAM) collaboration.

PhD thesis, University of Twente
ISBN: 978-90-365-3661-5
DOI: 10.3990/1.9789036536615
Cover: Balanced Rock, Arches National Park (UT, USA)
Cover design by: Liseth Siemons
Printed by: Gildeprint Drukkerijen, Enschede, The Netherlands

The publication of this thesis was financially supported by the Medisch Spectrum Twente and the Dutch Arthritis Foundation.

Medisch Spectrum Twente

Reumafonds
Dutch Arthritis Foundation

MEASURING DISEASE ACTIVITY IN PATIENTS WITH EARLY RHEUMATOID ARTHRITIS

PROEFSCHRIFT

ter verkrijging van
de graad van doctor aan de Universiteit Twente,
op gezag van de rector magnificus,
prof. dr. H. Brinksma,
volgens besluit van het College voor Promoties
in het openbaar te verdedigen
op vrijdag 4 juli 2014 om 14.45 uur

door

Liseth Siemons

geboren op 18 juni 1985
te Sneek

Dit proefschrift is goedgekeurd door:

Prof. dr. M.A.F.J. van de Laar (promotor)
Prof dr. C.A.W. Glas (promotor)
Dr. P.M. ten Klooster (assistent-promotor)

For Charles, Reini, Wolter, Arnout, and Kien

Acknowledgements

Finishing a PhD thesis is not something you achieve on your own, there are many people involved. I want to take this opportunity to thank everyone who supported me during these past 4 years and I would like to mention some of them in particular.

First of all, my thanks go to my daily supervisor, Peter ten Klooster. Peter, your door was always open and you always had (or made) time to help, give advice, or have a critical discussion. With your scientific understanding, keen eye for detail, and extensive statistical knowledge, your involvement certainly raised this thesis to a higher level. I'm grateful for your guidance and support, and I thank you for the opportunity you have given to me to learn from you during the past 4 years.

Next, I would like to thank my supervisors, Mart van de Laar and Cees Glas. Mart, thank you for your confidence when you offered me this PhD position and for making this valuable experience possible. During the project you always kept track of the "bigger picture", determining whether we were going in the right direction or whether adaptations or our original plans were necessary. I have much respect for your professionalism in and dedication to clinical research and the freedom that you were willing to give me over the years, including the opportunity to go abroad. I experienced your optimism and involvement as very valuable and your clinical insights were definitely of great importance throughout this thesis.
Cees, your expertise on item response theory and generalizability theory was indispensable for the realization of this thesis. It wasn't always easy for me to understand the, in my eyes, complex statistical formulas, but you always made sure that I (at least) understood the general idea behind it. I have always experienced your enthusiasm about statistical "problems" or "complexities" as inspiring. More than once I went to you with a question about an article we were working on and you did not only give me an answer, but a whole bunch of new research ideas as well. You were always looking ahead to things we could aspire to after we finished the current study. There is always more to discover and more to learn. Thank you for sharing your knowledge and enthusiasm with me.

It should not go unnoticed that this research would not have been possible without the funding of the Dutch RhEumatoid Arthritis Monitoring registry (DREAM), the rheumatologists and nurses who recruited patients and collected data, and the patients who gave their consent to participate. A special thanks must also go to Nancy, for answering all my questions about the database and the code book, and to Mirjam, for helping me with the organization of the supplementary case history examinations. Also, a word of thanks to all the co-authors who have contributed to my publications. Erik Taal, Harald Vonkeman, Ina Kuper, and Piet van Riel, thank you for your valuable scientific input and the nice collaboration. Harald, I'm glad that your involvement in my project grew over the years. I really value your clinical expertise and appreciate your openness and keen eye for detail.

I would like to express my gratitude to the members of my Graduation committee as well: Eswar Krishnan, Piet van Riel, Gerhard Zielhuis, Harald Vonkeman, Job van der Palen, and Robbert Sanderman. Thank you for your willingness to be part of my committee. A special thanks to Eswar, who granted me the opportunity to stay with his group at Stanford University for a period of 3 months. I very much appreciate the way you work, your involvement in the weekly meetings, and your patience when explaining things to the group or person-to-person. I'm thankful that I could be part of that for a little while. It was an amazing experience and I enjoyed it very much. Narinder, Weiqi and Linjun, thank you for the very warm welcome you gave me and for the fantastic trip to Alcatraz. Gwen, thank you for your kindness and the many nice chats we had near the coffee machine. And finally, I want to express my gratitude to Joni, who was willing to share her home with me during my stay. I know I was often busy, but I'm grateful for your hospitality and the great trips that we were able to make during my stay.

Then, of course, I also want to thank my Dutch colleagues. Thank you all for the very good time, for all the "gezelligheid", and the many nice chats. Without you these 4 years would not have been so great. You created a good working environment and a friendly atmosphere. I've had many roommates over the years and I appreciated the company of every one of you, but special thanks should go to my fellow-tower-colleagues and lunch-break buddy's. Between the hard work there was always room for a little chat or a serious conversation. I really enjoyed your company, the lunches we had together, and the dinner or wine tasting evenings at each other's homes. I hope we will stay in touch and will continue to have these social gatherings in the future. I would also like to thank our secretaries. You always stood by to assist with flight reservations, the ordering of materials, the arrangement of certain payments, and many, many other things. Even the organization of a conference or summerschool was no problem. Marieke, thanks for lending me a hand with that.

Then last, but definitely not least, I would like to thank my family and friends for their interest in my work and the support they offered me and there are certain people who I want to mention in particular. First, I couldn't have wished myself a better family. Mom and dad, you are the best. Thank you so much for always being there for me and believing in me. Thank you for your never-ending love, relentless support, and for helping me to relax when I needed to get my mind away from work. Arnout and Wolter, it is really nice that you wanted to be my paranimfs, but what makes me really happy and immensely proud is the fact that I have such great older brothers. Then, finally, my dearest Kien. I cannot thank you enough for your unconditional love, endless support, and infinite patience and positivism. Thanks for always being there for me, both now and in the future.

Enschede, May 2014

Contents

Acknowledgements vii

Chapter 1 Introduction 1

Chapter 2 Modern psychometrics applied in rheumatology: A systematic review 15

Chapter 3 Validating the 28-tender joint count using item response theory 47

Chapter 4 Contribution of assessing forefoot joints in early rheumatoid arthritis patients: Insights from item response theory 65

Chapter 5 Interchangeability of 28-joint disease activity scores using the erythrocyte sedimentation rate or the C-reactive protein as inflammatory marker 81

Chapter 6 How age and sex affect the erythrocyte sedimentation rate and C-reactive protein in early rheumatoid arthritis 93

Chapter 7 Distinct trajectories of disease activity over the first year in early rheumatoid arthritis patients following a treat-to-target strategy 105

Chapter 8 Further optimization of the reliability of the 28-joint Disease Activity Score in patients with early rheumatoid arthritis 117

Chapter 9 General discussion 131

List of publications 141

Summary / Samenvatting 143

Chapter 1

Introduction

Rheumatoid arthritis

Rheumatoid arthritis (RA) is a chronic autoimmune disease, which is characterized by symmetric inflammations of the peripheral joints [1-3]. The course of the disease is highly variable and has a substantial impact on the physical, psychological, and social health of the patients and, in turn, on society in terms of health care costs and decreased productivity. Predominant signs and symptoms of the disease include joint pain, joint swelling, joint damage or deformations, stiffness, fatigue, diminished physical and social functioning, and depression [1-6]. The worldwide prevalence rate of RA is estimated to be around 0.5% to 1.0% of the adult population in developed countries [1, 2, 7, 8], tends to increase with age, and is higher in females than in males.

Although the precise aetiology of RA remains unknown, the pathogenesis is assumed to be multifactorial, involving both genetic as well as environmental factors [1, 2, 7, 9]. New insights into the biological mechanisms have resulted in new medicines and drastic changes in the treatment strategies over the past decade, which have led to significant improvements in the future prospects of patients. Where previous treatments were mostly aimed at controlling symptoms and lacked effectiveness in the prevention of joint destruction and maintenance of functional status, current treatments have a more aggressive approach aiming to suppress RA disease activity as fast as possible, as completely as possible, and as long as possible [2, 3]. For this purpose, different relatively strict monitoring and treatment strategies are used [10, 11]. These strategies emphasize a strong interference early on in the disease and have proven to be effective in early RA patients in daily clinical practice [12], resulting in more and faster remissions compared to usual care. In particular, there are significantly larger improvements in: a) physical functioning, b) the (reduced) number of tender and swollen joints, and c) the patient-reported assessments of pain and general health [12].

Disease activity

These disease-activity-driven, or treat-to-target, treatment strategies require optimal measurement of disease activity. A widely used measure for monitoring disease control is the Disease Activity Score for 28 joints (DAS28). The DAS28 is a simplified form of its predecessor, the Disease Activity Score (DAS).

DAS

RA disease activity is a complex and multidimensional construct. Traditionally, rheumatologists based their clinical judgment of a patient's disease activity on a combination of factors, including laboratory measures, clinical assessments, and radiographic information [13]. However, this unstructured assessment method resulted in large discrepancies between rheumatologists and, in an attempt to formalize this process and reduce discrepancies, a quantitative disease activity index was proposed: the DAS [13].

Starting with a large collection of previously reported potential markers of disease activity, factor analysis was used to classify these variables into groups. The factors that discriminated best between patients with low and high disease activity (as defined by the rheumatologists' treatment decisions) were retained and their most influential variables were extracted using multiple regression analysis. This resulted in a continuous composite score reflecting a patient's current disease activity [13], based on (in decreasing importance) the graded Ritchie score (i.e. a tender joint count in 53 joints), a binary rated swollen joint count in 44 joints, a non-specific measure of inflammation called the erythrocyte sedimentation rate, and a patient-reported global assessment of general health.

DAS28

Though the development of the DAS aided the quantification of a patient's disease activity, its practical usefulness in clinical practice was limited due to the elaborate measurement of all 53 tender and 44 swollen joints. This is not only a tedious and time-consuming job for the clinician, but for the patient as well. Furthermore, some of the included joints are very difficult to assess reliably or might show abnormalities because of processes beyond the rheumatic disease [14-16]. Consequently, several studies examined the usability of reduced joint counts and found that they were as valid and reliable as more extensive joint counts [14, 16-19], which led to the development of a modified DAS score, i.e. the DAS28. The erythrocyte sedimentation rate (ESR) and general health measure were retained, but the extensive joint counts were reduced to only 28 joints [18]. Figure 1 gives an overview of the joints administered, including the shoulders [2 joints], elbows [2 joints], wrists [2 joints], metacarpophalangeals (MCPs) [10 joints], proximal interphalangeals (PIPs) [10 joints], and knees [2 joints] [20]. All joints are measured on a binary scale, where 0 reflects no pain/no swelling and 1 reflects the presence of pain/swelling, and are summed into a 28-tender joint count (TJC28) and 28-swollen joint count (SJC28).

These more or less objective, semi-objective and subjective measures were combined into a continuous scale of RA disease activity, using the following equation [18, 20]:

$$\textit{DAS28-ESR} = 0.56 * \sqrt{TJC28} + 0.28 * \sqrt{SJC28} + 0.70 * Ln(ESR) + 0.014 * GH.$$

The resulting DAS28 score reflects present disease activity, ranging from 0 to approximately 10, and can be used to classify patients as being in remission [DAS28 <2.6], or as having low disease activity [2.6 ≤ DAS28 ≤ 3.2], moderate disease activity [3.2 < DAS28 ≤ 5.1], or high disease activity [DAS28 >5.1] [20].

When following a disease-activity-driven treatment strategy, therapy adjustments defined by protocol are often based on these resulting DAS28 scores. However, these DAS28 scores have their limitations. Apart from disease activity, all individual component scores might be influenced by comorbidities [21-23], joint counts can also be affected by physician and patient related factors [24], and a patient's general health rating can be elevated because of non-inflammatory or personal factors [25]. On a population level the DAS28 score gives probably a good estimation of disease activity in RA patients but in individual RA patients inconsistencies can occur. Moreover, a rheumatologist might still observe disease activity in omitted joints when the DAS28 indicates a state of remission. Therefore, to better interpret a DAS28 score, it is essential to have thorough insight into its individual components and their shortcomings.

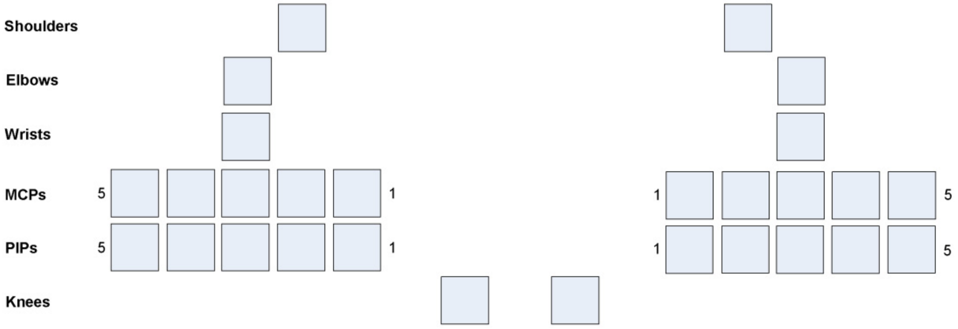


Figure 1 - The 28 joints included in the tender and swollen joint counts of the DAS28.
MCP = metacarpophalangeal, PIP = proximal interphalangeal

Individual components DAS28

Joint counts

RA is an inflammatory disease which predominantly manifests itself in warm, reddish, painful and swollen joints. Joints are often referred to as the major "organ" involved in RA

[26, 27]. Consequently, joint counts are considered essential for assessing RA disease activity. They belong to the core set of disease activity measures and their use is recommended by the American College of Rheumatology as well as the European League Against Rheumatism [20, 27, 28].

Since the assessment of all joints is unfeasible and inaccurate in clinical practice, mainly due to time constraints, reduced joints counts have been proposed over time. The DAS28 uses 28-joint counts to assess pain and swelling, which have been shown to correlate strongly with more extensive joint counts [14, 16-19]. Each visit, the presence of both joint pain and swelling is assessed by a trained rheumatologist or nurse practitioner by exerting a certain amount of pressure on the patient's joints. This is called a semi-objective assessment method because the degree of joint pain depends on the patient's pain threshold as well as the amount of pressure exerted by the rheumatologist, whereas the assessment of joint swelling depends on the physician's perceptions [24]. Large intra- and inter-observer variability have been reported in both joints counts, although reliability was generally found to be higher for the tender joint count [29]. Rating differences might be explained by factors as the assessors' levels of training and experience, a lack of standardization in examination methods, unclear definitions of what a swollen joint looks like, or the degree of joint deformity [29, 30]. Reliability can be improved by training, the use of the same assessor at each assessment, or the use of standardized guidelines for joint assessment [29, 31].

Another issue related to these reduced joint counts concerns the presence of residual joint activity in omitted joints [32-34], which has raised the question whether the 28-joint counts are sufficient to reflect a patient's current disease activity. The omission of certain joints seems warranted though because of major assessment difficulties or because of influential processes that go beyond RA. For instance, it is very difficult to assess swelling in the hip joints, and foot abnormalities might result from fluid retention or wearing the wrong footwear [14] instead of inflammation. When including these joints, they might provide relevant clinical information but they might also add unwanted random error variance to the measurement.

Acute phase reactants

Because RA is an inflammatory disease, it is not surprising that the DAS28 includes an acute phase reactant as an indicator of disease activity. Even though a single laboratory measure can only reflect part of the whole inflammatory response, it can be clinically helpful for evaluating inflammatory severity and for monitoring disease activity over time [22]. Although the DAS28 was originally developed with the ESR as acute phase reactant, it has been argued that the C-reactive protein (CRP) might give a better reflection of current disease activity because of its faster response to inflammatory stimuli [22, 35]. Consequently, another DAS28 formula was developed using CRP instead of ESR [36]:

$$\textit{DAS28-CRP} = 0.56 * \sqrt{TJC28} + 0.28 * \sqrt{SJC28} + 0.36 * Ln(CRP + 1) + 0.014 * GH + 0.96.$$

The ability to calculate DAS28 scores based on different acute phase reactants can be convenient, but it does raise the question whether these two measures can be used interchangeably because both inflammatory markers measure different aspects of the disease. Where ESR values are assumed to reflect the patient's disease activity over the past few weeks, CRP values may be a better reflection of short-term changes in disease activity [20, 22, 36, 37].

The ESR is the rate at which red blood cells fall through plasma in a tube of blood in one hour time. It is an indirect way of screening for elevated concentrations of acute phase plasma proteins in the blood (e.g. fibrinogen) because these cause the red blood cells to settle down more rapidly [22]. ESR concentrations respond slowly to inflammatory stimuli and, consequently, to changes in a patient's disease activity [22, 35, 37]. The CRP, on the other hand, is a direct measure of the acute phase reaction [35, 37-39], responding more rapidly to changes in inflammatory stimuli [22, 35, 37, 39, 40]. It is a protein that is produced in the liver as a reaction to certain biologic ligands that appear when inflammation develops [22] and is part of the body's defense mechanisms against damaged cells or pathogens. Assessment of both inflammatory markers is relatively easy, quick, and inexpensive [22, 35, 38-40], but the assessment of the ESR requires a fresh blood sample [22, 40], while the measurement of CRP can also been performed on stored frozen specimens in a central laboratory [22, 37, 40].

Their interpretation, however, is complicated by the fact that both are non-specific acute phase reactants of systemic inflammation, which means that elevated levels are not necessarily (solely) due to the inflammation of the rheumatic disease but can also be influenced by other factors. ESR, for instance, can be increased by other inflammatory conditions and non-inflammatory-related biological phenomena like paraproteinemia, abnormally shaped or sized red blood cells, changes in plasma composition, anemia, pregnancy, or certain drugs [22, 35, 37, 39, 40]. CRP production, on the other hand, can be hampered in liver disease and concentrations can be elevated in the presence of comorbid diseases like osteoarthritis, gout, and systemic lupus erythematosus, in the presence of bacterial infections, or because of non-inflammatory influences like sleep deprivation or unhealthy diets [22]. Finally, some medications, such as tocilizumab, have a direct effect on CRP production bypassing the disease processes of RA.

General health

Where the severity and impact of most rheumatic conditions were initially typically evaluated with clinical measures, patient-reported outcome measures (PROs) have gained attention since the eighties of the last century [41, 42] and are now even part of the core set of disease activity measures as defined by the American College of Rheumatology [28] as well as the ACR/EULAR remission criteria [27]. The DAS28 also includes a patient-reported outcome measure – a visual analog scale of general health (GH) – which asks the patient to give an overall assessment of how they currently feel, considering all the ways in which the RA influences their lives (Figure 2).

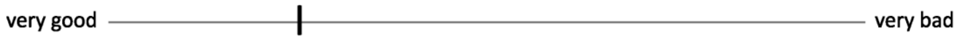


Figure 2 – Visual analog scale of general health on which patients indicate how they currently feel, considering all the ways in which the RA influences their lives.

A visual analog scale can be administered quick and easy, yet score interpretation can be very difficult because the patient is asked to consider all the ways in which the RA influences his or her life, touching upon different dimensions of the disease. As defined by the World Health Organization, health is not merely the absence of disease but involves the whole spectrum of physical, mental and social wellbeing [43], all of which can affect the patient's GH score. Consequently, a score of 20 for one person does not need to have the same meaning as a score of 20 in another person. The inclusion of the GH component has often been criticized. It is not only the most subjective component of the DAS28, but GH scores can also be elevated when none of the other DAS28 components point to an active disease, possibly due to certain non-inflammatory effects of RA like pain or fatigue [44]. Additionally, the patient's perception of health has been shown to differ across patients with equal DAS28 scores, which might be explained by the occurrence of response shifts during the disease course [45]. Still, it is believed that (semi-)objective clinical measures and subjective patient-reported outcome measures address different aspects of the disease [46] and that both should be administered to evaluate a multifactorial concept like RA disease activity.

Objective of this thesis

Although the DAS28 has proven its value in rheumatology, several doubts have been raised about residual disease activity in omitted joints [32-34], the influence of external factors on the level of inflammatory markers [22, 47-62], and the inclusion of a subjective patient-reported outcome measure [44, 45]. The aim of this thesis is to address some of these issues to improve our understanding of the complexities behind the measurement of disease activity in early rheumatoid arthritis patients.

Outline of this thesis

Modern psychometrics and joint counts

Statistical methods are essential for developing or evaluating outcome measures. So far, the majority of clinical measures have been developed using methods from classical test theory. In recent years, however, item response theory (IRT) emerged as an alternative, complementary psychometric method. IRT enables a more thorough evaluation of an instrument's psychometric characteristics by providing more detailed information on the item level [63].

Although IRT originated from the field of educational measurement, it is gaining attention in the medical field as well [64-67] and the first study of this thesis provides an overview of its use within rheumatology (**Chapter 2)**. This study shows that IRT applications have markedly increased over the past decades, although they are primarily being applied to patient-reported measures. To examine its applicability to clinical disease activity measures more thoroughly, it was evaluated whether IRT could be used to internally validate the 28-tender joint count (**Chapter 3**).

The focus on the item level provides the interesting opportunity to evaluate the measurement range and precision of the 28 included joints. A commonly raised criticism of the DAS28 is that patients might experience residual disease activity in excluded joints, especially in the feet. To address this issue, IRT information curves were used to evaluate the contribution of the forefoot joints to the measurement range of the 28-joint counts (**Chapter 4)**.

Inflammatory markers

When working with two DAS28 scores, depending on a preference for or the availability of the ESR or CRP, it is important to know whether these scores can be used interchangeably. Score discrepancies might lead to different interpretations of a patient's level of disease activity and can have far-reaching effects on treatment decisions. As such, the interchangeability of the two DAS28 scores is examined in **Chapter 5**.

Although the CRP levels are supposed to be less sensitive to external influences than the ESR [22, 35, 38, 39], an extensive number of studies have demonstrated similar dependencies on gender, age, and body mass index for both acute phase reactants, suggesting that these factors should be taken into account when interpreting a patient's DAS28 score. Yet, the extent of these effects remain unknown and are evaluated in **Chapter 6**.

Heterogeneity of the early RA population

As mentioned earlier, DAS28 scores are often embedded within the treatment protocols with the current emphasis on early aggressive treatments. They are used as a criterion variable for treatment decisions or for evaluating treatment effectiveness [12, 68]. In this line of thought it is important to evaluate whether individual differences exist in the course of disease activity (**Chapter 7**), because patients with different response patterns might be in need of different therapeutic interventions.

DAS28 reliability

The DAS28 established itself as a valid measure [18], but since reliability is a prerequisite for validity, it should be reliable as well. Yet common internal consistency measures like Cronbach's alpha are ill-suited for the assessment of composite reliability, i.e. when different components of disease activity are combined into one index measure. Therefore, **Chapter 8** applies principles from generalizability theory to determine the reliability of the DAS28 in patients with early RA.

Finally, a general discussion about the findings in chapter 2 to 8 is given in **Chapter 9**. It elaborates on clinical implications and explores future research directions.

References

1. Sangha O. Epidemiology of rheumatic diseases. Rheumatology. 2000;39(2):3-12.
2. Turkiewicz AM, Moreland LW. Rheumatoid arthritis. In: Bartlett SJ, editor. Clinical care in the rheumatic diseases. Atlanta (GA): Association of Rheumatology Health Professionals; 2006. p. 157-66.
3. Lee DM, Weinblatt ME. Rheumatoid arthritis. The Lancet. 2001;358:903-11.
4. Uhlig T, Loge JH, Kristiansen IS, Kvien TK. Quantification of reduced health-related quality of life in patients with rheumatoid arthritis compared to the general population. J Rheumatol. 2007;34(6):1241-7.
5. Pollard L, Choy EH, Scott DL. The consequences of rheumatoid arthritis: quality of life measures in the individual patient. Clin Exp Rheumatol. 2005;23(5 Suppl 39):S43-52.
6. Bath J, Hooper J, Giles M, Steel D, Reed E, Woodland J. Patient perceptions of rheumatoid arthritis. Nurs Stand. 1999;14(3):35-8.
7. Scott DL, Wolfe F, Huizinga TWJ. Rheumatoid arthritis. The Lancet. 2010;376:1094-108.

8. Gabriel SE, Michaud K. Epidemiological studies in incidence, prevalence, mortality, and comorbidity of the rheumatic diseases. Arthritis Res Ther. 2009;11(3):229.
9. Tobón GJ, Youinou P, Saraux A. The environment, geo-epidemiology, and autoimmune disease: Rheumatoid arthritis. J Autoimmun. 2010;35:10-4.
10. Smolen JS, Aletaha D, Bijlsma JW, Breedveld FC, Boumpas D, Burmester G, et al. Treating rheumatoid arthritis to target: recommendations of an international task force. Ann Rheum Dis. 2010;69(4):631-7.
11. Combe B, Landewe R, Lukas C, Bolosiu HD, Breedveld F, Dougados M, et al. EULAR recommendations for the management of early arthritis: report of a task force of the European Standing Committee for International Clinical Studies Including Therapeutics (ESCISIT). Ann Rheum Dis. 2007;66(1):34-45.
12. Schipper LG, Vermeer M, Kuper HH, Hoekstra MO, Haagsma CJ, Den Broeder AA, et al. A tight control treatment strategy aiming for remission in early rheumatoid arthritis is more effective than usual care treatment in daily clinical practice: a study of two cohorts in the Dutch Rheumatoid Arthritis Monitoring registry. Ann Rheum Dis. 2012;71(6):845-50.
13. Van der Heijde DMFM, van ’t Hof MA, van Riel PLCM, Theunisse LAM, Lubberts EW, van Leeuwen MA, et al. Judging disease activity in clinical practice in rheumatoid arthritis: First step in the development of a disease activity score. Annals of the Rheumatic Diseases. 1990;49:916-20.
14. Fuchs HA, Brooks RH, Callahan LF, Pincus T. A simplified twenty-eight-joint quantitative articular index in rheumatoid arthritis. Arthritis Rheum. 1989;32:531.
15. van Tuyl LHD, Britsemmer K, Wells GA, Smolen JS, Zhang B, Funovits J, et al. Remission in early rheumatoid arthritis defined by 28 joint counts: limited consequences of residual disease activity in the forefeet on outcome. Ann Rheum Dis. 2012;71(1):33-7.
16. Fuchs HA, Pincus T. Reduced joint counts in controlled clinical trials in rheumatoid arthritis. Arthritis Rheum. 1994;37(4):470-5.
17. Prevoo MLL, van Riel PLCM, van ‘t Hof MA, van Rijswijk MH, van Leeuwen MA, Kuper HH, et al. Validity and reliability of joint indices. A longitudinal study in patients with recent onset rheumatoid arthritis. Br J Rheumatol. 1993;32(7):589-94.
18. Prevoo ML, van ’t Hof MA, Kuper HH, van Leeuwen MA, van de Putte LB, van Riel PL. Modified disease activity scores that include twenty-eight–joint counts: development and validation in a prospective longitudinal study of patients with rheumatoid arthritis. Arthritis Rheum. 1995;38(1):44-8.
19. 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(1):38-43.
20. Van Riel PLCM, Fransen J, Scott DL. Eular handbook of clinical assessments in rheumatoid arthritis. Alphen aan den Rijn: van Zuiden Communications; 2004.
21. van Tuyl LH, Boers M. Patient's global assessment of disease activity: what are we measuring? Arthritis Rheum. 2012;64(9):2811-3.
22. Firestein GS, Budd RC, Harris Jr ED, McInnes IB, Ruddy S, Sergent JS. Kelly's textbook of rheumatology. Philadelphia: Saunders Elsevier; 2009.
23. Leeb BF, Andel I, Sautner J, Nothnagl T, Rintelen B. The DAS28 in rheumatoid arthritis and fibromyalgia patients. Rheumatology. 2004;43(12):1504-7.
24. Thompson PW, Kirwan JR. Joint count: A review of old and new articular indices of joint inflammation. Br J Rheumatol. 1995;34:1003-8.

25. Masri KR, Shaver TS, Shahouri SH, Wang S, Anderson JD, Busch RE, et al. Validity and reliability problems with patient global as a component of the ACR/EULAR remission criteria as used in clinical practice. J Rheumatol. 2012;39(6):1139-45.
26. Kapral T, Dernoschnig F, Machold KP, Stamm T, Schoels M, Smolen JS, et al. Remission by composite scores in rheumatoid arthritis: are ankles and feet important? Arthritis Res Ther. 2007;9(4):R72.
27. Felson DT, Smolen JS, Wells G, Zhang B, van Tuyl LHD, Funovits J, et al. American College of Rheumatology/European League Against Rheumatism provisional definition of remission in rheumatoid arthritis for clinical trials. Arthritis Rheum. 2011;63(3):573-86.
28. Felson DT, Anderson JJ, Boers M, Bombardier C, Chernoff M, Fried B, et al. The American College of Rheumatology preliminary core set of disease activity measures for rheumatoid arthritis clinical trials. The Committee on Outcome Measures in Rheumatoid Arthritis Clinical Trials. Arthritis Rheum. 1993;36(6):729-40.
29. Cheung PP, Gossec L, Mak A, March L. Reliability of joint count assessment in rheumatoid arthritis: A systematic literature review. Semin Arthritis Rheum. In press.
30. Marhadour T, Jousse-Joulin S, Chalès G, Grange L, Hacquard C, Loeuille D, et al. Reproducibility of joint swelling assessments in long-lasting rheumatoid arthritis: influence on Disease Activity Score-28 values (SEA-Repro study part I). J Rheumatol. 2010;37(5):932-7.
31. Pincus T. Limitations of a quantitative swollen and tender joint count to assess and monitor patients with rheumatoid arthritis. Bull NYU Hosp Jt Dis. 2008;66(3):216-23.
32. Landewé R, van der Heijde D, van der Linden S, Boers M. Twenty-eight-joint counts invalidate the DAS28 remission definition owing to the omission of the lower extremity joints: A comparison with the original DAS remission. Ann Rheum Dis. 2006;65:637-41.
33. Mäkinen H, Kautiainen H, Hannonen P, Sokka T. Is DAS28 an appropriate tool to assess remission in rheumatoid arthritis? Ann Rheum Dis. 2005;64:1410-3.
34. van der Leeden M, Steultjens MP, van Schaardenburg D, Dekker J. Forefoot disease activity in rheumatoid arthritis patients in remission: results of a cohort study. Arthritis Res Ther. 2010;12(1):R3.
35. Kushner I. C-reactive protein in rheumatology. Arthritis Rheum. 1991;34(8):1065-68.
36. Fransen J, Welsing PMJ, de Keijzer RMH, van Riel PLCM. Disease activity scores using C-reactive protein: CRP may replace ESR in the assessment of RA disease activity. Ann Rheum Dis. 2003;62 (Suppl. 1):151.
37. Wolfe F. Comparative usefulness of C-reactive protein and erythrocyte sedimentation rate in patients with rheumatoid arthritis. J Rheumatol. 1997;24(8):1477-85.
38. Crowson CS, Rahman MU, Matteson EL. Which measure of inflammation to use? A comparison of erythrocyte sedimentation rate and C-reactive protein measurements from randomized clinical trials of golimumab in rheumatoid arthritis. J Rheumatol. 2009;36(8):1606-10.
39. Husain TM, Kim DH. C-reactive protein and erythrocyte sedimentation rate in orthopaedics. UPOJ. 2002;15:13-6.
40. Paulus H, E., Brahn E. Is erythrocyte sedimentation rate the preferable measure of the acute phase response in rheumatoid arthritis? J Rheumatol. 2004;31(5):838-40.
41. Fries JF. The promise of the future, updated: better outcome tools, greater relevance, more efficient study, lower research costs. Fut Rheumatol. 2006;1(4):415-21.
42. Fries JF, Bruce B, Cella D. The promise of PROMIS: Using item response theory to improve assessment of patient-reported outcomes. Clin Exp Rheumatol. 2005;23(5 SUPPL. 39).

43. World Health Organization. WHO definition of Health [cited March 1, 2014]; Available from: http://www.who.int/about/definition/en/print.html
44. Vermeer M, Kuper HH, van der Bijl AE, Baan H, Posthumus MD, Brus HLM, et al. The provisional ACR/EULAR definition of remission in RA: a comment on the patient global assessment criterion. Rheumatology (Oxford). 2012;51(6):1076-80.
45. Kievit W, Welsing PMJ, Adang EMM, Eijsbouts AM, Krabbe PFM, van Riel PLCM. Comment on the use of self-reporting instruments to assess patients with rheumatoid arthritis: the longitudinal association between the DAS28 and the VAS general health. Arthritis Rheum. 2006;55(5):745-50.
46. Wolfe F. The prognosis of rheumatoid arthritis: assessment of disease activity and disease severity in the clinic. Am J Med. 1997;103(6A):12S-8S.
47. Nestel AR. ESR changes with age - a forgotten pearl. BMJ. 2012;344:e1403.
48. Miller A, Green M, Robinson D. Simple rule for calculating normal erythrocyte sedimentation rate. Br Med J (Clin Res Ed). 1983;286(6361):266.
49. Shearn MA, Kang IY. Effect of age and sex on the erythrocyte sedimentation rate. J Rheumatol. 1986;13(2):297-8.
50. De Silva DA, Woon FP, Chen C, Chang HM, Wong MC. Serum erythrocyte sedimentation rate is higher among ethnic South Asian compared to ethnic Chinese ischemic stroke patients. Is this attributable to metabolic syndrome or central obesity? J Neurol Sci. 2009;276(1-2):126-9.
51. Böttiger LE, Svedberg CA. Normal erythrocyte sedimentation rate and age. Br Med J. 1967;2(5544): 85-7.
52. Radovits BJ, Fransen J, van Riel PL, Laan RF. Influence of age and gender on the 28-joint Disease Activity Score (DAS28) in rheumatoid arthritis. Ann Rheum Dis. 2008;67(8):1127-31.
53. Ranganath VK, Elashoff DA, Khanna D, Park G, Peter JB, Paulus HE. Age adjustment corrects for apparent differences in erythrocyte sedimentation rate and C-reactive protein values at the onset of seropositive rheumatoid arthritis in younger and older patients. J Rheumatol. 2005;32(6):1040-2.
54. Hayes GS, Stinson IN. Erythrocyte sedimentation rate and age. Arch Ophthalmol. 1976;94(6):939-40.
55. Oeser A, Chung CP, Asanuma Y, Avalos I, Stein CM. Obesity is an independent contributor to functional capacity and inflammation in systemic lupus erythematosus. Arthritis Rheum. 2005;52(11):3651-9.
56. Kawamoto R, Kusunoki T, Abe M, Kohara K, Miki T. An association between body mass index and high-sensitivity C-reactive protein concentrations is influenced by age in community-dwelling persons. Ann Clin Biochem. 2013;50(Pt5):457-64.
57. Piéroni L, Bastard JP, Piton A, Khalil L, Hainque B, Jardel C. Interpretation of circulating C-reactive protein levels in adults: body mass index and gender are a must. Diabetes Metab. 2003;29(2 Pt 1): 133-8.
58. Lee YJ, Lee JH, Shin YH, Kim JK, Lee HR, Lee DC. Gender difference and determinants of C-reactive protein level in Korean adults. Clin Chem Lab Med. 2009;47(7):863-9.
59. Lakoski SG, Cushman M, Criqui M, Rundek T, Blumenthal RS, D'Agostino Jr RB, et al. Gender and C-reactive protein: data from the Multiethnic Study of Atherosclerosis (MESA) cohort. Am Heart J. 2006;152(3):593-8.
60. Rommel J, Simpson R, Mounsey JP, Chung E, Schwartz J, Pursell I, et al. Effect of body mass index, physical activity, depression, and educational attainment on high-sensitivity C-reactive protein in patients with atrial fibrillation. Am J Cardiol. 2013;111(2):208-12.
61. Kao TW, Lu IS, Liao KC, Lai HY, Loh CH, Kuo HK. Associations between body mass index and serum levels of C-reactive protein. S Afr Med J. 2009;99(5):326-30.

62. Rawson ES, Freedson PS, Osganian SK, Matthews CE, Reed G, Ockene IS. Body mass index, but not physical activity, is associated with C-reactive protein. Med Sci Sports Exerc. 2003;35(7):1160-6.
63. Reeve BB, Fayers P. Applying item response theory modeling for evaluating questionnaire item and scale properties. In: Fayers P, Hays RD, editors. Assessing Quality of Life in Clinical Trials: Methods of Practice. Oxford, NY: Oxford University Press; 2005. p. 55-73.
64. Belvedere SL, de Morton NA. Application of Rasch analysis in health care is increasing and is applied for variable reasons in mobility instruments. J Clin Epidemiol. 2010;63:1287-97.
65. McHorney CA. Generic health measurement: Past accomplishments and a measurement paradigm for the 21st century. Ann Intern Med. 1997;127(5):743-50.
66. McHorney CA. Ten recommendations for advancing patient-centered outcomes measurement for older persons. Ann Intern Med. 2003;139:403-9.
67. Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. Med Care. 2000;38(9):II28-II42.
68. Vermeer M, Kuper HH, Hoekstra M, Haagsma CJ, Posthumus MD, Brus HLM, et al. Implementation of a treat-to-target strategy in very early rheumatoid arthritis. Arthritis Rheum. 2011;63(10):2865-72.

Chapter 2

Modern psychometrics applied in rheumatology

A systematic review

L. Siemons
P.M. ten Klooster
E. Taal
C.A.W. Glas
M.A.F.J. van de Laar

BMC Musculoskelet Disord. 2012, 13: 216

Abstract

Background: Although item response theory (IRT) appears to be increasingly used within health care research in general, a comprehensive overview of the frequency and characteristics of IRT analyses within the rheumatic field is lacking. An overview of the use and application of IRT in rheumatology to date may give insight into future research directions and highlight new possibilities for the improvement of outcome assessment in rheumatic conditions. Therefore, this study systematically reviewed the application of IRT to patient-reported and clinical outcome measures in rheumatology.

Methods: Literature searches in PubMed, Scopus and Web of Science resulted in 99 original English-language articles which used some form of IRT-based analysis of patient-reported or clinical outcome data in patients with a rheumatic condition. Both general study information and IRT-specific information were assessed.

Results: Most studies used Rasch modeling for developing or evaluating new or existing patient-reported outcomes in rheumatoid arthritis or osteoarthritis patients. Outcomes of principle interest were physical functioning and quality of life. Since the last decade, IRT has also been applied to clinical measures more frequently. IRT was mostly used for evaluating model fit, unidimensionality and differential item functioning, the distribution of items and persons along the underlying scale, and reliability. Less frequently used IRT applications were the evaluation of local independence, the threshold ordering of items, and the measurement precision along the scale.

Conclusion: IRT applications have markedly increased within rheumatology over the past decades. To date, IRT has primarily been applied to patient-reported outcomes, however, applications to clinical measures are gaining interest. Useful IRT applications not yet widely used within rheumatology include the cross-calibration of instrument scores and the development of computerized adaptive tests which may reduce the measurement burden for both the patient and the clinician. Also, the measurement precision of outcome measures along the scale was only evaluated occasionally. Performed IRT analyses should be adequately explained, justified, and reported. A global consensus about uniform guidelines should be reached concerning the minimum number of assumptions which should be met and best ways of testing these assumptions, in order to stimulate the quality appraisal of performed IRT analyses.

Background

Since there is no gold standard for the assessment of disease severity and impact in most rheumatic conditions, it is common practice to administer multiple outcome measures to patients. Initially, the severity and impact of most rheumatic conditions was typically evaluated with clinical measures (CMs) [1, 2] such as laboratory measures of inflammation like the erythrocyte sedimentation rate [3] and physician-based joint counts [4, 5]. Since the eighties of the last century, however, rheumatologists have increasingly started to use patient-reported outcomes (PROs) [1, 2]. As a result, a wide variety of PROs are currently in use, varying from single item visual analog scales (e.g. pain or general health) to multiple item scales like the health assessment questionnaire (HAQ) [6] which measures a patient's functional status and the 36-item short form health survey (SF-36) which measures eight dimensions of health related quality of life [7].

Statistical methods are essential for the development and evaluation of all outcome measures. By far, most health outcome measures have been developed using methods from classical test theory (CTT). In recent years, however, an increase in the use of statistical methods based on item response theory (IRT) can be observed in health status assessment [8-10]. Extensive and detailed descriptions of IRT can be found in the literature [11-14]. In short, IRT is a collection of probabilistic models, describing the relation between a patient's response to a categorical question/item and the underlying construct being measured by the scale [11, 15]. IRT supplements CTT methods, because it provides more detailed information on the item level and on the person level. This enables a more thorough evaluation of an instrument's psychometric characteristics [15], including its measurement range and measurement precision. The evaluation of the contribution of individual items facilitates the identification of the most relevant, precise, and efficient items for the assessment of the construct being measured by the instrument. This is very useful for the development of new instruments, but also for improving existing instruments and developing alternate or short form versions of existing instruments [16]. Additionally, IRT methods are particularly suitable for equating different instruments intended to measure the same construct [17] and for cross-cultural validation purposes [18]. Finally, IRT provides the basis for developing item banks and patient-tailored computerized adaptive tests (CATs) [9, 19, 20].

Although IRT appears to be increasingly used within health care research in general, a comprehensive overview of the frequency and characteristics of IRT analyses within the rheumatic field is lacking. The Outcome Measures in Rheumatology (OMERACT) network recently initiated a special interest group aimed at promoting the use of IRT methods in rheumatology [21]. An overview of the use and application of IRT in rheumatology to date may give insight into future research directions and highlight new possibilities for the improvement of outcome assessment in rheumatic conditions. Therefore, the aim of this study was to systematically review the application of IRT to clinical and patient-reported outcome measures within rheumatology.

Methods

Search strategy

Figure 1 presents an overview of the various stages followed during the search process, starting with an extensive literature search in April 2012 to identify all eligible studies up to and including the year 2011. Electronic database searches of PubMed, Scopus, and Web of Science were carried out, using the terms 'Item response theor*' OR 'Item response model*' OR 'latent trait theor*' OR Rasch OR Mokken, in combination with Rheumat* OR Arthros* OR arthrit*.

Inclusion and exclusion criteria

Only original research articles written in English were included. Articles were considered original when they included original data and when they performed analyses on this data in order to achieve a defined study objective. To be included, studies should present an application of IRT in a sample of which at least 50% had some kind of rheumatic disease. In cases where less than 50% of the study sample consisted of rheumatic patients (i.e. inflammatory rheumatism, arthrosis, soft tissue rheumatism), the study was only included when the rheumatic sample was analysed separately from the rest of the sample. Reviews, letters, editorials, opinion papers, abstracts, posters, and purely descriptive studies were excluded. No limitations were set for study design.

Study identification and selection

The search strategy resulted in a total of 385 studies. After the removal of 189 duplicates, 196 unique articles were identified. Two reviewers independently screened all 196 studies for relevance based on the abstract and title identified from the initial search. If no evident inclusion or exclusion reasons were identified, the full-text was examined. In total, 103 studies did not meet inclusion criteria and were excluded. The main reasons for exclusion were: the study population (i.e. the study population was not clearly defined or the study contained a rheumatic sample <50% of the total sample which was not separately analysed), the statistical analyses (i.e. no IRT application), and the article type (i.e. non-original research). Figure 1 includes an overview of the exclusion reasons followed by the number of articles removed.

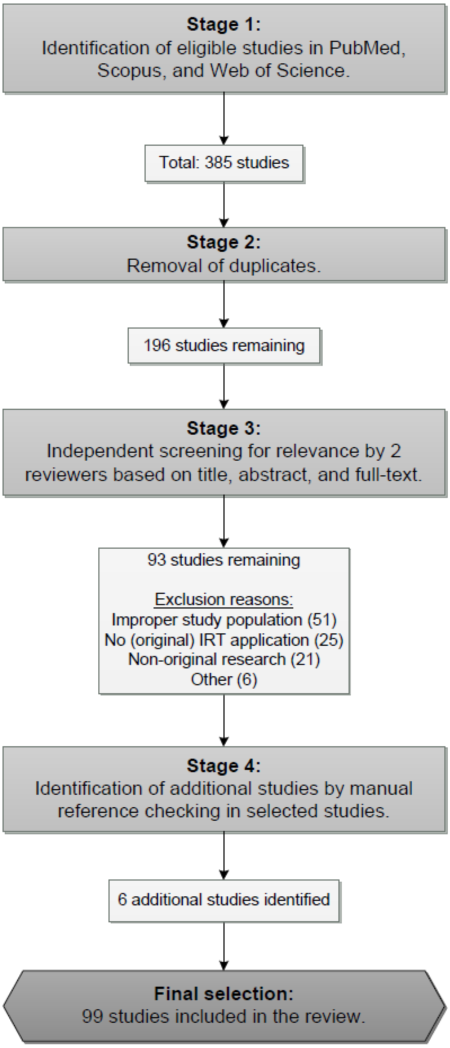

Stage 1:
Identification of eligible studies in PubMed, Scopus, and Web of Science.
Total: 385 studies
Stage 2:
Removal of duplicates.
196 studies remaining
Stage 3:
Independent screening for relevance by 2 reviewers based on title, abstract, and full-text.
93 studies remaining
Exclusion reasons:
Improper study population (51)
No (original) IRT application (25)
Non-original research (21)
Other (6)
Stage 4:
Identification of additional studies by manual reference checking in selected studies.
6 additional studies identified
Final selection:
99 studies included in the review.

Figure 1. Flowchart of the search process.

Data extraction

First, two reviewers independently evaluated a random sample of 15 articles. Both general study information as well as IRT-specific information were extracted, using a purpose-made checklist (Additional file 1) based on both expert input and important issues as mentioned in Tennant and Conaghan [22], Reeve and Fayers [15], and Orlando [23]. Inter-rater agreement of the evaluated variables was moderate to high, with Cohen's kappa ranging from 0.60 to 1.00. Most of the disagreements were caused by differing interpretations of some of the extracted variables. For instance, one of the reviewers interpreted the checklist on “performed analyses” as performed analyses using IRT based methods only, whereas the other reviewer interpreted it more broadly including classical test theory methods as well (the latter being the correct method). Consensus about these differences was reached by discussion. Next, one of these reviewers also evaluated the remaining 84 articles.

General study information

General information concerned the author(s), publication year, study population, the populations’ country of origin, total number of participants (N), study design of the IRT analyses (i.e. cross-sectional or longitudinal), type of outcome measure (PRO or CM), and main measurement intention (e.g. quality of life, pain, overall physical functioning).

Purpose of analyses

The purpose of the analyses was determined by the main goal the author(s) of the article pursued (e.g. the development, evaluation, comparison, or cross-cultural validation of instruments).

Specific IRT analyses

Before a researcher can perform IRT analyses, an appropriate IRT model should be selected. Unidimensional models are most widely applied, the simplest being the Rasch model which assumes that the items are equally discriminating and vary only in their difficulty. The 2-parameter logistic model (2-PL model) extends the Rasch model by assuming that the items have a varying ability to discriminate among people with different levels of the underlying construct [11, 15, 19, 23]. These models can be specified further for polytomous items. The rating scale model, graded response model, modified graded response model, partial credit model, and generalized partial credit model can be applied in case of ordered categorical responses. The nominal response model can be applied when response categories are not necessarily ordered [11, 15, 19, 23, 24]. The rating scale model and the partial credit model are generalizations of the Rasch model, the other models are generalizations of the 2-PL model. In addition to these unidimensional models, multidimensional models and specific non-parametric models like the Mokken model [25, 26] have been developed. Differences in model assumptions should be taken into account when choosing a model and model choice should be motivated by taking the discrimination equality of the items and the number of (ordered) response categories into consideration [15, 22-24].

The applied IRT software and the corresponding item and person parameter estimation method(s) should also be cited, since not all software packages report the findings in the same way [22] and because the use of different estimation methods may result in different parameter estimates [11].

To make IRT results interpretable and trustworthy, three principal assumptions should be evaluated when applying a unidimensional IRT model [15, 23]. The first assumption concerns unidimensionality, meaning that the set of test items measure only a single construct [11, 15, 22, 23]. Analyses for checking the unidimensionality can include different types of factor analysis of the items or the residuals. A more advanced method would be to compare a unidimensional IRT model with a multidimensional IRT model, for instance using a likelihood ratio test. The second (related) assumption concerns local independence of the items. When this assumption is violated this may indicate that the items have more in common with each other than just the single underlying construct [11, 15, 22, 23]. This may either point to response dependency (e.g. overlapping items in the scale) or to multidimensionality of the scale [22]. It can lead to biased parameter estimates and wrong decisions about, for instance, item selection [15]. Local independence can be tested by a factor analysis of the residual covariations, or with more specific statistics targeted at responses to pairs of items [12]. The third assumption concerns the model's appropriateness to reflect the true relationship among the underlying construct and the item responses [11, 15, 22, 23]. This can be examined with both item and person fit statistics. More information about these assumptions and suggestions about which aspects to report can be found in the literature [11, 15, 22, 23].

Other useful IRT applications include the evaluation of the presence of differential item functioning, the reliability and measurement precision, the ordering of the response categories or item thresholds, and the hierarchical ordering and distribution of persons and items along the scale of the underlying construct.

Differential item functioning (DIF, also called item bias) is present when patients with similar levels of the underlying construct being measured respond differently to an item [15, 22]. Commonly examined types of DIF are DIF across gender and age [22].

Global IRT reliability is equivalent to Cronbach's alpha, with the difference that not the raw score but the IRT score is being used in its calculation. Which specific global reliability statistics are presented usually depends on the software package used. Contrary to CTT methods, IRT also provides information about the local reliability [12] and, related to this, the instrument's measurement precision along the scale of the underlying construct.

With rating scale analysis, the ordering of the response categories or item thresholds can be checked, enabling the evaluation of the appropriateness or redundancy of the response categories [15]. Likewise, the hierarchical ordering and/or distribution of persons and items along the scale can be analysed to determine the measurement range of the outcome measure and to determine whether the items function well for the included population sample or whether there is a mismatch between them [23].

Results

General information of included studies

The initial database search yielded a total of 93 eligible studies. Six additional studies were identified by manual reference checks of the selected studies. This resulted in a final selection of 99 studies (Additional file 2). Figure 2 shows that the prevalence of IRT analysis within rheumatology increased markedly over the past decades. This is consistent with conclusions from Hays et al. [19], and with findings from Belvedere and Morton [8] who examined the frequency of Rasch analyses in the development of mobility instruments.

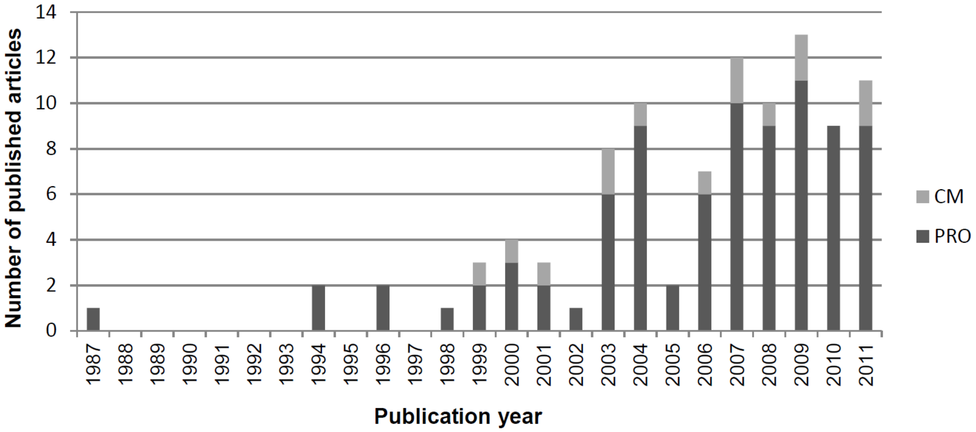


Figure 2. Number of published articles reporting the application of IRT within rheumatology.

Table 1 presents an overview of the most prominent results. By far, most research was carried out with patients from the United States or the United Kingdom, but data from patients from The Netherlands and Canada were also regularly used. The vast majority of studies involved cross-sectional IRT analyses. It could also be observed that an increasing number of studies perform longitudinal IRT analyses since the 21st century, as represented by a rise of DIF testing over time.

Study samples varied from as little as 18 persons in the study of Penta et al. [27] to as many as 16519 persons in the study conducted by Wolfe et al. [28]. Most studies (92.9%) performed analyses on a population sample of at least 50 persons.

In 85 of the 99 studies IRT analyses were applied to PROs. The remaining 14 studies applied IRT to CMs. The vast majority of the studies applied IRT to data gathered from patients suffering from rheumatoid arthritis (RA) or osteoarthritis (OA).

Outcome measures of overall physical functioning and quality of life were most frequently being analysed. To a lesser extent, studies applied IRT to PRO measures of specific functioning [27, 29-37], pain [35, 38-43], psychological constructs [44-46], and work disability [47-51]. Studies also applied IRT to CMs such as measures of disease activity [52-54] and disease damage or radiographic severity [55-57].

Purpose of analyses

Most common main goals for both the PRO- and the CM-studies were the development or evaluation of new measures, the evaluation of existing measures, and the development or evaluation of alternate or short form versions of an existing measure. In addition, several studies aimed to cross-culturally validate a patient-reported or clinical measure. IRT was rarely applied for the development of item banks [17, 58] or computerized adaptive tests [59, 60].

Specific IRT analyses

IRT model and software

The vast majority of IRT applications within rheumatology involved Rasch analyses, although a clear specification and rationale of the applied Rasch model was not always given. Few studies used a two-parameter IRT model or Mokken analysis. Most analyses were carried out with the software packages Bigsteps/Winsteps or RUMM.

A motivation of the model choice was only provided in 27.3% of the studies. Likewise, the item and person parameter estimation methods were rarely specified (8.1% and 4.0% of the studies, respectively).

Table 1 – Overview of the most prominent results.

Variable	PRO-studies		CM-studies		Total-studies	
	n	% *	n	% *	n	% *
Country						
US	26	30.6	4	28.6	30	30.3
UK	24	28.2	2	14.3	26	26.3
The Netherlands	11	12.9	4	28.6	15	15.2
Canada	10	11.8	1	7.1	11	11.1
Other	32	37.6	5	35.7	37	37.4
Design						
Cross-sectional	76	89.4	14	100.0	90	90.9
Longitudinal	13	15.3	2	14.3	15	15.2
Disease condition						
RA	43	50.6	5	35.7	48	48.5
OA	31	36.5	3	21.4	34	34.3
Other	31	36.5	7	50.0	38	38.4
Measurement intention						
Overall physical functioning	33	38.8	2	14.3	35	35.4
Quality of life	26	30.6	2	14.3	28	28.3
Specific functioning	10	11.8	0	0.0	10	10.1
Pain	7	8.2	0	0.0	7	7.1
Psychological constructs	3	3.5	0	0.0	3	3.0
Work disability	5	5.9	0	0.0	5	5.0
Disease activity	0	0.0	3	21.4	3	3.0
Disease damage or radiographic severity	0	0.0	3	21.4	3	3.0
Other	11	12.9	4	28.6	15	15.2
Main goal						
Development/evaluation new measures	25	29.4	2	14.3	27	31.4
Evaluation existing measures	31	36.5	6	42.9	37	37.4
Development/evaluation alternate/short form	11	12.9	2	14.3	13	13.1
Development item bank or CAT	4	4.7	0	0.0	4	4.0
Cross-cultural validation	7	8.2	2	14.3	9	9.1
Other	11	12.9	3	21.4	14	14.1
Software						
Bigsteps/Winsteps	28	32.9	3	21.4	31	31.3
RUMM	29	34.1	6	42.9	35	35.4
Other / not specified	29	34.1	5	35.7	34	34.3
IRT model						
Rasch	72	84.7	12	85.7	84	84.8
2-PLM	13	15.3	1	7.1	14	14.1
Mokken	3	3.5	1	7.1	4	4.0
IRT analyses						
Unidimensionality	65	76.5	10	71.4	75	75.8
Local independence	16	18.8	1	7.1	17	17.2
Appropriateness model (fit analyses)	77	90.6	13	92.9	90	90.9
DIF	50	58.8	6	42.9	56	56.6
Person/item separation/reliability	52	61.2	10	71.4	62	62.6
Hierarchical ordering/distribution of items/persons	57	67.1	9	64.3	66	66.7
Rating scale analysis	30	35.3	7	50.0	37	37.4
Measurement precision of the scale	10	11.8	1	7.1	11	11.1

**Note that some studies can be assigned to multiple subcategories, therefore, the sum of the percentages within a category exceeds 100%. PRO: patient-reported outcome (N=85), CM: clinical measure (N=14), RA: rheumatoid arthritis, OA: osteoarthritis, CAT: computerized adaptive test, IRT: item response theory, 2-PLM: 2 parameter logistic model, DIF: differential item functioning*

IRT assumptions

The assumption of unidimensionality was tested in approximately three quarters of the studies. Methods used for this purpose mainly concerned some type of factor analysis (confirmatory/exploratory factor analysis or principal component analysis) or the examination of specific IRT statistics (e.g. whether the overall model fit or the item fit values were larger than a pre-specified cut-off point). No studies were found where unidimensional IRT models were contrasted with multidimensional IRT models.

A possible violation of the assumption of local independence was evaluated in only one of the CM studies, and in only 18.8% of the studies concerning a PRO. Evaluation of the studies also indicated there was no clear agreement on how to evaluate this assumption, given the variety of methods used.

The assumption of the appropriateness of the model was evaluated by approximately 91% of the studies. When applied, roughly half of the cases evaluated overall fit (PRO: 51.9%, CM: 53.8%), almost all evaluated item fit (PRO: 97.4%, CM: 100.0%), but a much smaller percentage evaluated person fit statistics (PRO: 33.8%, CM: 30.8%).

Additional IRT analyses

More than half of the studies used IRT to examine DIF. When applied, analyses varied from cross-sectional DIF across gender (PRO: 80.0%, CM: 66.7%), age (PRO: 76.0%, CM: 66.7%), disease duration (PRO: 36.0%, CM: 16.7), countries/cultures/ethnicity (PRO: 18.0%, CM: 16.7%), and disease type (PRO: 10.0%, CM: 16.7%), to longitudinal DIF analyses over time (PRO: 28.0%, CM: 33.3%).

Other commonly performed IRT analyses included analyses of the global reliability, the hierarchical ordering and distribution of items and persons, and rating scale analyses (i.e. the ordering of the response categories or item thresholds). In addition, a small number of PRO-studies reported IRT analyses regarding the measurement precision of the scale, whereas only 1 of the CM studies evaluated this.

Discussion

IRT offers a powerful framework for the evaluation or development of existing and new outcome measures. This is the first study that systematically reviewed the extent to which IRT has been applied to measurements from rheumatology. Results showed a marked increase in IRT applications within the rheumatic field from the late eighties up to now. Even though most research focussed on PROs, IRT also appeared to be useful for application to CMs. Some opportunities for further IRT applications and improvements in the analyses and reporting of IRT studies were also pointed out.

IRT can be applied for various purposes. First, IRT analysis is useful for the development and evaluation of new measures [22]. For instance, Helliwell et al. [32] developed a foot impact scale to assess foot status in RA patients. Rasch modeling was used to facilitate item reduction by selecting items which were free of DIF and fitted model expectations. Where the CTT methods often discard items at the extremes of the measurement range because too few patients answer them affirmatively, IRT includes these items since they provide important information at the extremes of the measurement range [61].

IRT is also suitable for the evaluation of existing (ordinal) outcome measures. For example, when evaluating an instrument's included response categories it can be determined whether they perform as intended or whether categories should be collapsed into fewer options or expanded into more options [22]. Furthermore, it can be evaluated whether the items in the outcome measure form a unidimensional scale as expected or whether item deletion is necessary [22].

Another favourable feature of IRT is that it is expressed at the item level instead of test level as in CTT [11]. By evaluating the performance of individual items, alternate or short form versions of existing measures can be developed. For example, Wolfe et al. [62] developed an alternate version of the HAQ [6, 63], known as the HAQ-II, specifically targeted at patients with a relatively high physical functioning.

Another commonly used feature of modeling at the item level is the robust assessment of DIF, as reflected in the high proportion of performed DIF analyses. Nevertheless, the full potential of modeling at the item level is not yet being used, given the low percentage of studies evaluating the items' performance (i.e. measurement precision and local reliability) along the scale.

When comparing the studies focusing on RA patients with those focusing on OA patients, the measurement intensions of the analysed instruments and the applied IRT models were highly comparable. However, a notable difference was found in the main goals of these studies. Where the RA studies pursued widely varying main goals, including the development of new instruments, the evaluation of existing instruments, the comparison of different instruments, and cross-cultural validation, the studies on OA patients generally focused on the evaluation of existing instruments only.

There are several IRT applications which have not yet been (frequently) used within rheumatology. One IRT application which appears to be still in its infancy within rheumatology, but which is likely to gain importance in the future, is the development of computerized adaptive tests (CATs) [2]. When testing by means of a CAT, every patient receives a test which is tailored (adapted) to his or her level on the underlying construct being measured. Consequently, each patient can be administered different sequences and numbers of items, drawn from a large item bank. By applying CATs, tests can be shortened without any loss of measurement precision, reducing measurement burden for both the patient and the rheumatologist [1, 2, 9-11, 16].

The potential advantages of cross-calibration is another IRT application which has not yet been recognized within rheumatology. As opposed to CTT methods, the item responses are regressed on separate item and person parameters in IRT [11]. This means that the definition of item parameters is independent of the sample receiving the test and the definition of person parameters is independent of the test items given. This separation of parameters facilitates the cross-calibration of various outcome measures based on the same underlying construct [11, 64], making their scores comparable with each other.

As discussed earlier, it is important to test the assumptions of unidimensionality, local independence, and model appropriateness when analysing data by means of IRT methods. Items which violate one or more of these assumptions should be combined, rephrased, or deleted [22, 23], since they complicate the interpretation of model outcomes. A promising observation was that the majority of the studies tested the assumption of unidimensionality and the appropriateness of the IRT model, albeit some studies did not report any fit statistics. Although comparisons between unidimensional and multidimensional IRT models provide a much more rigorous test of unidimensionality than factor analyses, such comparisons were not made. Analyses of model fit mainly involved overall fit statistics or item fit statistics, and to a lesser extent the evaluation of person fit. Person fit, however, is also important since deviant response patterns of patients may seriously affect the item fit. The removal of patients with such response patterns from the analysis may improve the scale's internal construct validity significantly [22]. Most studies, however, did not check the assumption of local independence. The importance of local independence has only more recently been recognized and, consequently, only some of the most recent studies (from the year 2007) did evaluate this assumption. Future studies should continue to pay attention to this assumption, since locally dependent items could cause parameter estimates to be biased, which may lead to wrong decisions concerning item selection when constructing a certain outcome measure [15].

The results also showed room for improvement in the reporting of made choices and the rationale for specific decisions. For instance, the applied IRT model is often not specified and, if specified, the reasons behind the selected IRT model and used estimation methods are often not clearly motivated. This complicates the quality appraisal and replication of performed analyses.

Where Belvedere and de Morton [8] examined the application of Rasch analysis only, this study included the whole spectrum of IRT models. A notable finding of this review was that the Rasch models dominate within rheumatology, and that two-parameter IRT models were applied in only a few studies. This may be due to the ease of use of a Rasch model and the easiness with which its results can be interpreted. However, this advantage of Rasch modeling comes with the strict assumption that every item of the measure is equally discriminative. Whether this assumption is appropriate can be tested by comparing the Rasch model fit with the 2-parameter model fit. Since the studies of Pham et al. [65] and Siemons et al. [54] are the only studies in which such a comparison was made, this is a point of interest for future studies.

Although IRT is becoming increasingly popular in health status assessment, IRT is quite complex to understand and is not yet a main-stream technique for most researchers and rheumatologists. To increase common understanding and to improve the interpretation of outcomes resulting from the performed IRT analyses, (bio)statisticians, rheumatologists, and researchers should closely collaborate. Clear guidelines on the quality appraisal of performed IRT analyses might increase the use and understanding of IRT in rheumatology even further. Currently, there are no clear guidelines available for rating the methodological quality of the performed IRT analyses. Although standardized tools like the COSMIN (COnsensusbased Standards for the selection of health status Measurement INstruments) checklist [66] can be used for evaluating the methodological quality of studies on measurement properties, this checklist only contains a few questions regarding IRT analyses and is, therefore, more suitable for analyzing the quality of performed classical test theory analyses. Even though the quality checklist used in this study was based on both expert input and important issues from the literature, it was not exhaustive and, consequently, it might have some limitations. For example, when the sample size was considered, only the absolute number was reported. It was not checked whether the authors also justified the sample size for the analyses they wanted to perform. The varying sample size of the analysed patient groups which was found between studies, might be due to the absence of clear guidelines regarding sample size requirements. It is argued that the most simple Rasch analyses already require a minimum size of 50-100 persons [15, 23]. However, many issues are involved in determining the right sample size for a certain study, including the model choice, the number of response categories, and the purpose of the study [15, 23]. These issues should be carefully considered to determine the sample size which is minimally needed to achieve reliable model estimates. Consensus and clear guidelines on quality aspects concerning IRT analyses might guide the choice of an adequate sample size and might also stimulate the development of uniform guidelines for performing and reporting IRT studies, and the development of a checklist for evaluating the quality of the performed and reported IRT analyses.

The formulation of such guidelines will provide a strong foundation to future IRT studies. Tennant et al. already provided such guidelines for performing Rasch analyses [22]. However, given the large diversity of approaches, models, and software used in the field of IRT it is difficult to recommend a single set of guidelines for all types of studies, and an expansion or modification of their guidelines might be needed. In order to get sufficient support for these guidelines it is important to first attempt to reach a more global consensus about recommendations. This article could provide input for such attempts and the COSMIN checklist [66] can serve as an example of how such an international approach can lead to the development of a consensus-based checklist. Agreement should be reached on the minimum number of assumptions which should be met (e.g. unidimensionality, model fit, and DIF analysis) and best ways of testing these assumptions. Additionally, this review showed that IRT methods are rarely being applied for the evaluation of an instrument’s local reliability and measurement precision along the scale of the underlying construct and the construction of item banks and CATs, all unique features of IRT. Therefore, it is recommended that more emphasis will be placed on these features in the guidelines and in future studies.

Conclusions

A marked increase of IRT applications could be observed within rheumatology. IRT has primarily been applied to patient-reported outcomes, but it also appeared to be a useful technique for the evaluation of clinical measures. To date, IRT has mainly been used for the development of new static outcome measures and the evaluation of existing measures. In addition, alternate or short forms were created by evaluating the fit and performance of individual items. Useful IRT applications which are not yet widely used within rheumatology include the cross-calibration of instrument scores and the development of computerized adaptive tests which may reduce the measurement burden for both the patient and the clinician. Also, the measurement precision of outcome measures along the scale has only been evaluated occasionally. The fact that IRT has not yet experienced the same level of standardization and consensus on methodology as CTT methods stresses the importance to adequately explain, justify, and report performed IRT analyses. A global consensus on uniform guidelines should be reached about the minimum number of assumptions which should be met and best ways of testing these assumptions, in order to stimulate the quality appraisal of performed and reported IRT analyses.

References

1. Fries JF. The promise of the future, updated: better outcome tools, greater relevance, more efficient study, lower research costs. Fut Rheumatol. 2006;1(4):415-21.
2. Fries JF, Bruce B, Cella D. The promise of PROMIS: Using item response theory to improve assessment of patient-reported outcomes. Clin Exp Rheumatol. 2005;23(39):S53-S7.
3. Van Riel PLCM, Fransen J, Scott DL. Eular handbook of clinical assessments in rheumatoid arthritis. Alphen aan den Rijn: van Zuiden Communications; 2004.
4. Pala O, Cavaliere LF. Joint counts. In: Bartlett SJ, editor. Clinical care in the rheumatic diseases. Atlanta (GA): Association of Rheumatology Health Professionals; 2006. p. 39-41.
5. Scott DL, Houssien DA. Joint assessment in rheumatoid arthritis. Br J Rheumatol. 1996;35:14-8.
6. Fries JF, Spitz P, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. Arthritis Rheum. 1980;23(2):137-45.
7. Ware JE, Sherbourne CD. The MOS 36-item short form health survey (SF-36): I. Conceptual framework and item selection. . Med Care. 1992;30(6):473-83.
8. Belvedere SL, de Morton NA. Application of Rasch analysis in health care is increasing and is applied for variable reasons in mobility instruments. J Clin Epidemiol. 2010;63:1287-97.
9. McHorney CA. Generic health measurement: Past accomplishments and a measurement paradigm for the 21st century. Ann Intern Med. 1997;127(5):743-50.
10. McHorney CA. Ten recommendations for advancing patient-centered outcomes measurement for older persons. Ann Intern Med. 2003;139:403-9.
11. Hambleton RK, Swaminathan H, Rogers HJ. Fundamentals of item response theory. Newbury Park (CA): Sage Publications; 1991.
12. Scheerens J, Glas CAW, Thomas SM. Educational evaluation, assessment, and monitoring. A systematic approach. Lisse: Swets & Zeitlinger; 2003.
13. Baker FB, Kim S-H. Item response theory. Parameter estimation techniques. New York: Marcel Dekker; 2004.
14. Baker FB. The basics of item response theory. College Park (MD): ERIC Clearinghouse on Assessment and Evaluation; 2001.
15. Reeve BB, Fayers P. Applying item response theory modeling for evaluating questionnaire item and scale properties. In: Fayers P, Hays RD, editors. Assessing Quality of Life in Clinical Trials: Methods of Practice. Oxford, NY: Oxford University Press; 2005. p. 55-73.
16. Fries JF, Bruce B, Bjorner J, Rose M. More relevant, precise, and efficient items for assessment of physical function and disability: moving beyond the classic instruments. Ann Rheum Dis. 2006;65(Suppl III):iii16–iii21.
17. McHorney CA, Cohen AS. Equating health status measures with item response theory: illustrations with functional status items. Med Care. 2000;38(9):II-43-II-59.
18. Tennant A, Penta M, Tesio L, Grimby G, Thonnard JL, Slade A, et al. Assessing and adjusting for cross-cultural validity of impairment and activity limitation scales through differential item functioning within the framework of the Rasch model: the PRO-ESOR project. Med Care. 2004;42:I37-48.
19. Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. Med Care. 2000;38(9):II28-II42.
20. Revicki DA, Cella DF. Health status assessment for the twenty-first century: item response theory, item banking and computer adaptive testing. Qual Life Res. 1997;6(595-600).

21. Tugwell P, Boers M, Brooks M, Simon L, Strand V, Idzerda L. OMERACT: An international initiative to improve outcome measurement in rheumatology. Trials. 2007;8(38).
22. Tennant A, Conaghan PG. The rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a rasch paper? Arthritis Rheum. 2007;57:1358-62.
23. Orlando M. Critical issues to address when applying item response theory (IRT) models. Conference on Improving Health Outcomes Assessment Based on Modern Measurement Theory and Computerized Adaptive Testing. Bethesda, MD: Hyatt; 2004.
24. Embretson SE, Reise SP. Item response theory for psychologists. Mahwah, NJ: Lawrence Erlbaum Associates; 2000.
25. Mokken RJ. A theory and procedure of scale analysis with applications in political research. The Hague: Mouton; 1971.
26. Sijtsma K, Molenaar IW. Introduction to nonparametric item response theory. Thousand Oaks, CA: Sage; 2002.
27. Penta M, Thonnard JL, Tesio L. ABILHAND: a Rasch-built measure of manual ability. Arch Phys Med Rehabil. 1998;79(9):1038-42.
28. Wolfe F, Michaud K, Kahler K, Omar M. The Short Arthritis Assessment Scale: a brief assessment questionnaire for rapid evaluation of arthritis severity in research and clinical practice. J Rheumatol. 2004;31(12):2472-9.
29. Budiman-Mak E, Conrad K, Stuck R, Matters M. Theoretical model and Rasch analysis to develop a revised Foot Function Index. Foot Ankle Int. 2006;27(7):519-27.
30. Conaghan PG, Emerton M, Tennant A. Internal construct validity of the Oxford knee scale: Evidence from Rasch measurement. Arthritis Care Res. 2007;57(8):1363-7.
31. Durez P, Fraselle V, Houssiau F, Thonnard JL, Nielens H, Penta M. Validation of the ABILHAND questionnaire as a measure of manual ability in patients with rheumatoid arthritis. Ann Rheum Dis. 2007;66(8):1098-105.
32. Helliwell P, Reay N, Gilworth G, Redmond A, Slade A, Tennant A, et al. Development of a foot impact scale for rheumatoid arthritis. Arthritis Rheum. 2005;53(3):418-22.
33. Paulsen T, Grotle M, Garratt A, Kjeken I. Development and psychometric testing of the patient-reported measure of activity performance of the hand (MAP-Hand) in rheumatoid arthritis. J Rehabil Med. 2010;42(7):636-44.
34. Vanthuyne M, Smith V, Arat S, Westhovens R, Keyser FD, Houssiau FA, et al. Validation of a manual ability questionnaire in patients with systemic sclerosis. Arthritis Care Res. 2009;61(5):695-703.
35. Haugen IK, Moe RH, Slatkowsky-Christensen B, Kvien TK, van der Heijde D, Garratt A. The AUSCAN subscales, AIMS-2 hand/finger subscale, and FIOHA were not unidimensional scales. J Clin Epidemiol. 2011;64(9):1039-46.
36. Ko Y, Lo N-N, Yeo S-J, Yang K-Y, Yeo W, Chong H-C, et al. Rasch analysis of the Oxfort Knee Score. Osteoarthr Cartilage. 2009;17:1163-9.
37. Woodburn J, Vliet Vlieland TP, van der Leeden M, Steultjens MP. Rasch analysis of Dutch-translated version of the Foot Impact Scale for rheumatoid arthritis. Rheumatology. 2011;50(7):1315-9.
38. Boeckstyns MEH. Development and construct validity of a knee pain questionnaire. Pain. 1987;31(1):47-52.
39. Kersten P, White PJ, Tennant A. The Visual Analogue WOMAC 3.0 scale - internal validity and responsiveness of the VAS version. BMC Musculoskelet Disord. 2010;11(80).

40. O'Malley KJ, Suarez-Almazor M, Aniol J, Richardson P, Kuykendall DH, Moseley JB, Jr., et al. Joint-specific multidimensional assessment of pain (J-MAP): factor structure, reliability, validity, and responsiveness in patients with knee osteoarthritis. J Rheumatol. 2003;30(3):534-43.
41. Roorda LD, Jones CA, Waltz M, Lankhorst GJ, Bouter LM, van der Eijken JW, et al. Satisfactory cross cultural equivalence of the Dutch WOMAC in patients with hip osteoarthritis waiting for arthroplasty. Ann Rheum Dis. 2004;63(1):36-42.
42. 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(2):369-78.
43. Davis AM, Badley EM, Beaton DE, Kopec J, Wright JG, Young NL, et al. Rasch analysis of the Western Ontario McMaster (WOMAC) Osteoarthritis Index: results from community and arthroplasty samples. J Clin Epidemiol. 2003;56:1076-83.
44. Cieza A, Hilfiker R, Boonen A, Chatterji S, Kostanjsek N, Ustun BT, et al. Items from patient-oriented instruments can be integrated into interval scales to operationalize categories of the International Classification of Functioning, Disability and Health. J Clin Epidemiol. 2009;62(9):912-21.
45. Covic T, Pallant JF, Conaghan PG, Tennant A. A longitudinal evaluation of the Center for Epidemiologic Studies-Depression scale (CES-D) in a rheumatoid arthritis population using Rasch analysis. Health Qual Life Outcomes. 2007;5(41).
46. Covic T, Pallant JF, Tennant A, Cox S, Emery P, Conaghan PG. Variability in depression prevalence in early rheumatoid arthritis: a comparison of the CES-D and HAD-D Scales. BMC Musculoskelet Disord. 2009;10(18).
47. Gilworth G, Chamberlain MA, Harvey A, Woodhouse A, Smith J, Smyth MG, et al. Development of a work instability scale for rheumatoid arthritis. Arthritis Rheum. 2003;49(3):349-54.
48. Gilworth G, Emery P, Barkham N, Smyth MG, Helliwell P, Tennant A. Reducing work disability in Ankylosing Spondylitis: development of a work instability scale for AS. BMC Musculoskelet Disord. 2009;10(68).
49. Gilworth G, Emery P, Gossec L, Vliet Vlieland TP, Breedveld FC, Hueber AJ, et al. Adaptation and cross-cultural validation of the rheumatoid arthritis work instability scale (RA-WIS). Ann Rheum Dis. 2009;68(11):1686-90.
50. Tang K, Beaton DE, Lacaille D, Gignac MAM, Zhang W, Anis AH, et al. The Work Instability Scale for Rheumatoid Arthritis (RA-WIS): Does it work in osteoarthritis? Qual Life Res. 2010;19(7):1057-68.
51. Tang K. Disease-related differential item functioning in the work instability scale for rheumatoid arthritis: converging results from three methods. Arthritis Care Res. 2011;63(8):1159-69.
52. Bode RK, Klein-Gitelman MS, Miller ML, Lechman TS, Pachman LM. Disease activity score for children with juvenile dermatomyositis: Reliability and validity evidence. Arthritis Rheum. 2003;49(1):7-15.
53. Lawton G, Bhakta BB, Chamberlain MA, Tennant A. The Behcet's disease activity index. Rheumatology. 2004;43(1):73-8.
54. Siemons L, ten Klooster PM, Taal E, Kuper IH, van Riel P, van de Laar M, et al. Validating the 28-Tender Joint Count Using Item Response Theory. J Rheumatol. 2011;38(12):2557-64.
55. Brunner HI, Feldman BM, Urowitz MB, Gladman DD. Item weightings for the Systemic Lupus International Collaborating Clinics/American College of Rheumatology Disease Damage Index using Rasch analysis do not lead to an important improvement. J Rheumatol. 2003;30(2):292-7.
56. Wolfe F, van der Heijde DM, Larsen A. Assessing radiographic status of rheumatoid arthritis: introduction of a short erosion scale. J Rheumatol. 2000;27(9):2090-9.

57. Conaghan PG, Tennant A, Peterfy CG, Woodworth T, Stevens R, Guermazi A, et al. Examining a whole-organ magnetic resonance imaging scoring system for osteoarthritis of the knee using Rasch analysis. Osteoarthritis Cartilage. 2006;14 Suppl A:A116-21.
58. Kopec JA, Sayre EC, Davis AM, Badley EM, Abrahamowicz M, Sherlock L, et al. Assessment of health-related quality of life in arthritis: conceptualization and development of five item banks using item response theory. Health Qual Life Outcomes. 2006;4(33).
59. Jette AM, McDonough CM, Haley SM, Ni PS, Olarsch S, Latham N, et al. A computer-adaptive disability instrument for lower extremity osteoarthritis research demonstrated promising breadth, precision, and reliability. J Clin Epidemiol. 2009;62(8):807-15.
60. Kosinski M, Bjorner JB, Ware JE, Sullivan E, Straus WL. An evaluation of a patient-reported outcomes found computerized adaptive testing was efficient in assessing osteoarthritis impact. J Clin Epidemiol. 2006;59(7):715-23.
61. Tennant A, P. MS, Hagell P. Application of Rasch analysis in the development and application of quality of life instruments. Value Health. 2004;7(1):S22-S6.
62. Wolfe F, Michaud K, Pincus T. Development and validation of the health assessment questionnaire II: a revised version of the health assessment questionnaire. Arthritis Rheum. 2004;50(10):3296-305.
63. Fries JF, Spitz PW, Young DY. The dimensions of health outcomes: The Health Assessment Questionnaire, disability and pain scales. J Rheumatol. 1982;9:789-93.
64. Dorans NJ. Linking scores from multiple health outcome instruments. Qual Life Res. 2007;16:85-94.
65. Pham T, van der Heijde DM, Pouchot J, Guillemin F. Development and validation of the French ASQoL questionnaire. Clin Exp Rheumatol. 2010;28(3):379-85.
66. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Qual Life Res. 2010;19:539-49.

Additional file 1

Checklist

List of abbreviations

1-PLM	1-parameter logistic model
2-PLM	2-parameter logistic model
3-PLM	3-parameter logistic model
ADL	Activities of daily living
CAT	Computerized adaptive testing
CFA	Confirmatory factor analysis
CML	Conditional maximum likelihood
DIF	Differential item functioning
EAP	Expected a posteriori estimator
EFA	Exploratory factor analysis
GPCM	Generalized partial credit model
GRM	Graded response model
IADL	Instrumental activities of daily living
IRT	Item response theory
MML	Marginal maximum likelihood
PCA	Principal component analysis
PCM	Partial credit model
PRO	Patient reported outcome
QoL	Quality of life
RSM	Rating scale model
UML	Unconditional maximum likelihood
(W)ML	(Weighted) maximum likelihood

General information

First author ……………………………………………………………………………

Publication year □□□□

Country / Countries of study execution ……………………………………………………………………………..

……………………………………………………………………………..

Design IRT analyses

□ Cross-sectional

□ Longitudinal

Disease condition(s) ……………………………………………………………………………

Total N ……………………………………………………………………………

Application

□ PRO measure

□ Clinical measure / other

Instrument

Instrument examined ……………………………………………………………………………...........

Main measurement intention

□ QoL / Health Status

□ Overall physical function (ADL, IADL, Mobility tasks)

□ Specific functioning (e.g. hand function)

□ Disease activity

□ Pain

□ Psychosocial construct (e.g. anxiety, depression)

□ Other: ………………………………………………………………….

Main goal of the study

□ Development/evaluation of new instrument

□ Development/evaluation of alternate / short form version

□ Evaluation of existing instrument

□ Comparison of instruments

□ Comparison of the psychometric properties of a single instrument in various patient groups

□ Scoring of instrument

□ Cross-calibration / equation of instruments

□ Calibration / evaluation of item bank

□ Development / evaluation of CAT

□ Cross-cultural validation

□ Other: ………………………………………………………………….

Applied IRT model

General **Specific**

□ 1-PLM / Rasch ➔ □ Dichotomous

- □ Polytomous
  - □ RSM
  - □ PCM
  - □ Other: ………………………………………………….....
  - □ Not specified

□ 2-PLM ➔ □ Dichotomous

- □ Polytomous
  - □ GRM
  - □ GPCM
  - □ Other: ………………………………………………….....
  - □ Not specified

□ 3-PLM

□ Mokken

□ Not specified

Applied IRT software

□ Bigsteps / Winsteps
□ RUMM
□ Multilog / Bilog
□ Parscale
□ MPlus
□ MSP
□ SAS
□ GLAMM / STATA
□ Conquest
□ Other: …………………………………………………………..
□ Not specified

Performed analyses

□ Unidimensionality

□ Local independence

□ Item fit

□ Person fit

□Person/item/subscale separation and/or reliability

□ Measurement precision (e.g. item/test information)

□ DIF

□ Rating scale analysis (response category ordering/item thresholds)

□ Hierarchical ordering and/or distribution of persons/items

□ Cross-calibration / equation

□ Other: ………………………………………………………………………………

Quality appraisal

Sample size ………………………..

Number of items ………………………...

Model description

□ Yes, with model specification (e.g. Rasch-PCM, Rasch dichotomous)

□ Yes, without model specification

□ No

Model choice adequately explained (e.g., number or type of response categories / common discrimination parameter)

□ Yes

□ No

Applied IRT software is cited

□ Yes

□ No

Various IRT models were tested

□ Yes

□ No

Assumptions adequately tested

□ Unidimensionality ➔

□ IRT residuals

□ EFA / PCA of item scores

□ CFA of item scores

□ PCA of residuals

□ IRT statistics (e.g. model fit)

□ Other: ………………………………

	□ Local independence ➔	□ IRT residuals
		□ Residual covariation CFA (Modification indices)
		□ Residual covariation PCA / EFA
		□ Other: ………………………………
	□ Fit ➔	□ Overall fit statistics
		□ Item fit statistics
		□ Person fit statistics
	□ DIF ➔	□ Age
		□ Gender
		□ Disease duration
		□ Countries / cultures
		□ Diseases
		□ Time points
		□ Other: ………………………………
Item parameter estimation method	□ JML	
	□ UML	
	□ MML	
	□ CML	
	□ Bayesian	
	□ Not specified	
Person/population parameter estimation method	□ (W)ML	
	□ Bayesian (e.g. EAP)	
	□ Not specified	

Important flaws:

………

………

Additional file 2

List of included articles

1. Ayis S, Dieppe P. The natural history of disability and its determinants in adults with lower limb musculoskeletal pain. J Rheumatol. 2009;36(3):583-91.
2. Bode RK, Klein-Gitelman MS, Miller ML, Lechman TS, Pachman LM. Disease activity score for children with juvenile dermatomyositis: Reliability and validity evidence. Arthritis Rheum. 2003;49(1):7-15.
3. Boeckstyns MEH. Development and construct validity of a knee pain questionnaire. Pain. 1987;31(1):47-52.
4. Brunner HI, Feldman BM, Urowitz MB, Gladman DD. Item weightings for the Systemic Lupus International Collaborating Clinics/American College of Rheumatology Disease Damage Index using Rasch analysis do not lead to an important improvement. J Rheumatol. 2003;30(2):292-7.
5. Budiman-Mak E, Conrad K, Stuck R, Matters M. Theoretical model and Rasch analysis to develop a revised Foot Function Index. Foot Ankle Int. 2006;27(7):519-27.
6. Chiou CF, Sherbourne CD, Cornelio I, Lubeck DP, Paulus HE, Dylan M, et al. Development and validation of the revised Cedars-Sinai health-related quality of life for rheumatoid arthritis instrument. Arthritis Rheum. 2006;55(6):856-63.
7. Chiou CF, Sherbourne CD, Ofman J, Lee M, Lubeck DP, Paulus HE, et al. Development and validation of Cedars-Sinai Health-Related Quality of Life in Rheumatoid Arthritis (CSHQ-RA) short form instrument. Arthritis Rheum. 2004;51(3):358-64.
8. Chiou CF, Suarez-Almazor ME, Sherbourne CD, Chang CH, Reyes C, Dylan M, et al. Development and validation of a preference weight multiattribute health outcome measure for rheumatoid arthritis. J Rheumatol. 2006;33(12):2409-11.
9. Chogle AR, Mistry KJ, Deo SS. Comparison of the Indian version of Health Assessment Questionnaire Score and Short Form 36 Physical Function Score in rheumatoid arthritis using Rasch analysis. Indian J Rheumatol. 2008;3(2):52-7.
10. Cieza A, Hilfiker R, Boonen A, Chatterji S, Kostanjsek N, Ustun BT, et al. Items from patient-oriented instruments can be integrated into interval scales to operationalize categories of the International Classification of Functioning, Disability and Health. J Clin Epidemiol. 2009;62(9):912-21.
11. Cieza A, Hilfiker R, Boonen A, van der Heijde D, Braun J, Stucki G. Towards an ICF-based clinical measure of functioning in people with ankylosing spondylitis: A methodological exploration. Disabil Rehabil. 2009;31(7):528-37.
12. Conaghan PG, Emerton M, Tennant A. Internal construct validity of the Oxford knee scale: Evidence from Rasch measurement. Arthritis Care Res. 2007;57(8):1363-7.
13. Conaghan PG, Tennant A, Peterfy CG, Woodworth T, Stevens R, Guermazi A, et al. Examining a whole-organ magnetic resonance imaging scoring system for osteoarthritis of the knee using Rasch analysis. Osteoarthritis Cartilage. 2006;14 Suppl A:A116-21.
14. Covic T, Pallant JF, Conaghan PG, Tennant A. A longitudinal evaluation of the Center for Epidemiologic Studies-Depression scale (CES-D) in a rheumatoid arthritis population using Rasch analysis. Health Qual Life Outcomes. 2007;5:41.
15. Covic T, Pallant JF, Tennant A, Cox S, Emery P, Conaghan PG. Variability in depression prevalence in early rheumatoid arthritis: a comparison of the CES-D and HAD-D Scales. BMC Musculoskelet Disord. 2009;10:18.

16. Davis AM, Badley EM, Beaton DE, Kopec J, Wright JG, Young NL, et al. Rasch analysis of the Western Ontario McMaster (WOMAC) Osteoarthritis Index: results from community and arthroplasty samples. J Clin Epidemiol. 2003;56:1076-83.
17. Davis AM, Perruccio AV, Canizares M, Tennant A, Hawker GA, Conaghan PG, et al. The development of a short measure of physical function for hip OA HOOS-Physical Function Shortform (HOOS-PS): an OARSI/OMERACT initiative. Osteoarthritis Cartilage. 2008;16(5):551-9.
18. Doward LC, McKenna SP, Meads DM, Twiss J, Revicki D, Wong RL, et al. Translation and validation of non-English versions of the Ankylosing Spondylitis Quality of Life (ASQOL) questionnaire. Health Qual Life Outcomes. 2007;5.
19. Doward LC, McKenna SP, Whalley D, Tennant A, Griffiths B, Emery P, et al. The development of the L-QoL: a quality-of-life instrument specific to systemic lupus erythematosus. Ann Rheum Dis. 2009;68(2):196-200.
20. Doward LC, Spoorenberg A, Cook SA, Whalley D, Helliwell PS, Kay LJ, et al. Development of the ASQoL: A quality of life instrument specific to ankylosing spondylitis. Ann Rheum Dis. 2003;62(1):20-6.
21. Durez P, Fraselle V, Houssiau F, Thonnard JL, Nielens H, Penta M. Validation of the ABILHAND questionnaire as a measure of manual ability in patients with rheumatoid arthritis. Ann Rheum Dis. 2007;66(8):1098-105.
22. El Miedany Y, El Gaafary M, El Aroussy N, Ahmed I, Youssef S, Palmer D. Patient reported outcomes in ankylosing spondylitis: development and validation of a new questionnaire for functional impairment and quality of life assessment. Clin Exp Rheumatol. 2011;29(5):801-10.
23. Eyres S, Tennant A, Kay L, Waxman R, Helliwell PS. Measuring disability in ankylosing spondylitis: comparison of bath ankylosing spondylitis functional index with revised Leeds Disability Questionnaire. J Rheumatol. 2002;29(5):979-86.
24. Gilworth G, Chamberlain MA, Bhakta B, Haskard D, Silman A, Tennant A. Development of the BD-QoL: a quality of life measure specific to Behcet's disease. J Rheumatol. 2004;31(5):931-7.
25. Gilworth G, Chamberlain MA, Harvey A, Woodhouse A, Smith J, Smyth MG, et al. Development of a work instability scale for rheumatoid arthritis. Arthritis Rheum. 2003;49(3):349-54.
26. Gilworth G, Emery P, Barkham N, Smyth MG, Helliwell P, Tennant A. Reducing work disability in Ankylosing Spondylitis: development of a work instability scale for AS. BMC Musculoskelet Disord. 2009;10:68.
27. Gilworth G, Emery P, Gossec L, Vliet Vlieland TP, Breedveld FC, Hueber AJ, et al. Adaptation and cross-cultural validation of the rheumatoid arthritis work instability scale (RA-WIS). Ann Rheum Dis. 2009;68(11):1686-90.
28. Goetz C, Ecosse E, Rat AC, Pouchot J, Coste J, Guillemin F. Measurement properties of the osteoarthritis of knee and hip quality of life OAKHQOL questionnaire: an item response theory analysis. Rheumatology. 2011;50(3):500-5.
29. Haley SM, McHorney CA, Ware JE. Evaluation of the MOS SF-36 physical functioning scale (PF-10): I. Unidimensionality and reproducibility of the Rasch item scale. J Clin Epidemiol. 1994;47(6):671-84.
30. Hassett AL, Li T, Buyske S, Savage SV, Gignac MA. The multi-faceted assessment of independence in patients with rheumatoid arthritis: preliminary validation from the ATTAIN study. Curr Med Res Opin. 2008;24(5):1443-53.
31. Haugen IK, Moe RH, Slatkowsky-Christensen B, Kvien TK, van der Heijde D, Garratt A. The AUSCAN subscales, AIMS-2 hand/finger subscale, and FIOHA were not unidimensional scales. J Clin Epidemiol. 2011;64(9):1039-46.
32. Haywood KL, Garratt AM, Jordan KP, Healey EL, Packham JC. Evaluation of ankylosing spondylitis quality of life (EASi-QoL): Reliability and validity of a new patient-reported outcome measure. J Rheumatol. 2010;37(10):2100-9.
33. Helliwell P, Reay N, Gilworth G, Redmond A, Slade A, Tennant A, et al. Development of a foot impact scale for rheumatoid arthritis. Arthritis Rheum. 2005;53(3):418-22.

34. Hirsch JD, Lee SJ, Terkeltaub R, Khanna D, Singh J, Sarkin A, et al. Evaluation of an Instrument Assessing Influence of Gout on Health-Related Quality of Life. J Rheumatol. 2008;35(12):2406-14.
35. Jette AM, McDonough CM, Haley SM, Ni PS, Olarsch S, Latham N, et al. A computer-adaptive disability instrument for lower extremity osteoarthritis research demonstrated promising breadth, precision, and reliability. J Clin Epidemiol. 2009;62(8):807-15.
36. Katz PP, Radvanski DC, Allen D, Buyske S, Schiff S, Nadkarni A, et al. Development and validation of a short form of the valued life activities disability questionnaire for rheumatoid arthritis. Arthritis Care Res. 2011;63(12):1664-71.
37. Keenan AM, McKenna SP, Doward LC, Conaghan PG, Emery P, Tennant A. Development and validation of a needs-based quality of life instrument for osteoarthritis. Arthritis Care Res. 2008;59(6):841-8.
38. Kelly PA, Kallen MA, Suarez-Almazor ME. A combined-method psychometric analysis recommended modification of the multidimensional health locus of control scales. J Clin Epidemiol. 2007;60(5):440-7.
39. Kersten P, White PJ, Tennant A. The Visual Analogue WOMAC 3.0 scale - internal validity and responsiveness of the VAS version. BMC Musculoskelet Disord. 2010;11.
40. Ko Y, Lo N-N, Yeo S-J, Yang K-Y, Yeo W, Chong H-C, et al. Rasch analysis of the Oxfort Knee Score. Osteoarthr Cartilage. 2009;17:1163-9.
41. Kopec JA, Sayre EC, Davis AM, Badley EM, Abrahamowicz M, Sherlock L, et al. Assessment of health-related quality of life in arthritis: conceptualization and development of five item banks using item response theory. Health Qual Life Outcomes. 2006;4:33.
42. Köse SK, Öztuna D, Kutlay S, Elhan AH, Tennant A, Kücükdeveci AA. Psychometric properties of the Health Assessment Questionnaire Disability Index (HAQ-DI) and the Modified Health Assessment Questionnaire (MHAQ) in patients with knee osteoarthritis. Turk J Rheumatol. 2010;25:147-55.
43. Kosinski M, Bjorner JB, Ware JE, Sullivan E, Straus WL. An evaluation of a patient-reported outcomes found computerized adaptive testing was efficient in assessing osteoarthritis impact. J Clin Epidemiol. 2006;59(7):715-23.
44. Kristjansson E, Tugwell PS, Wilson AJ, Brooks PM, Driedger SM, Gallois C, et al. Development of the effective musculoskeletal consumer scale. J Rheumatol. 2007;34(6):1392-400.
45. Kucukdeveci AA, Sahin H, Ataman S, Griffiths B, Tennant A. Issues in cross-cultural validity: example from the adaptation, reliability, and validity testing of a Turkish version of the Stanford Health Assessment Questionnaire. Arthritis Rheum. 2004;51(1):14-9.
46. Kurtais Y, Oztuna D, Küçükdeveci AA, Kutlay S, Hafiz M, Tennant A. Reliability, construct validity and measurement potential of the ICF comprehensive core set for osteoarthritis. Bmc Musculoskel Dis. 2011;12:255.
47. Kutlay S, Kucukdeveci AA, Gonul D, Tennant A. Adaptation and validation of the Turkish version of the Rheumatoid Arthritis Quality of Life Scale. Rheumatol Int. 2003;23(1):21-6.
48. Lawton G, Bhakta BB, Chamberlain MA, Tennant A. The Behcet's disease activity index. Rheumatology. 2004;43(1):73-8.
49. Lee YS, Douglas J, Chewning B. Techniques for developing health quality of life scales for point of service use. Soc Indic Res. 2007;83(2):331-50.
50. Leong KP, Kong KO, Thong BY, Koh ET, Lian TY, Teh CL, et al. Development and preliminary validation of a systemic lupus erythematosus-specific quality-of-life instrument (SLEQOL). Rheumatology. 2005;44(10):1267-76.
51. Leung YY, Tam LS, Kun EW, Ho KW, Li EK. Comparison of 4 functional indexes in psoriatic arthritis with axial or peripheral disease subgroups using Rasch analyses. J Rheumatol. 2008;35(8):1613-21.
52. Martin M, Kosinski M, Bjorner JB, Ware JE, Jr., Maclean R, Li T. Item response theory methods can improve the measurement of physical function by combining the modified health assessment questionnaire and the SF-36 physical function scale. Qual Life Res. 2007;16(4):647-60.

53. McHorney CA, Cohen AS. Equating health status measures with item response theory. Med Care. 2000;38(9):II-43-II-59.
54. McKenna SP, Doward LC, Whalley D, Tennant A, Emery P, Veale DJ. Development of the PsAQoL: a quality of life instrument specific to psoriatic arthritis. Ann Rheum Dis. 2004;63(2):162-9.
55. McTaggart-Cowan HM, Brazier JE, Tsuchiya A. Clustering Rasch results: A novel method for developing rheumatoid arthritis states for use in valuation studies. Value Health. 2010;13(6):787-95.
56. Mielenz TJ, Edwards MC, Callahan LF. First item response theory analysis on Tampa Scale for Kinesiophobia (fear of movement) in arthritis. J Clin Epidemiol. 2010;63(3):315-20.
57. Mielenz TJ, Edwards MC, Callahan LF. Item-response-theory analysis of two scales for self-efficacy for exercise behavior in people with arthritis. J Aging Phys Act. 2011;19(3):239-48.
58. Moorer P, Suurmeije Th P, Foets M, Molenaar IW. Psychometric properties of the RAND-36 among three chronic diseases (multiple sclerosis, rheumatic diseases and COPD) in The Netherlands. Qual Life Res. 2001;10(7):637-45.
59. Ndosi M, Tennant A, Bergsten U, Kukkurainen ML, Machado P, de la Torre-Aboki J, et al. Cross-cultural validation of the Educational Needs Assessment Tool in RA in 7 European countries. BMC Musculoskelet Disord. 2011;12:110.
60. Niedermann K, Forster A, Ciurea A, Hammond A, Uebelhart D, de Bie R. Development and psychometric properties of a joint protection self-efficacy scale. Scand J Occup Ther. 2011;18(2):143-52.
61. Niedermann K, Forster A, Hammond A, Uebelhart D, de Bie R. Development and validation of a German version of the joint protection behavior assessment in patients with rheumatoid arthritis. Arthritis Rheum. 2007;57(2):249-55.
62. Nordenskiold U, Grimby G, Hedberg M, Wright B, Linacre JM. The structure of an instrument for assessing the effects of assistive devices and altered working methods in women with rheumatoid arthritis. Arthritis Care Res. 1996;9(5):358-67.
63. O'Malley KJ, Suarez-Almazor M, Aniol J, Richardson P, Kuykendall DH, Moseley JB, Jr., et al. Joint-specific multidimensional assessment of pain (J-MAP): factor structure, reliability, validity, and responsiveness in patients with knee osteoarthritis. J Rheumatol. 2003;30(3):534-43.
64. Osborne RH, Elsworth GR, Whitfield K. The Health Education Impact Questionnaire (heiQ): an outcomes and evaluation measure for patient education and self-management interventions for people with chronic conditions. Patient Educ Couns. 2007;66(2):192-201.
65. Pallant JF, Keenan AM, Misajon R, Conaghan PG, Tennant A. Measuring the impact and distress of osteoarthritis from the patients' perspective. Health Qual Life Outcomes. 2009;7.
66. Paulsen T, Grotle M, Garratt A, Kjeken I. Development and psychometric testing of the patient-reported measure of activity performance of the hand (MAP-Hand) in rheumatoid arthritis. J Rehabil Med. 2010;42(7):636-44.
67. Penta M, Thonnard JL, Tesio L. ABILHAND: a Rasch-built measure of manual ability. Arch Phys Med Rehabil. 1998;79(9):1038-42.
68. Perkins K, Hoffman RW, Bezruczko N. A Rasch analysis for classification of systemic lupus erythematosus and mixed connective tissue disease. J Appl Meas. 2008;9(2):136-50.
69. Perruccio AV, Lohmander LS, Canizares M, Tennant A, Hawker GA, Conaghan PG, et al. The development of a short measure of physical function for knee OA KOOS-Physical Function Shortform (KOOS-PS) - an OARSI/OMERACT initiative. Osteoarthritis Cartilage. 2008;16(5):542-50.
70. Pham T, van der Heijde DM, Pouchot J, Guillemin F. Development and validation of the French ASQoL questionnaire. Clin Exp Rheumatol. 2010;28(3):379-85.
71. Pollard B, Dixon D, Dieppe P, Johnston M. Measuring the ICF components of impairment, activity limitation and participation restriction: an item analysis using classical test theory and item response theory. Health Qual Life Outcomes. 2009;7:41.

72. Pouchot J, Ecosse E, Coste J, Guillemin F. Validity of the childhood health assessment questionnaire is independent of age in juvenile idiopathic arthritis. Arthritis Rheum. 2004;51(4):519-26.
73. Rauch A, Cieza A, Boonen A, Ewert T, Stucki G. Identification of similarities and differences in functioning in persons with rheumatoid arthritis and ankylosing spondylitis using the International Classification of Functioning, Disability and Health (ICF). Clin Exp Rheumatol. 2009;27(4 Suppl 55):S92-101.
74. Roorda LD, Jones CA, Waltz M, Lankhorst GJ, Bouter LM, van der Eijken JW, et al. Satisfactory cross cultural equivalence of the Dutch WOMAC in patients with hip osteoarthritis waiting for arthroplasty. Ann Rheum Dis. 2004;63(1):36-42.
75. Ryser L, Wright BD, Aeschlimann A, Mariacher-Gehler S, Stucki G. A new look at the Western Ontario and McMaster Universities Osteoarthritis Index using Rasch analysis. Arthritis Care Res. 1999;12(5):331-5.
76. Sheehan TJ, DeChello LM, Garcia R, Fifield J, Rothfield N, Reisine S. Measuring disability: application of the Rasch model to activities of daily living (ADL/IADL). J Outcome Meas. 2001;5(1):839-63.
77. Sheehan TJ, DuBrava S, Fifield J, Reisine S, DeChello L. Rate of change in functional limitations for patients with rheumatoid arthritis: effects of sex, age, and duration of illness. J Rheumatol. 2004;31(7):1286-92.
78. Siemons L, ten Klooster PM, Taal E, Kuper IH, van Riel P, van de Laar M, et al. Validating the 28-Tender Joint Count Using Item Response Theory. J Rheumatol. 2011;38(12):2557-64.
79. Steultjens MPM, Dekker J, van Baar ME, Oostendorp RAB, Bijlsma JWJ. Internal consistency and validity of an observational method for assessing disability in mobility in patients with osteoarthritis. Arthritis Care Res. 1999;12(1):19-25.
80. Suurmeijer TP, Doeglas DM, Moum T, Briancon S, Krol B, Sanderman R, et al. The Groningen Activity Restriction Scale for measuring disability: its utility in international comparisons. Am J Public Health. 1994;84(8):1270-3.
81. Tammaru M, McKenna SP, Meads DM, Maimets K, Hansen E. Adaptation of the rheumatoid arthritis quality of life scale for Estonia. Rheumatol Int. 2006;26(7):655-62.
82. Tang K. Disease-related differential item functioning in the work instability scale for rheumatoid arthritis: converging results from three methods. Arthritis Care Res. 2011;63(8):1159-69.
83. Tang K, Beaton DE, Lacaille D, Gignac MAM, Zhang W, Anis AH, et al. The Work Instability Scale for Rheumatoid Arthritis (RA-WIS): Does it work in osteoarthritis? Qual Life Res. 2010;19(7):1057-68.
84. Taylor WJ, Colvine K, Gregory K, Collis J, McQueen FM, Dalbeth N. The Health Assessment Questionnaire Disabillity Index is a valid measure of physical function in gout. Clin Exp Rheumatol. 2008;26(4):620-6.
85. Taylor WJ, McPherson KM. Using Rasch analysis to compare the psychometric properties of the Short Form 36 physical function score and the Health Assessment Questionnaire disability index in patients with psoriatic arthritis and rheumatoid arthritis. Arthritis Rheum. 2007;57(5):723-9.
86. ten Klooster PM, Taal E, van de Laar MA. Rasch analysis of the Dutch Health Assessment Questionnaire disability index and the Health Assessment Questionnaire II in patients with rheumatoid arthritis. Arthritis Rheum. 2008;59(12):1721-8.
87. Tennant A, Hillman M, Fear J, Pickering A, Chamberlain MA. Are we making the most of the Stanford Health Assessment Questionnaire? Br J Rheumatol. 1996;35(6):574-8.
88. Tennant A, Kearns S, Turner F, Wyatt S, Haigh R, Chamberlain MA. Measuring the function of children with juvenile arthritis. Rheumatology. 2001;40(11):1274-8.
89. Uhlig T, Lillemo S, Moe RH, Stamm T, Cieza A, Boonen A, et al. Reliability of the ICF Core Set for rheumatoid arthritis. Ann Rheum Dis. 2007;66(8):1078-84.
90. van Groen MM, ten Klooster PM, Taal E, van de Laar MAFJ, Glas CAW. Application of the health assessment questionnaire disability index to various rheumatic diseases. Qual Life Res. 2010;19(9):1255-63.
91. Vanthuyne M, Smith V, Arat S, Westhovens R, Keyser FD, Houssiau FA, et al. Validation of a manual ability questionnaire in patients with systemic sclerosis. Arthritis Care Res. 2009;61(5):695-703.

92. Wolfe F. Which HAQ is best? A comparison of the HAQ, MHAQ and RA-HAQ, a difficult 8 item HAQ (DHAQ), and a rescored 20 item HAQ (HAQ20): analyses in 2,491 rheumatoid arthritis patients following leflunomide initiation. J Rheumatol. 2001;28(5):982-9.
93. 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(2):369-78.
94. Wolfe F, Hawley DJ, Goldenberg DL, Russell IJ, Buskila D, Neumann L. The assessment of functional impairment in fibromyalgia (FM): Rasch analyses of 5 functional scales and the development of the FM Health Assessment Questionnaire. J Rheumatol. 2000;27(8):1989-99.
95. Wolfe F, Kong SX. Rasch analysis of the Western Ontario MacMaster questionnaire (WOMAC) in 2205 patients with osteoarthritis, rheumatoid arthritis, and fibromyalgia. Ann Rheum Dis. 1999;58(9):563-8.
96. Wolfe F, Michaud K, Kahler K, Omar M. The Short Arthritis Assessment Scale: a brief assessment questionnaire for rapid evaluation of arthritis severity in research and clinical practice. J Rheumatol. 2004;31(12):2472-9.
97. Wolfe F, Michaud K, Pincus T. Development and validation of the health assessment questionnaire II: a revised version of the health assessment questionnaire. Arthritis Rheum. 2004;50(10):3296-305.
98. Wolfe F, van der Heijde DM, Larsen A. Assessing radiographic status of rheumatoid arthritis: introduction of a short erosion scale. J Rheumatol. 2000;27(9):2090-9.
99. Woodburn J, Vliet Vlieland TP, van der Leeden M, Steultjens MP. Rasch analysis of Dutch-translated version of the Foot Impact Scale for rheumatoid arthritis. Rheumatology. 2011;50(7):1315-9.

Chapter 3

Validating the 28-tender joint count using item response theory

L. Siemons
P.M. ten Klooster
E. Taal
H.H. Kuper
P.L.C.M. van Riel
M.A.F.J. van de Laar
C.A.W. Glas

J Rheumatol. 2011;38(12):2557-64.

Abstract

Objective: To examine the construct validity of the 28-tender joint count (TJC28) using item response theory (IRT)-based methods.

Methods: A total of 457 patients with early stage rheumatoid arthritis (RA) were included. Internal construct validity of the TJC28 was evaluated by determining whether the TJC28 fitted a 2-parameter logistic IRT model. As well, we tested whether the discrimination and difficulty parameters of the joints properly reflected the known left-right symmetry of joint involvement. External validity was evaluated by correlations with other established measures of disease activity, including pain, disability, general health, erythrocyte sedimentation rate (ESR), and the 28-swollen joint count.

Results: The TJC28 showed a good fit with the 2-parameter logistic model, with no relevant differential item functioning across sex, age and time and with excellent reliability. The 28 joints covered a reasonable range of disease activity, even though they were mainly targeted at patients with moderate or high disease activity levels. The joint parameters reflected the left-right symmetry of joint involvement for all pairs of joints except one. All disease activity measures, except ESR, were significantly correlated with the TJC28. Most correlations were of the expected magnitude.

Conclusion: The TJC28 showed good internal and acceptable external construct validity for patients with early-stage RA. The IRT analyses did point to some potential limitations of the instrument, a major problem being its limited measurement range. Future research should examine whether instrument modifications might lead to a more robust assessment of disease activity in patients with RA.

Introduction

Rheumatoid arthritis (RA) is a systemic autoimmune disease, decreasing life expectancy by 3 to 10 years compared to the general population [1, 2]. People with RA experience chronic inflammation of joints and periarticular tissues [3], characterized by symmetric pain and swelling in the joints [4-7]. The disease generally follows an unpredictable course, often with alternating periods of mild and severe disease activity [3].

RA treatments are aimed at reaching a state of remission as soon as possible [8]. Because joint tenderness is an important characteristic of RA, joint counts that measure the extent of joint tenderness are used for the assessment of RA severity [9]. A joint count is a specific quantitative clinical measure to assess the status of a patient with RA [10]. Therefore, it forms a major component of indices of disease activity [11] and remission [12]. Although various joint counts have been developed, ranging from the evaluation of 28 to 80 joints, the 28-joint count is currently the most widely used measurement instrument.

Earlier studies showed the 28-tender joint count (TJC28) to be a reliable and valid joint index [9, 13-15]. However, these studies have only used classical test theory (CTT) methods. To date, the construct validity of the TJC28 has never been analyzed using item response theory (IRT)-based methods. IRT is a sophisticated psychometric approach that has been adopted to supplement the more traditional approaches [16] to enable a more thorough evaluation of an instrument's psychometric characteristics. IRT has already been frequently and successfully applied in evaluating and improving health outcome questionnaires [17], but it has rarely been applied to clinical measures, such as tender joint counts. Therefore, the aim of our study was to examine both the internal and external construct validity of the TJC28 using IRT-based methods.

Materials and methods

Patients

Patients with early-stage RA participating in the Dutch Rheumatoid Arthritis Monitoring remission induction cohort [18] were included in this study. This observational, multicenter cohort was established in 2006 to evaluate a treatment strategy aimed at reaching a state of remission. The patients were asked for inclusion in the cohort by their rheumatologists. Patients were qualified for inclusion at the moment of clinical diagnosis of RA. Symptoms duration was a maximum of 1 year, and patients had to be at least 18 years old. Any who had previously used disease-modifying antirheumatic drugs or prednisolone were excluded from the cohort.

The result was a total baseline sample of 457 patients. Measurements were performed during each hospital visit. The data from the first timepoint (i.e. at inclusion) were used for all analyses. In addition, one of the fit analyses (i.e. evaluating differential item functioning across time) was based on data from the first 3 timepoints (t_1= at inclusion, t_2= 8 weeks after inclusion, t_3= 12 weeks after inclusion). Because the duration since inclusion varied among patients, followup measurements involved a decreasing number of patients. At the third timepoint the remaining sample consisted of a total of 391 patients.

Measures

The TJC28 and the 28-swollen joint count (SJC28) were administered separately at each visit by a trained nurse practitioner or rheumatologist. The 28 joints were scored on a dichotomous scale, with 0 indicating “no pain” or “no swelling” in the joint, and 1 indicating “pain” or “swelling” in the joint [9, 19]. Both 28-joint counts include the shoulders, elbows, wrists, and knees, the 10 metacarpophalangeal (MCP) joints, and the 10 proximal interphalangeal (PIP) joints [20].

Besides the TJC28 and the SJC28, patients were asked to complete the Health Assessment Questionnaire – Disability Index (HAQ-DI) [21] which measures physical function, and visual analog scales for pain (VAS pain) and general health (VAS GH). The alternative disability index (HAQ-ADI) [22], which does not correct for the use of aids and devices, was derived from the HAQ-DI and was scored on a scale from 0 to 3 (higher scores indicating more physical disability). Pain and general health were measured on a 100-mm VAS scale, 0 indicating “no pain” or “very good”, and 100 indicating “unbearable pain” or “very bad”.

Laboratory samples were collected before each hospital visit, including the erythrocyte sedimentation rate (ESR), which is a nonspecific measure of inflammation [19].

Statistical analyses

When using an IRT framework, the relationship between item scores and the underlying construct of interest (i.e. the latent trait variable θ, representing the degree of joint tenderness in our study) can be modeled. When applying CTT approaches, sum scores of different combinations of joints can be obtained. However, this does not imply that the sum score reflects a meaningful underlying construct. IRT has several beneficial properties compared to the traditional CTT approach, enabling a more thorough evaluation of an instrument’s psychometric characteristics. If the TJC28 fits an IRT model, this supports the construct validity of the instrument, because this shows that the observed responses can be explained by the underlying structure of the instrument [23]. Further, if attenuation is present (i.e. underestimated correlations between measurements due to unreliability caused by measurement error) IRT can deal with this problem more precisely than the CTT approach since it considers latent correlations instead of sum-score based observed correlations. In addition, IRT can successfully handle incomplete item administration designs and missing data, and where CTT often assumes a normal distribution of the true scores, IRT can deal with various distributions of latent variables [23].

IRT models the probability of a joint being scored as tender on the basis of characteristics of the patient (the degree of joint tenderness: θ) and the item (such as the difficulty and discrimination level). Each single joint is regarded as an item and has a corresponding IRT model curve. Two widely applied IRT models are the Rasch model (also known as the 1-parameter logistic model) and the 2-parameter logistic (2-PL) model, both shown in Figure 1. The y-axis shows the probability of a joint to be scored as tender, while the x-axis shows the latent trait that corresponds to the degree of joint tenderness a patient experiences (θ, scaled around zero). Figure 1A shows the Rasch model, including 3 joints with different difficulty parameters [24]. The value of the difficulty parameter of a specific joint equals the point on the x-axis at which the patient has a probability of 0.5 of having a painful joint [24, 25]. So, for joint 1 its value will be equal to -1. Figure 1B shows the 2-PL model. In it, the curves intersect because of the addition of the discrimination parameter. This parameter is proportional to the slope of the curve; the higher its value, the steeper the slope, and the better the joint discriminates between patients with various degrees of joint tenderness [25].

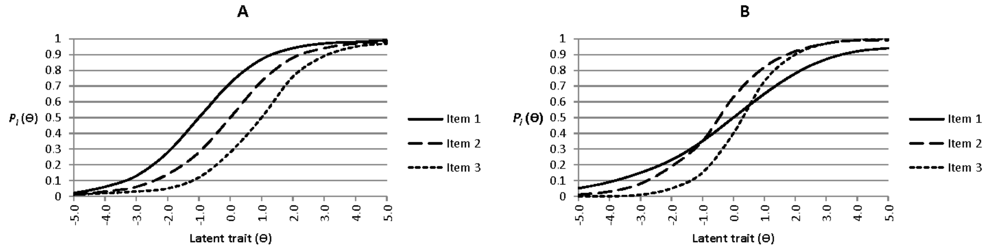


Figure 1 – Graphic representation of the Rasch model (A) and the 2-parameter logistic model (B), where Pi(θ) is the probability of a joint to be scored as tender.

In our study, a 2-PL model was used to analyze the construct validity of the TJC28. This was motivated by both practical and empirical reasons. First, we wanted to examine whether the symmetry that characterizes RA is reflected in both the difficulty and the discrimination parameters of the IRT model. Second, a log-likelihood ratio test showed that the 2-PL model had a significantly better fit to the TJC28 than the Rasch model (log-likelihood ratio test=163.81, df=27, *p*<0.01).

Internal construct validity

This was assessed by evaluating whether the TJC28 could be fitted to the 2-PL model, whether the joint parameters truly reflected the known left-right symmetry of joint involvement in patients with RA [4, 5], and whether the TJC28 had an acceptable reliability.

A) Fit analyses

IRT models rely on several assumptions. One of these concerns the shape of the response curves. Using a Lagrange multiplier test, the LM-Q1-test [26], it was determined whether the shape of the curves belonging to the TJC28 fitted the shape of the curves assumed by the 2-PL model. This means the joint curves have various difficulty parameters, various discrimination parameters, and a lower zero asymptote. Two outcome values considered important for determining the fit of the curves with the LM-Q1-test are the *p*-value of the test and the effect size [27]. A *p*-value >0.05 indicates a good item-model fit, but this statistic is sensitive to large sample sizes [27]. For large sample sizes, the absolute effect size should also be evaluated. The effect size is given by the difference between the observed and the expected average score on an item in a specific group and can, therefore, range between 0 and 1. An effect size of <0.10 has been previously used as an acceptable measure for item model fit [28]. Well-fitting response curves can also be seen as strong evidence for unidimensionality of the TJC28 [29].

Additionally, it was examined whether differential item functioning (DIF) across sex, age and time was present. A joint shows DIF across sex or age if individuals from different groups (e.g. men vs. women) but with the same latent trait value, do not have the same probability of reporting a joint as being tender [24]. DIF across time is present when the joint difficulty parameters are unstable over time [27]. The stability of the parameters was examined over the first 3 timepoints.

B) Left-right symmetry of joint parameters

The symmetry of the difficulty and discrimination parameters was simultaneously tested for each pair of joints using a Wald test [30]. This test determines whether the parameter values of the left-side joint and the parameter values of the right-side joint are equal. Nonsignificant results ($p > 0.05$) indicate that the joint parameters properly reflect the known left-right symmetry of joint involvement.

C) Reliability and measurement precision

In IRT, the reliability of the TJC28 is estimated as the ratio of the expectation of the posterior variance of the latent variable θ given the instrument-score, and the total variance of θ [29]. A value >0.70 is considered acceptable for group use, while a value of 0.85 or higher is required for individual use [16]. The IRT reliability coefficient is equivalent to Cronbach's alpha.

When applying IRT, the range of θ for which a joint or the total TJC28 is most reliable for measuring patients' levels of joint tenderness can be depicted in an information curve. An information curve shows the range over θ where the individual joint or the total TJC28 can best discriminate among individual patients [31]. Ideally, the instrument includes joints with high discrimination parameters that cover a broad spectrum of joint difficulties. In this way, the spectrum of joint tenderness can be measured as precisely as possible. The higher the information level of a joint, the more the joint contributes to the measurement precision of joint tenderness. Information curves of individual joints were plotted for evaluation of the performance of each single joint. The test information curve of the TJC28 and its associated reliability levels [r = 1-(1/test information at θ)] were plotted to evaluate the performance of the total TJC28.

External construct validity

Previous studies used sum scores of the TJC to determine its correlation with other established measures of disease activity, while IRT uses latent trait values (θ). The external construct validity of the TJC28 was evaluated by examining whether the baseline θ values and traditional sum scores of the TJC28 showed an expected pattern of correlations with 5 other established measures of disease activity [32]: VAS pain, HAQ-ADI, VAS GH, ESR, and the SJC28.

Correlations <0.3 were defined as weak (low), between 0.3 and 0.6 as moderate, and >0.6 as strong (high) [33]. All correlations were expected to be both positive and significant. Although highly variable correlations between the TJC and these variables were found in previous studies, moderate correlations were expected since they are all measures of disease activity [9, 14, 34-38].

Results

Demographics at inclusion

Baseline data were available from 457 patients (288 women and 169 men). The mean (SD) age at inclusion was 55.4 (15.2) years for the women and 59.8 (12.4) years for the men. Baseline measures of disease activity are summarized in Table 1. The TJC28 had a mean score of 5.7. For interpretation, a TJC28 score of 0 corresponded to an estimated θ score in the range of -1.65 to -0.69, and a TJC28 score of 28 corresponded to estimated θ scores in the range of 2.82 to 3.25.

Internal construct validity

Fit analyses

Table 2 presents the results of the fit analyses. Although some joints showed a statistically significant misfit (*p*<0.05), all effect sizes were well below 0.10. These results indicate that there was a good fit between the curves of the TJC28 and the 2-PL model. In addition, there was no relevant DIF across sex, age (median split: ≤59 vs. ≥60 years), and time.

Table 1 - Mean scores (SD) of established measures of disease activity at baseline in 457 patients with early-stage rheumatoid arthritis.

Measures	Scoring scale	Mean (SD)
TJC28	0-28	5.7 (5.7)
VAS pain	0-100	49.4 (25.4)
HAQ-ADI	0-3	1.0 (0.7)
VAS GH	0-100	49.9 (25.2)
SJC28	0-28	7.9 (5.7)
ESR	0-140	29.6 (22.0)
DAS28	0-10	4.7 (1.4)

TJC28: tender joint count for 28 joints, VAS: visual analog scale, HAQ-ADI: Health Assessment Questionnaire- Alternative Disability Index, GH: patient's general health assessment, SJC28: swollen joint count for 28 joints, ESR: erythrocyte sedimentation rate; DAS28: Disease Activity Score for 28 joints

Table 2 - Results of the fit analyses.

Fit analysis	No. joints with p≤0.05	Effect size
Fit of the curves	4	≤ 0.03
Sex differences	4	≤ 0.06
Age differences	1	≤ 0.03
Constancy of location parameters over time	7	≤ 0.06

Left-right symmetry of joint parameters

Table 3 presents the parameter estimates generated by the 2-PL model. The Wald test showed a nonsignificant result for all pair of joints except 1. This demonstrates that both the difficulty and the discrimination parameters properly reflected the left-right symmetry of joint involvement, which is characteristic of RA.

Reliability and measurement precision

The reliability of the TJC28 was acceptable for group use as well as for individual use (r = 0.874).

Table 3 presents the discrimination parameter values, ranging from 0.670 to 1.049 for larger joints (shoulders, elbows, wrists, knees), and from 1.369 to 2.269 for smaller joints (the MCPs and PIPs). Joint difficulties covered only the positive half of the spectrum, ranging from 0.613 to 3.659, reflecting low response probabilities. This limited range of joint difficulties was also reflected in the information curves (Figure 2). The test information curve showed that the scale measured the patient's level of θ with a reliability level acceptable for group use (r > 0.70) over the range from θ = -0.60 to θ = +3.05 [31]. Outside this range, the test information curve and the scale's reliability rapidly decreased, meaning that the corresponding levels of θ were estimated with reduced precision. Over the range from θ = 0.0 to θ = +2.5, the scale's reliability was also acceptable for individual use (r > 0.85). The reliability was at its highest point (r > 0.93) at θ = +1.3. The item information curves showed that smaller joints (MCPs and PIPs) provided more information to the test than larger joints (shoulders, elbows, wrists, knees).

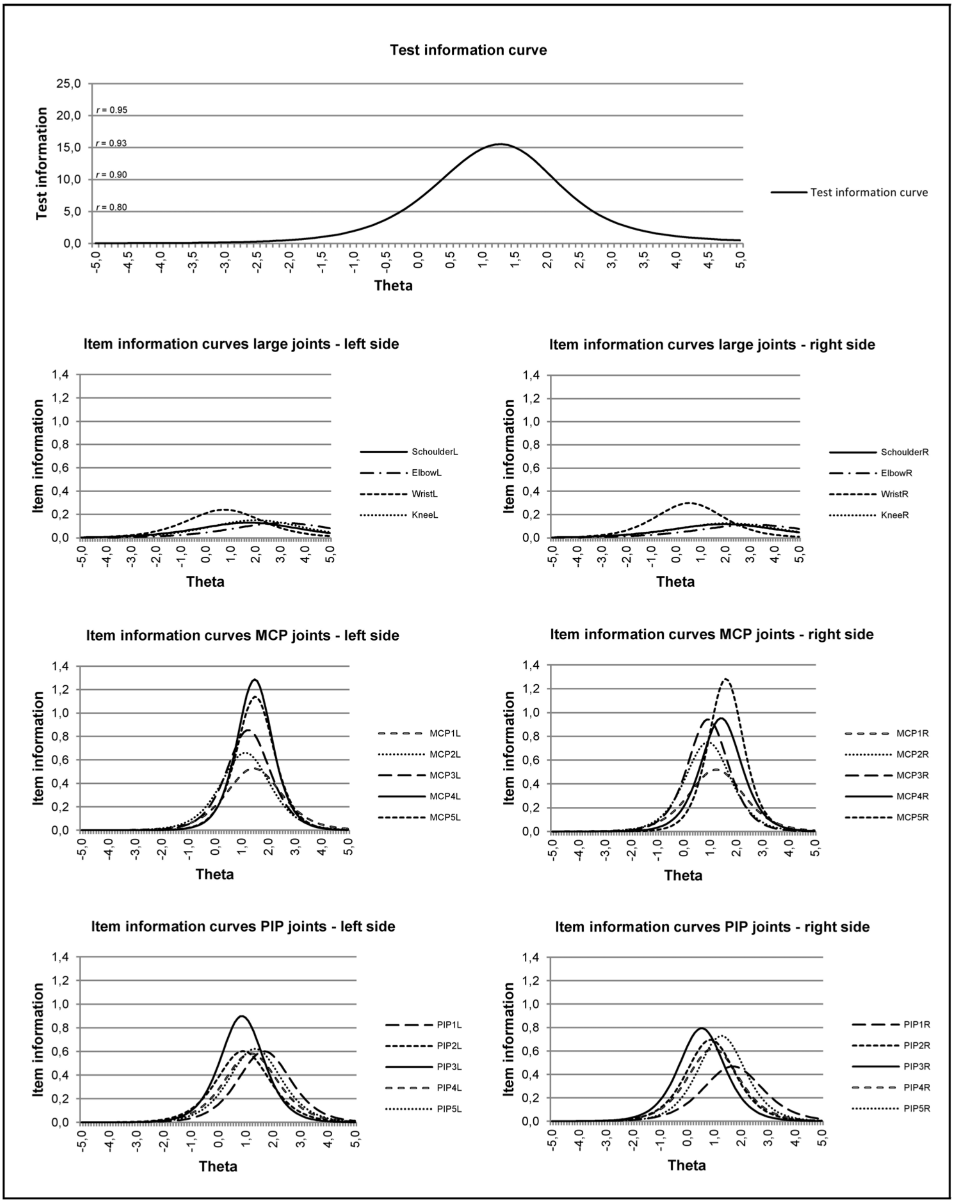


Figure 2 –Top graph shows the test information curve of the 28-tender joint count with its associated reliability level. The graphs below represent the item information curves for the joints on the left side (left column) and right side (right column) of the body.

Table 3 – Average joint scores, item response theory joint parameter values, and Wald test results.

Joint	Average joint score[*1]		Discrimination parameter		Difficulty parameter		Results	
	Left	Right	Left	Right	Left	Right	Wald test value[*2]	*p*-value[*3]
Shoulder	0.24	0.22	0.721	0.686	1.252	1.350	0.139	0.93
Elbow	0.12	0.13	0.713	0.670	2.212	2.101	0.187	0.91
Wrist	0.35	0.38	0.983	1.094	0.747	0.613	0.295	0.86
MCP1	0.17	0.21	1.453	1.443	2.130	1.796	2.499	0.29
MCP2	0.21	0.25	1.627	1.735	1.867	1.617	1.756	0.42
MCP3	0.18	0.24	1.850	1.943	2.336	1.822	7.056	0.03
MCP4	0.12	0.14	2.269	1.953	3.406	2.804	4.415	0.11
MCP5	0.12	0.11	2.135	2.266	3.255	3.659	2.580	0.28
PIP1	0.13	0.14	1.552	1.369	2.613	2.334	0.974	0.61
PIP2	0.26	0.26	1.554	1.669	1.415	1.469	0.071	0.96
PIP3	0.26	0.33	1.898	1.784	1.650	0.986	4.617	0.10
PIP4	0.21	0.26	1.528	1.577	1.828	1.541	1.328	0.51
PIP5	0.18	0.19	1.579	1.708	2.131	2.196	0.072	0.96
Knee	0.20	0.20	0.778	0.710	1.572	1.516	0.017	0.99

*[*1] Average score on a scale from 0 to 1, [*2] with 2 degrees of freedom, [*3] p-value for a simultaneous test for differences in difficulty and/or discrimination. MCP: metacarpophalangeal, PIP: proximal interphalangeal.*

External construct validity

Spearman's correlations with the other established measures of disease activity for both the θ estimations and the sum scores of the TJC28 are shown in Table 4.

Table 4 – Spearman's correlations of the TJC28 (using θ as well as the sum score) with the sum scores of the established measures of disease activity at baseline (N=457). Except where indicated, p≤0.01 (2-tailed).

Measurement of disease activity	Correlation with θ TJC28	Correlation with sum score TJC28
VAS pain	0.279	0.280
HAQ-ADI	0.405	0.416
VAS GH	0.305	0.302
SJC28	0.453	0.440
ESR	0.023 (*p*=0.626)	0.064 (*p*=0.177)

TJC28: tender joint count for 28 joints, VAS: visual analog scale, HAQ-ADI: Health Assessment Questionnaire-Alternative Disability Index, GH: patient's general health assessment, SJC28: swollen joint count for 28 joints, ESR: erythrocyte sedimentation rate.

The correlations based on the θ estimations of the TJC28 were very similar to the correlations based on the sum scores. As expected, all correlations were positive. However, for both the θ estimates and the sum scores of the TJC28 only 4 out of 5 correlations were significant. The HAQ-ADI, joint swelling, and the patient's general health assessment did show the expected moderate correlations. Pain correlated less strongly with joint tenderness than expected, but the correlation was only just below the cut-off point of 0.30. However, a very low correlation was found with ESR.

Discussion

This is the first study to examine the validity of the TJC28 by applying IRT-based methods. As a result, the instrument's psychometric characteristics can be evaluated more thoroughly than with CTT alone. The results showed that the TJC28 is a valid and reliable measure for patients with early-stage RA. An acceptable fit of the TJC28 to the 2-PL model was demonstrated, with no relevant DIF across sex, age and time, and with excellent reliability. The joints included in the TJC28 covered a reasonable range of disease activity, although measurement precision was limited for lower levels of disease activity. Additionally, the joint parameters properly reflected the left-right symmetry of joint involvement. Evaluation of the external validity showed that all correlations, except with ESR, were similar to the correlations found in previous studies.

Statistical transformations of the ESR values, such as square root and natural logarithm transformations as performed in the Disease Activity Score for 28 joints [15], did not improve the correlation with joint tenderness. A limited distribution of ESR values within the patient sample might explain the non-significance of this correlation. However, given the high SD (22.04) of the ESR values, this does not seem plausible. Moreover, secondary analyses did show significant and higher correlations between ESR and all other measures of disease activity (*r* between 0.17 for the VAS GH and 0.30 for the HAQ-ADI), and C-Reactive Protein (*r*=0.64), another measure of inflammation. Evaluation of the correlations with the individual joints showed that ESR was significantly correlated with the larger joints (*r* between 0.10 and 0.14), but not with the smaller joints that constitute the largest part of the TJC28. This higher correlation with larger joints is in accord with earlier findings [37] and suggests that the ESR mainly reflects the volume of inflammation in the larger joints, while the TJC28 is also in large part explained by the smaller joints. Future studies should evaluate the correlation between the TJC28 and ESR in an RA population in which the patients have more inflamed smaller joints than in the current sample, to determine whether joint size affects this correlation.

Examination of the correlations showed that those based on the θ estimations of the TJC28 were very similar to the correlations based on the sum scores of the TJC28. This indicates that the θ scores and the sum scores corresponded highly to each other and that attenuation did not pose any serious problems in our study, diminishing the actual advantage of using IRT-based scores instead of sum scores for evaluating the external construct validity of the TJC28. It also suggests that it is adequate to use sum scores for the calculation of a patient's TJC.

The unequal discrimination parameters give additional support for use of the 2-PL instead of the Rasch model, since those parameters are assumed to be equal in the Rasch model. The parameter results also showed that especially the smaller joints showed high discrimination parameters, indicating that the MCP and PIP joints discriminate better between patients with different degrees of joint tenderness (θ) than do larger joints (shoulder, elbow, wrist, knee; Table 3). This is in line with the clinical experience of healthcare providers treating patients with RA. A point of interest regarding the joint difficulties is that the wrists show the lowest values (Table 3). They also have the highest average score (left wrist: 0.35, right wrist: 0.38), which is consistent with the clinical experience that the wrist is a commonly affected joint in RA [39, 40]. This is also reflected in the minor degree of information the wrists provide to the test. However, the wrists do provide some information at the lower levels of disease activity, which can be regarded as a positive property given the limited measurement range of the instrument along the lower range of disease activity.

The results concerning the reflection of left-right symmetry of joint involvement in the joint parameters reflect several studies that emphasize that symmetry of joint involvement characterizes RA [4-7], providing additional support for the construct validity of the TJC28. From a strict test perspective it can be argued that this would imply that half of the joints can be removed from the TJC28. After all, it can point to redundant items, which might be locally dependent, and that make the test unnecessary long. However, removing items might have an effect on the psychometric characteristics of the test by reducing the test information and its corresponding reliability. Moreover, from a clinical perspective it is probably undesirable to remove half of the joints, because a patient's total number of tender joints are being used for individual diagnosis and treatment decisions.

The IRT analyses showed that the TJC28 is a highly reliable instrument; however, this does not imply that the scale also has high interrater reliability. Interrater bias might still be embedded in the inaccuracies of the measure. It is expected, however, that this type of bias did not pose any serious problems in our study, since problems with interrater reliability have mainly been reported for graded or weighted joint counts [19, 41, 42], while a nongraded TJC was used our study.

The accuracy and broadness of the test, given the high discrimination parameters and the range of joint difficulties covered, make accurate measurement of change over time possible. The advantage of using IRT instead of the more traditional approaches is that latent trait values are used instead of sum scores. Even when there are data missing, the latent trait values can still be estimated.

IRT has been successfully applied for the evaluation and improvement of questionnaires of health outcome measures [17]. Since the focus of IRT is at the item level instead of the test-level, the contribution of each single joint can be evaluated without knowledge of the other joints in the instrument [24], a feature that is not available in procedures based on CTT methods. Among others, this feature makes it possible to obtain joint counts with lesser joints without major loss of measurement precision [24]. However, IRT has rarely been applied for the evaluation or improvement of clinical measures, such as a TJC. One demonstration of the application of IRT in a clinical trial can be found in Glas, et al. [43]. They successfully applied IRT to tender *point* counts in fibromyalgia. They showed that tender point counts of patients diagnosed with fibromyalgia had a good fit with IRT models, and that items could be removed without facing a substantial loss of power. Our study extended this application to clinical measures by applying IRT to a TJC in patients with early-stage RA. Future studies could investigate whether a modified or shorter TJC will perform equally well or perhaps even better than the TJC28.

In contrast to CTT, IRT information curves can be obtained when applying IRT to the data. This provides insight into the performance of the total TJC28 and of the individual joints, and exposes opportunities for scale improvement. The covered range of joint difficulties demonstrated that the TJC28 mainly functions along the moderate and higher spectrum of disease activity. The test and item information functions also showed that θ is measured with the greatest precision for patients with a higher degree of joint tenderness, especially with joint tenderness in the smaller joints. This spectrum limitation was caused by the low number of painful joints experienced by the sample of patients with early-stage RA. Since the cohort we used represented a large sample size, and since it included patients from 6 hospitals from different regions in The Netherlands, it is expected that this cohort is representative of the patients with early-stage RA. However, to further examine the measurement precision of the TJC28 and to make the results more generalizable, future research should expand our study by applying IRT to RA samples with a longer disease duration.

The rationale concerning which joints to include in a joint count has not yet been clearly outlined in the literature. The joints included in the TJC28 were selected based on pragmatic logistic considerations and clinical experience [20]. Although the TJC28 appears to be a reliable and valid instrument to assess joint tenderness, it does not include the feet and ankle joints. There have been several discussions about whether the feet and ankles really can be omitted from the instrument [44, 45]. It has been argued that the 28-joint count might be useful in clinical trials, but that a more comprehensive joint count that includes the foot joints might be preferable for following the disease progress of patients in daily clinical practice [13, 19]. IRT may provide clarity in this discussion, since IRT provides an opportunity to evaluate the contribution of each single foot joint and ankle joint [24]. The joints differ in the degree of information they provide, shown by the inequality in the parameter values. This means the joints contribute unequally to the precision of measurement. By evaluating whether foot and ankle joints provide any significant information, it can be decided whether they truly can safely be omitted from the joint count. Future research should apply IRT to more extensive joint counts, such as the TJC68, to examine which joints provide important information to the instrument and should be included, and which joints provide limited information and can therefore be omitted from the joint count.

Our study confirmed that the TJC28 has good internal and acceptable external construct validity for patients with early-stage RA. However, the IRT analyses also pointed to some potential limitations of the instrument - a major problem being its limited measurement range. Since test information was low for lower levels of disease activity, it might be appropriate to modify the TJC28 to improve its measurement precision and range, for instance by expanding the TJC with joints that provide more information at the lower levels of disease activity. It is recommended that future studies examine both the TJC28 and more extensive joint indices in RA samples with a longer disease duration to confirm our findings and to explore possibilities for further improvements of the TJC.

Acknowledgements

The authors thank the respondents who participated in our study and the rheumatologists from the Medisch Spectrum Twente (Enschede), Ziekenhuisgroep Twente (Almelo, Hengelo), Isala Klinieken (Zwolle), Universitair Medisch Centrum Groningen (Groningen), UMC St Radboud (Nijmegen) and TweeSteden Ziekenhuis (Tilburg) for their assistance in patient recruitment and data collection.

References

1. Gonzalez A, Maradit Kremers H, Crowson CS, Nicola PJ, Davis JM, Therneau TM, et al. The widening mortality gap between rheumatoid arthritis patients and the general population. Arthritis Rheum. 2007;56(11):3583-7.
2. Tobón GJ, Youinou P, Saraux A. The environment, geo-epidemiology, and autoimmune disease: Rheumatoid arthritis. J Autoimmun. 2010;35:10-4.
3. Turkiewicz AM, Moreland LW. Rheumatoid arthritis. In: Bartlett SJ, editor. Clinical care in the rheumatic diseases. Atlanta (GA): Association of Rheumatology Health Professionals; 2006. p. 157-66.
4. Arnett FC, Edworthy SM, Bloch DA, McShane DJ, Fries JF, Cooper NS, et al. The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. Arthritis Rheum. 1988;31:315-24.
5. Helliwell PS, Hetthen J, Sokoll K, Green M, Marchesoni A, Lubrano E, et al. Joint symmetry in early and late rheumatoid and psoriatic arthritis. Arthritis Rheum. 2000;43:865-71.
6. Abramson JH. On the diagnostic criteria of active rheumatoid arthritis. J Chron Dis. 1967;20:275-90.
7. Ropes MW, Bennett GA, Cobb S, Jacox R, Jessar RA. Proposed diagnostic criteria for rheumatoid arthritis. Ann Rheum Dis. 1957;16:118-25.
8. Aletaha D, Smolen JS. Remission of rheumatoid arthritis: Should we care about definitions? Clin Exp Rheumatol. 2006;24:S45-S51.
9. Scott DL, Houssien DA. Joint assessment in rheumatoid arthritis. Br J Rheumatol. 1996;35:14-8.
10. Pala O, Cavaliere LF. Joint counts. In: Bartlett SJ, editor. Clinical care in the rheumatic diseases. Atlanta (GA): Association of Rheumatology Health Professionals; 2006. p. 39-41.
11. Felson DT, Anderson JJ, Boers M, Bombardier C, Chernoff M, Fried B, et al. The American College of Rheumatology preliminary core set of disease activity measures for rheumatoid arthritis clinical trials. The Committee on Outcome Measures in Rheumatoid Arthritis Clinical Trials. Arthritis Rheum. 1993;36(6):729-40.
12. Pinals RS, Masi AT, Larsen RA. Preliminary criteria for clinical remission in rheumatoid arthritis. Arthritis Rheum. 1981;24(10):1308-15.
13. 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.
14. Prevoo MLL, van Riel PLCM, van 't Hof MA, van Rijswijk MH, van Leeuwen MA, Kuper HH, et al. Validity and reliability of joint indices. A longitudinal study in patients with recent onset rheumatoid arthritis. Br J Rheumatol. 1993;32:589-94.
15. Prevoo MLL, van ‘t Hof MA, Kuper HH, van Leeuwen MA, van de Putte LBA, van Riel PLCM. Modified disease activity scores that include twenty-eight-joint counts. Arthritis Rheum. 1995;38:44-8.
16. Tennant A, Conaghan PG. The rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a rasch paper? Arthritis Rheum. 2007;57:1358-62.
17. Hays RD, Morales LS, Reise SP. Item response theory and health outcome measurement in the 21st century. Med Care. 2000;38:II28-II42.
18. Vermeer M, Kuper H, Hoekstra M, Bernelot Moens H, Van Riel P, van de Laar M. Remission in daily clinical practice: excellent results after one year of tight control in very early rheumatoid arthritis, results of the DREAM remission induction cohort. Ann Rheum Dis. 2010;69:506.

19. Van Riel PLCM, Fransen J, Scott DL. Eular handbook of clinical assessments in rheumatoid arthritis. Alphen aan den Rijn: van Zuiden Communications; 2004.
20. Fuchs HA, Brooks RH, Callahan LF, Pincus T. A simplified twenty-eight-joint quantitative articular index in rheumatoid arthritis. Arthritis Rheum. 1989;32:531.
21. Fries JF, Spitz P, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. Arthritis Rheum. 1980;23:137-45.
22. Bruce B, Fries JF. The stanford health assessment questionnaire: a review of its history, issues, progress, and documentation. J Rheumatol. 2003;30:167-78.
23. Van den Berg SM, Glas CAW, Boomsma DI. Variance decomposition using an IRT measurement model. Behav Genet. 2007;37:604-16.
24. Hambleton RK, Swaminathan H, Rogers HJ. Fundamentals of item response theory. Newbury Park (CA): Sage Publications; 1991.
25. Baker FB. The basics of item response theory. College Park (MD): ERIC Clearinghouse on Assessment and Evaluation; 2001.
26. Glas CAW. Modification indices for the 2-pl and the nominal response model. Psychometrika. 1999;64:273-94.
27. Te Marvelde JM, Glas CAW, van Landeghem G, van Damme J. Application of multidimensional item response theory models to longitudinal data. Educ Psychol Meas. 2006;66:5-34.
28. van Groen MM, ten Klooster PM, Taal E, van de Laar MAFJ, Glas CAW. Application of the health assessment questionnaire disability index to various rheumatic diseases. Qual Life Res. 2010:1-9.
29. Scheerens J, Glas CAW, Thomas SM. Educational evaluation, assessment, and monitoring. A systematic approach. Lisse: Swets & Zeitlinger; 2003.
30. Glas CAW, Verhelst ND. Testing the Rasch model. In: Fischer GH, Molenaar IW, editors. Rasch models Their foundations, recent developments and applications. New York (NY): Springer; 1995. p. 69-96.
31. Reeve BB, Fayers P. Applying item response theory modelling for evaluating questionnaire item and scale properties. In: Fayers PM, Hays RD, editors. Assessing Quality of Life in Clinical Trials: Methods and Practice. Oxford: Oxford University Press; 2005. p. 55 - 73.
32. Van Riel PLCM. Provisional guidelines for measuring disease activity in clinical trials on rheumatoid arthritis. Br J Rheumatol. 1992;31:793-4.
33. Browne JP, van der Meulen JH, Lewsey JD, Lamping DL, Black N. Mathematical coupling may account for the association between baseline severity and minimally important difference values. J Clin Epidemiol. 2010;63:865-74.
34. El Miedany Y, Youssef SS, El Gaafary M. Short-term outcome after anti-tumor necrosis factor-α therapy in rheumatoid arthritis: Do we need to revise our assessment criteria? J Rheumatol. 2006;33:490-6.
35. Sokka T. Assessment of pain in rheumatic diseases. Clin Exp Rheumatol. 2005;23:S77-S84.
36. Plant MJ, O'Sullivan MM, Lewis PA, Camilleri JP, Coles EC, Jessop JD. What factors influence functional ability in patients with rheumatoid arthritis. Do they alter over time? Rheumatology. 2005;44:1181-5.
37. Van Leeuwen MA, van der Heijde DMFM, van Rijswijk MH, Houtman PM, van Riel PLCM, van de Putte LBA, et al. Interrelationship of outcome measures and process variables in early rheumatoid arthritis. A comparison of radiologic damage, physical disability, joint counts, and acute phase reactants. J Rheumatol. 1994;21:425-9.
38. Dwyer KA, Coty MB, Smith CA, Dulemba S, Wallston KA. A comparison of two methods of assessing disease activity in the joints. Nurs Res. 2001;50:214-21.

39. Filippucci E, Iagnocco A, Salaffi F, Cerioni A, Valesini G, Grassi W. Power Doppler sonography monitoring of synovial perfusion at the wrist joints in patients with rheumatoid arthritis treated with adalimumab. Ann Rheum Dis. 2006;65:1433-7.
40. Baan H, Hoekstra M, Veehof M, van de Laar M. Ultrasound findings in rheumatoid wrist arthritis highly correlate with function. Disabil Rehabil. 2010:1-5.
41. Thompson PW, Hart LE, Goldsmith CH, Spector TD, Bell MJ, Ramsden MF. Comparison of four articular indices for use in clinical trials in rheumatoid arthritis: Patient, order and observer variation. J Rheumatol. 1991;18(5):661-5.
42. 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(4):716-7.
43. Glas CAW, Geerlings H, van de Laar MAFJ, Taal E. Analysis of longitudinal randomized clinical trials using item response models. Contemp Clin Trials. 2009;30:158-70.
44. Landewé R, van der Heijde D, van der Linden S, Boers M. Twenty-eight-joint counts invalidate the DAS28 remission definition owing to the omission of the lower extremity joints: A comparison with the original DAS remission. Ann Rheum Dis. 2006;65:637-41.
45. Van der Leeden M, Steultjens MPM, Ursum J, Dahmen R, Roorda LD, van Schaardenburg D, et al. Prevalence and course of forefoot impairments and walking disability in the first eight years of rheumatoid arthritis. Arthritis Rheum. 2008;59:1596-602.

Chapter 4

Contribution of assessing forefoot joints in early rheumatoid arthritis patients

Insights from item response theory

L. Siemons
P.M. ten Klooster
E. Taal
H.H. Kuper
P.L.C.M. van Riel
C.A.W. Glas
M.A.F.J. van de Laar

Arthrit Care Res. 2013, 65(2): 212-219

Abstract

Objective: To evaluate the contribution of assessing forefoot joints to the measurement range and measurement precision of joint counts in early rheumatoid arthritis (RA) using item response theory.

Methods: Baseline measures of tender and swollen joint counts were analyzed in 459 early RA patients from the Dutch Rheumatoid Arthritis Monitoring remission induction cohort. The contribution of forefoot joints was studied by evaluating their effect on the measurement range and measurement precision of measures based on 28-joint counts. In addition, the alignment between the patient and joint distributions was investigated to determine whether the forefoot joints were informative for measuring joint tenderness or swelling of an early RA patient.

Results: In total, 233 patients (50.76%) experienced tenderness and 200 patients (43.57%) experienced swelling in ≥ 1 forefoot joint. Forefoot joints were more informative for measuring joint tenderness than joint swelling, but did not significantly improve the measurement range and measurement precision of the 28-joint counts. Furthermore, including forefoot joints did not remove the existing discrepancy between the joint and patient distributions in both joint counts.

Conclusion: Forefoot joints were frequently affected on an individual level, but did not significantly improve the measurement range or precision of 28-joint counts in patients with early RA. From a measurement perspective, reduced joint counts are appropriate for use on a population level. The contribution of assessing forefoot joints on an individual level requires further investigation. Additionally, the results should be cross-validated in patients with longer disease durations to determine whether the pattern of joint involvement is similar in later stages of RA.

Introduction

The Disease Activity Score for 28 joints (DAS28) [1] is a widely used index measure for assessing the disease activity of individual rheumatoid arthritis (RA) patients and for evaluating treatment effectiveness aimed at reaching or sustaining a state of remission. However, despite the frequent involvement of the 28 joints included in the DAS28, there are data to support that the omission of the forefoot joints causes the DAS28 to underestimate actual disease activity in early RA patients predominantly with disease activity in the feet [2]. Additionally, the DAS28 remission criteria might not reflect a true state of remission, due to the presence of residual tenderness or swelling in omitted joints [3]. Consequently, the cut off point for DAS28 remission often has been criticized [3, 4].

From a measurement perspective, the high prevalence rates of painful and swollen forefoot joints [5, 6] might provide relevant additional quantitative clinical information for assessing and monitoring the status of patients with RA. However, practical constraints, extra administration time, assessment difficulty, and the recognition that abnormalities in the feet may often result from processes other than RA are all reasons for their frequent exclusion from the joint counts [7-10]. Furthermore, various studies have shown that reduced joint counts that exclude the forefoot joints appear to be as reliable and valid as more comprehensive joint counts [1, 5, 7, 11]. Nevertheless, the exclusion of the forefoot joints remains a topic of debate and research.

When approaching this debate from a measurement perspective, most studies used methods from the classical test theory to examine the contribution of forefoot joints, for instance, by examining the correlation between reduced and extended joints counts [5]. However, classical test theory does not provide insight into the effect the inclusion of forefoot joints has on the measurement range and measurement precision of the total joint count, an insight that can be obtained by applying a different measurement perspective, called item response theory (IRT). Therefore, the aim of this study was to evaluate the contribution of assessing forefoot joints in patients with early RA using IRT.

Patients and methods

Patients

Patients participated in the Dutch Rheumatoid Arthritis Monitoring (DREAM) remission induction cohort [12], a multicenter cohort that was started in 2006 to evaluate the effect of a protocolized treatment strategy aiming at a state of remission in early RA patients in daily clinical practice. Patients were eligible for inclusion at the moment they were clinically diagnosed with RA, if they were ages ≥18 years, and if they had never received disease-modifying antirheumatic drugs or prednisolone before. The ethics committee of each participating hospital evaluated the study protocol. Since the study data were gathered during daily clinical practice, the ethics committees determined no approval was required, which is in accordance with the Dutch Law. Nevertheless, informed consent was obtained from each patient.

Measures

Disease activity was assessed by a trained rheumatologist or nurse practitioner at each visit using extensive joint counts, including 44 joints for the measurement of joint tenderness and joint swelling, the erythrocyte sedimentation rate (ESR), the C-reactive protein (CRP) level, and a 100-mm visual analog scale (VAS) for general health. The 28- and 38-joint counts for tenderness and swelling, used for analyzing the contribution of the forefoot joints, were derived from these 44-joint counts. The DAS28 score was computed with the 28 tender joint count (TJC28), 28 swollen joint count (SJC28), ESR, and VAS for general health scores. Only baseline measures were used for analyses.

The 28-joint counts on tenderness and swelling include the shoulder (n=2), elbow (n=2), wrist (n=2), metacarpophalangeal (MCP; n=10), proximal interphalangeal (PIP; n=10), and knee (n=2) joints. The 38-joint counts also include the 10 metatarsophalangeal (MTP) joints of the feet. All of the joints were scored dichotomously.

The Boolean-based American College of Rheumatology (ACR)/European League Against Rheumatism (EULAR) definition of RA remission was used to determine remission, since this criterion can be used with both the 28- as well as the 38-joint counts [13].

Statistical methods

To analyze the contribution of the forefoot joints, an IRT model (the generalized partial credit model [GPCM]) [14] was applied. Both the 28- and 38-joint counts of tenderness and swelling had to fit the GPCM before the contribution of the forefoot joints could be analyzed. This fit was analyzed with Lagrange Multiplier Q1 tests [15], where absolute effect sizes of <0.10 were seen as an indication of good item-model fit [16, 17].

Next, the reliability and corresponding measurement precision of the joint counts with and without forefoot joints were evaluated and compared. IRT global reliability is equivalent to Cronbach's alpha, but based on IRT scores (theta) instead of raw scores [18]. The theta scores are scaled around 0 and correspond to the degree of joint tenderness or swelling a patient experiences. Reliabilities of >0.7 were considered sufficient for group use, whereas values of >0.85 were deemed necessary for individual use [19].

Local reliability and measurement precision of both the total joint counts as well as their individual joints were derived from 2 types of information curves [18, 20]. First, the test information curves (TICs) were investigated to determine the range on the underlying scale where the total joint counts can reliably (precisely) measure a patient's degree of joint tenderness or swelling, where reliability = 1 - (1/test information at Θ) and the precision of the estimated theta is the reciprocal of the test information ($\mathrm{Var}(\Theta)$ = 1/test information at Θ) [18]. TICs of joint counts with and without forefoot joints were compared to determine the effect of including forefoot joints on the measurement precision and measurement range of the total instrument. A TIC is a sum of the individual item information curves (IICs). These show the contribution of the individual joints to the estimation of joint tenderness or swelling. IICs of the forefoot joints were compared to the IICs of the other joints to determine their individual contribution to the total instrument and to evaluate for which patients they provide the most information.

Finally, the alignment between the patient and the joint distributions was evaluated. Ideally, the mean scores of these distributions should be relatively close to each other and the distributions should approximately cover the same range of the scale. Joints measuring outside the range of the patient distribution are less informative about the patients' joint tenderness or swelling than joints measuring inside this range.

IRT analyses were performed with the statistical program MIRT [21]. Full item parameter calibrations are available from the corresponding author upon request. Patient and joint distributions were plotted with SPSS, version 18.0.

Results

Patient characteristics

The cohort included a total of 459 patients for analysis, predominantly consisting of women (62.31%). Most had active disease at inclusion, with a mean ± SD DAS28 score of 4.69 ± 1.31, and they experienced their general health as rather low, with a mean ± SD score of 49.83 ± 25.03 (Table 1). On average, patients had 6 tender and 7 swollen joints on the 28-joint counts, which increased to 8 tender and 9 swollen joints on the 38-joint counts. Forefoot joints were commonly affected, with 233 patients (50.76%) experiencing tenderness and 200 patients (43.57%) experiencing swelling in ≥1 of the forefoot joints. Based on both the 28- and the 38-joint counts, 7 patients were in remission according to the new Boolean-based ACR/EULAR definition of remission. None of the remission patients experienced tenderness or swelling in their forefoot joints.

Table 1 – Baseline characteristics of the 459 rheumatoid arthritis patients included in the cohort*.

Variable	Value
Age, years	57.76 ± 14.70
Female, no. (%)	286 (62.31)
DAS28 score	4.69 ± 1.31
TJC28	5.60 ± 5.55
TJC38	7.90 ± 7.26
SJC28	7.39 ± 5.60
SJC38	9.15 ± 6.73
VAS for general health, mm	49.83 ± 25.03
ESR, mm/hour	29.16 ± 20.89
CRP level, mg/l	21.11 ± 35.80
≥ 1 tender MTP joint, no. (%)	233 (50.76)
≥ 1 swollen MTP joint, no. (%)	200 (43.57)
> 1 tender MTP joint, no. (%)	196 (42.70)
> 1 swollen MTP joint, no. (%)	170 (37.04)

** Values are the mean ± SD unless otherwise indicated. DAS28 = Disease Activity Score in 28 joints, TJC = tender joint count, SJC = swollen joint count, VAS = visual analog scale, ESR = erythrocyte sedimentation rate, CRP = C-reactive protein, MTP = metatarsophalangeal*

Fit and global reliability

The results showed a good fit to the GPCM for both the 28- and 38-joints counts, since all of the effect sizes were well below 0.10 (Table 2).

Table 2 - Fit of the joint counts to the generalized partial credit model and associated global reliability levels*.

Instrument	Reliability	Maximum effect size of Lagrange Multiplier Q1 tests over items
TJC28	0.827	0.03
SJC28	0.860	0.04
TJC38	0.861	0.06
SJC38	0.869	0.04

** TJC = tender joint count, SJC = swollen joint count*

All of the joint counts showed reliabilities acceptable for individual use (*r*>0.85) except for the TJC28, which was sufficiently reliable for group use, but slightly below the level for individual use. Reliability of the 28-joint counts only marginally increased when the forefoot joints were added to the instrument.

Local reliability and measurement range

Both the TJC28 and the SJC28 only measured reliably for patients with a moderate to high degree of joint pain (*r* >0.80 for $-0.3 < \Theta < 2.8$ and $-0.7 < \Theta < 2.2$, respectively). Most information was provided by the small MCP and PIP joints, whereas larger joints contained only a limited amount of information. Although Figures 1 and 2 demonstrate that these measurement ranges became slightly broader after inclusion of the forefoot joints (TJC38: *r* >0.80 for $-0.5 < \Theta < 3.1$; SJC38: *r* >0.80 for $-0.8 < \Theta < 2.7$), the forefoot joints did not provide a high amount of information to the SJC, indicating that they do not contribute much to the estimation of joint swelling. Furthermore, in both joint counts, the forefoot joints mainly contained information along a range of the underlying scale already covered by the joints of the TJC28 and SJC28. (Graphical representations of the separate calibrations of the 28- and 38-joint counts are available from the corresponding author upon request.)

Patient-joint distributions

TJC38

Figure 3 shows a discrepancy between the patient and the joint distribution of the TJC38. Where the joint distribution spreads from 0.65 to 3.07, the patient distribution is much broader (spreading from -1.48 to 3.51). The items cluster together in a small range at the right half of the person distribution. Consequently, patients with only a minor degree of joint tenderness are not adequately measured by the included joints. The lower measurement precision for these patients also diminishes the instrument’s ability to discriminate between patients with varying degrees of joint tenderness.

The 2 joints located most to the right are the elbow joints, restricting their relevance to the small proportion of patients located in the right tail of the person distribution in particular. The forefoot joints, on the other hand, appear to be relevant for a larger proportion of the early RA sample, since they are located more in the middle of the patient distribution. However, their contribution to the measurement of disease activity is limited because they function along a range already covered by other joints of the instrument.

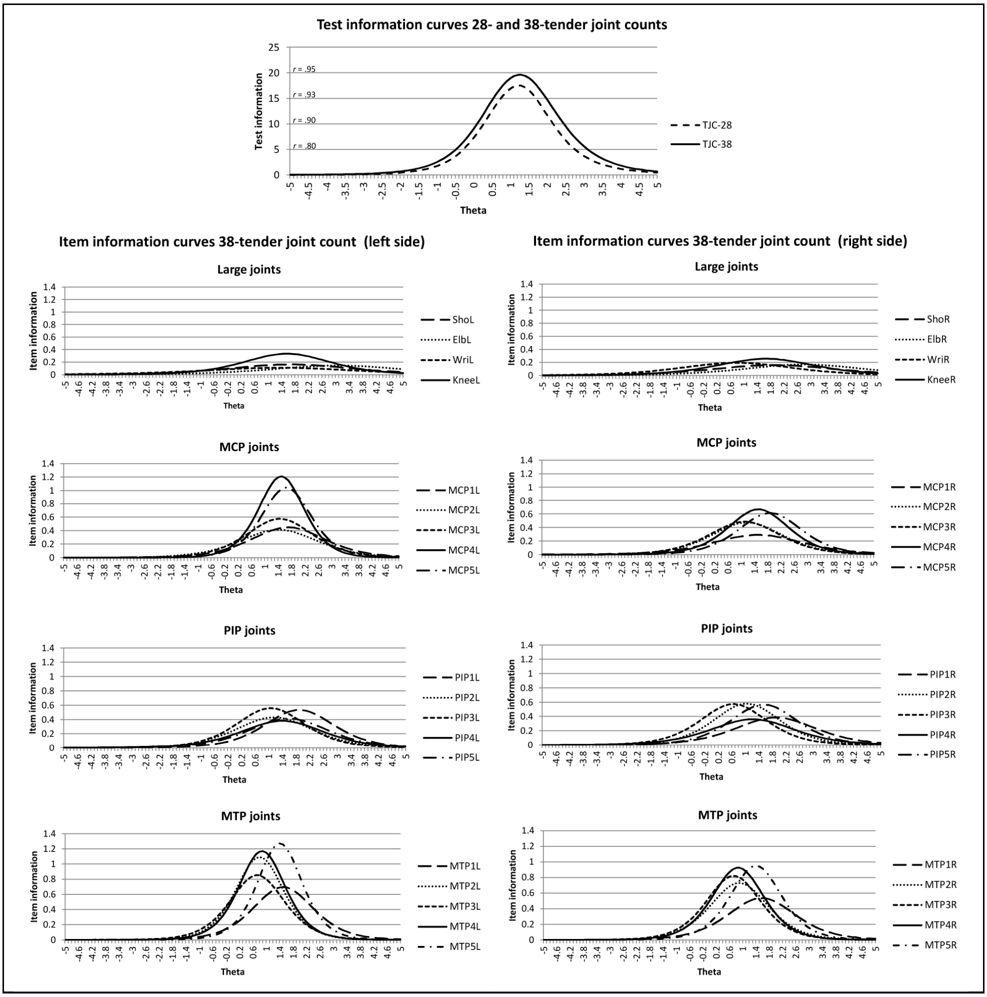


Figure 1 – Test and item information curves of the tender joint counts (TJCs). The top graph shows the test information curves of the TJC28 and TJC38 with associated local reliability levels (*r*). The bottom graphs show the item information curves of the TJC38 for the joints on the left side (left column) and right side (right column) of the body. Sho = shoulder, Elb = elbow, Wri = wrist, MCP = metacarpophalangeal, PIP = proximal interphalangeal, MTP = metatarsophalangeal.

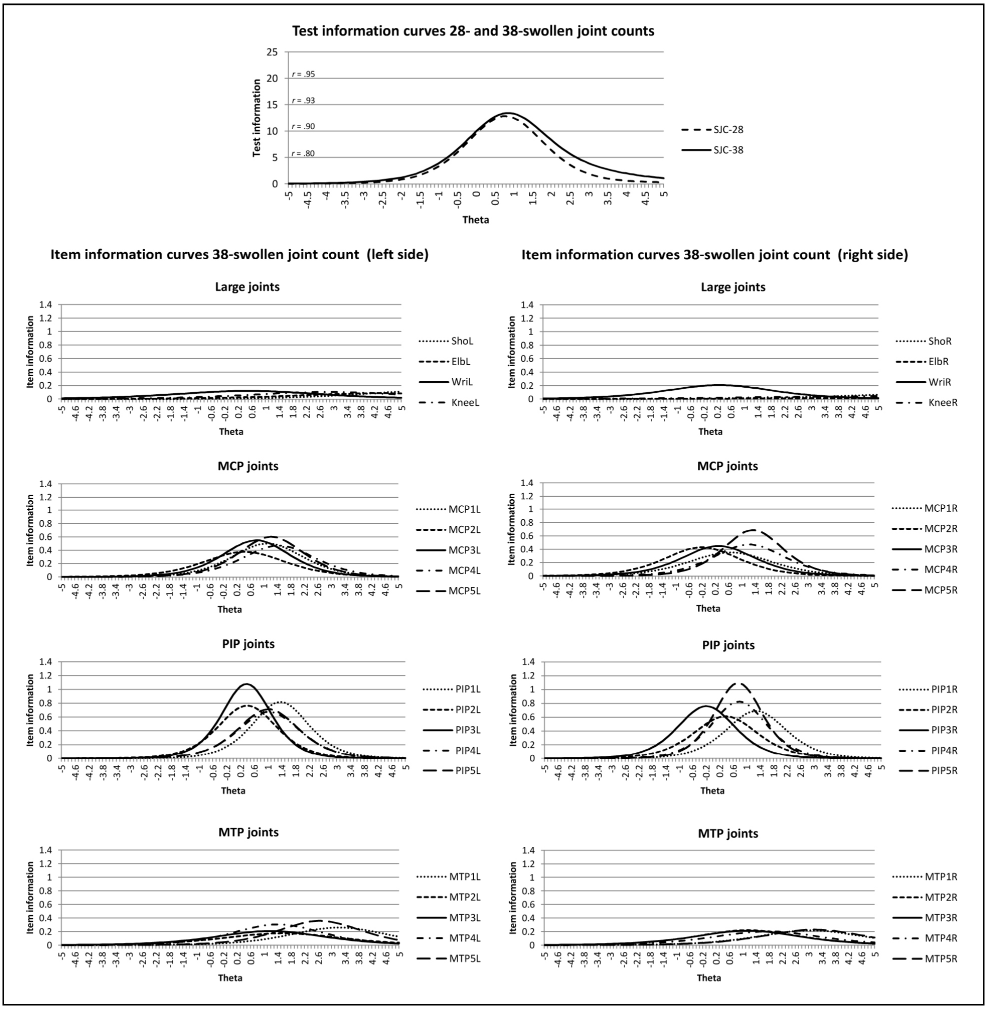


Figure 2 – Test and item information curves of the swollen joint counts (SJCs). The top graph shows the test information curves of the SJC28 and SJC38 with associated local reliability levels (r). The bottom graphs show the item information curves of the TJC38 for the joints on the left side (left column) and right side (right column) of the body. Sho = shoulder, Elb = elbow, Wri = wrist, MCP = metacarpophalangeal, PIP = proximal interphalangeal, MTP = metatarsophalangeal.

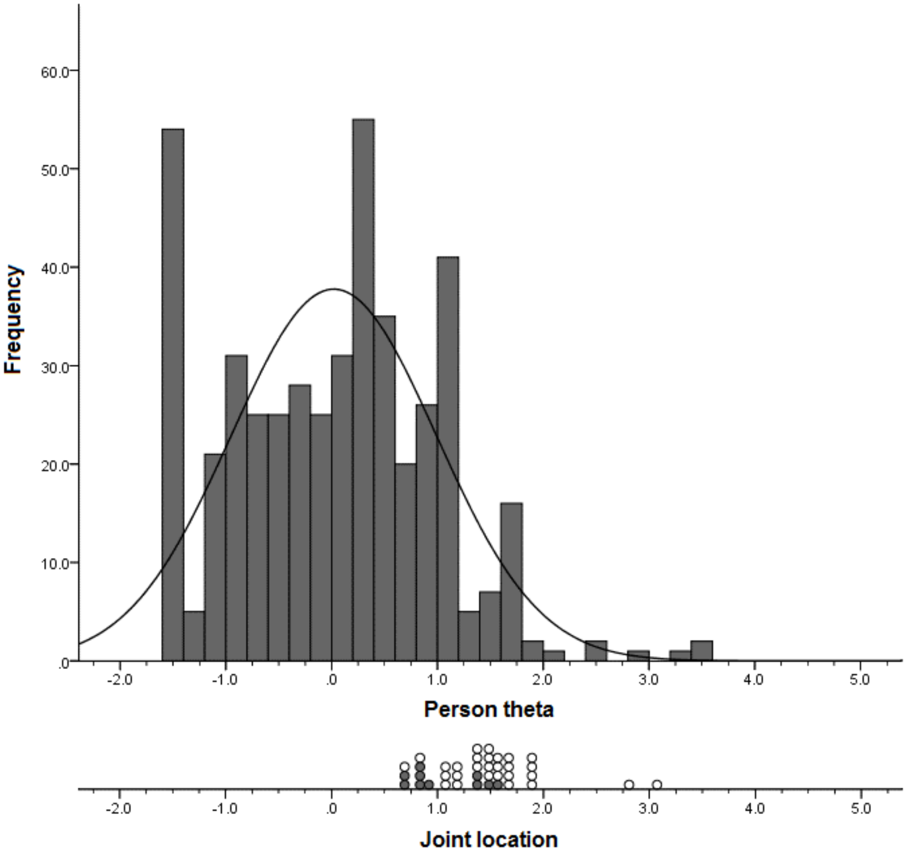


Figure 3 - Graphical representation of the patient distribution (top graph) and the joint distribution (bottom graph) of the 38 tender joint count. Shaded dots in the joint distribution show the forefoot joints. Measurement precision is optimal at the scale location of each joint.

SJC38

The SJC38 also shows a large discrepancy between the patient and joint distributions (Figure 4). The patient distribution ranges from -1.78 to 4.03, whereas the joint distribution spreads from -0.20 to 8.95. Five joints fall outside the range covered by the person distribution (elbows, shoulders and right knee), diminishing their relevance to the early RA sample. Of the 10 forefoot joints, 4 joints function along a range of the scale not yet covered by other joints. Nevertheless, inclusion of these forefoot joints is only relevant for a small proportion of the patients, since they measure most precisely for the patients located in the right tail of the person distribution.

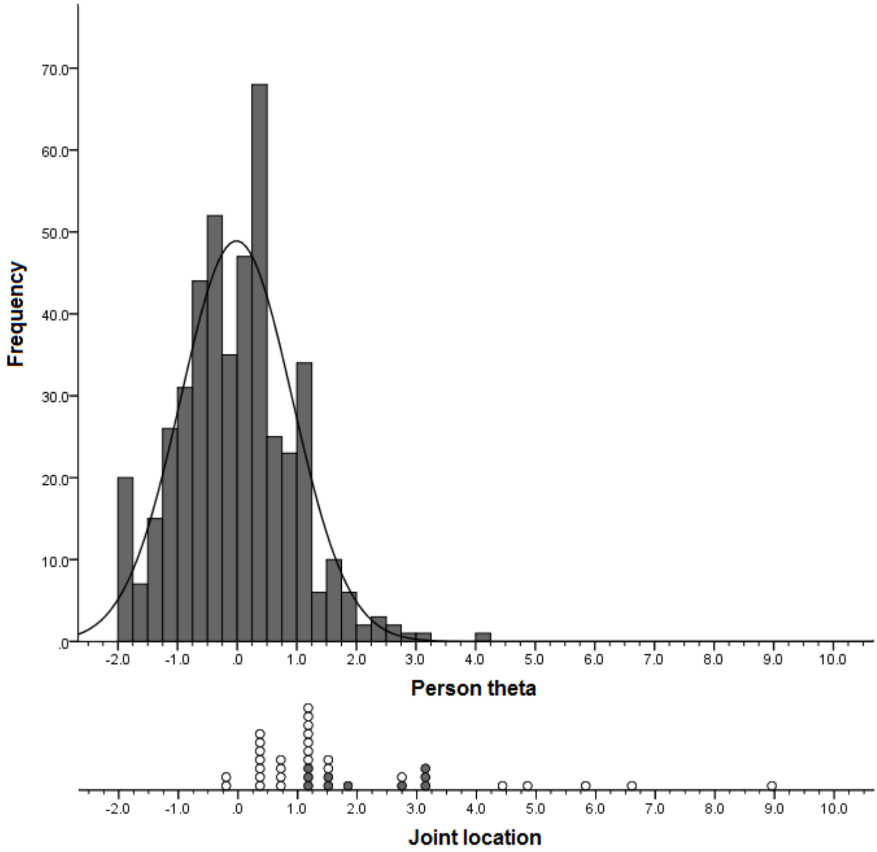


Figure 4 - Graphical representation of the patient distribution (top graph) and the joint distribution (bottom graph) of the 38 swollen joint count. Shaded dots in the joint distribution show the forefoot joints. Measurement precision is optimal at the scale location of each joint.

Discussion

The results showed no relevant effect of the inclusion of forefoot joints on a population level of patients with early RA, confirming conclusions from previous studies [8, 22]. The global reliability was already acceptable for group use when forefoot joints were excluded, the measurement range did not become significantly broader after the inclusion of forefoot joints, and the discrepancy between the joint and the patient distribution for both the TJC and the SJC remained large.

On an individual level, however, the assessment of forefoot joints might be relevant for a large proportion of the individual patients, given the high prevalence rates of affected forefoot joints in the early RA sample, which was also found by van der Leeden et al [6]. Since the anchor remission is rapidly becoming more important within rheumatology, these results show the importance of including forefoot joints. Moreover, researchers or clinicians interested in tracking the disease course of an individual patient might want to include the forefoot joints in both the TJC as well as the SJC, since these joints can contain important information about and might give a more complete image of the development of the patient's disease [23]. These results are consistent with earlier findings emphasizing the importance of forefoot joints on an individual level [2, 22]. Nevertheless, the information curves did show a large discrepancy between the forefoot joints' information value in the TJC and SJC. Where the forefoot joints appeared to be informative for the TJC, their information value for the SJC was almost negligible. These results suggest that inclusion of the forefoot joints might be more useful for measuring an individual patient's degree of joint tenderness than for measuring his or her degree of joint swelling. The limited information value of forefoot joints in the SJC might be explained by the clinical experience that the assessment of swelling in the forefoot joints is more difficult than in other joints [8]. Nevertheless, the results did show that 4 forefoot joints of the swollen joint count functioned along a range of the scale not yet covered by other joints. Although, inclusion of these forefoot joints was only relevant for a small proportion of the patients, diminishing its relevance on a population level, it might have significant implications on response criteria and early classification of individual patients. On the other hand, the joints of the elbows, shoulders, and right knee fall outside the range covered by the person distribution, casting doubt on their relevance for the early RA sample. Future work could focus more on the relevance of these large joints that are included in the reduced joint counts and on the relevance of the forefoot joints on an individual level.

This study gives new methodological support to earlier research showing that the reduced 28-joint counts can be useful for assessing baseline disease activity at a population level (e.g. as indicators of hospital performance or in clinical trials), but that more extensive joint counts might be preferable for following the disease course of the patients in daily clinical practice [5, 23]. Since foot joints, particularly the MTP joints, are commonly affected in RA [5], the main interest of this study was in the effect of the inclusion of forefoot joints on the measurement range and measurement precision of the joint count. Data were derived from 44-joint counts of tenderness and swelling, and consequently, other foot joints such as the tarsometatarsal joints and the tarsal joints within the midfoot, were not included in this study. To make study results comparable to and consistent with the study of van Tuyl et al [8], the ankle joints were excluded as well and analyses were based on 38-joint counts. If the outcome of interest is the whole foot, however, future work should also focus on investigation of midfoot and hindfoot joint swelling and tenderness.

A limitation of this study is that, where previous studies evaluated misclassifications of RA disease activity due to the omission of forefoot joints [2, 8], this could not be evaluated in this study because only baseline measures were used with almost no patients in remission. Data from followup measurements must be analyzed in future research to determine the impact of the inclusion or exclusion of forefoot joints on RA disease activity classifications.

Another reason for analyzing followup measurements is to examine whether the results of this study are also applicable beyond patients with a recent diagnosis. The pattern of joint involvement at diagnosis might not necessarily resemble the pattern at later measurement points or posttreatment. Consequently, the results might not apply equally to patients in a later stage of RA treatment.

All participating patients were clinically diagnosed with RA. The finding that 7 of these patients were already in a state of remission according to the ACR/EULAR criteria despite having symptoms that led to a diagnosis of RA can be explained by the use of different classification instruments for the clinical diagnosis of RA and the determination of remission. To be classified as being in a state of remission, the Boolean-based ACR/EULAR criteria were used [13], which means that a patient had to have TJC, SJC, CRP level (in mg/dl), and patient global assessment (on a scale of 0-10) scores that were all ≤1. For a clinical diagnosis of RA, on the other hand, the 2010 ACR/EULAR classification criteria for RA were used [24]. According to these criteria, RA is present when the patient experiences synovitis in at least 1 joint that cannot be explained any better by an alternative diagnosis. In addition, a total score of 6 or higher (with a maximum of 10) has to be obtained on the number and site of involved joints (score 0-5), the serologic abnormality (score 0-3), the symptom duration (score 0-1), and the elevated acute-phase response (score 0-1).

In conclusion, forefoot joints were frequently affected on an individual level, but the inclusion of forefoot joints did not significantly improve the measurement range or measurement precision of the TJC and SJC in patients with early RA. From a measurement perspective, reduced joint counts are appropriate to use on a population level. The contribution of assessing forefoot joints on an individual level could be important clinically, and requires further investigation. Additionally, the applicability of these study results should be examined beyond patients with a new diagnosis to determine whether the pattern of joint involvement at diagnosis resembles the pattern at later stages of RA.

Acknowledgements

The authors would like to thank all of the patients for their participation in this study and the rheumatologists from the Medisch Spectrum Twente (Enschede), Ziekenhuisgroep Twente (Almelo, Hengelo), Universitair Medisch Centrum Groningen (Groningen), Isala Klinieken (Zwolle), TweeSteden Ziekenhuis (Tilburg), and UMC St. Radboud (Nijmegen) for patient recruitment and data collection.

References

1. Prevoo MLL, van 't Hof MA, Kuper HH, van Leeuwen MA, van de Putte LBA, van Riel PLCM. Modified disease activity scores that include twenty-eight-joint counts. Arthritis Rheum. 1995;38:44-8.
2. Bakker MF, Jacobs JWG, Kruize AA, van der Veen MJ, van Booma-Frankfort C, Vreugdenhil SA, et al. Misclassification of disease activity when assessing individual patients with early rheumatoid arthritis using disease activity indices that do not include joints of feet. Ann Rheum Dis. 2011;In Press.
3. Landewé R, van der Heijde D, van der Linden S, Boers M. Twenty-eight-joint counts invalidate the DAS28 remission definition owing to the omission of the lower extremity joints: A comparison with the original DAS remission. Ann Rheum Dis. 2006;65:637-41.
4. Mäkinen H, Kautiainen H, Hannonen P, Sokka T. Is DAS28 an appropriate tool to assess remission in rheumatoid arthritis? Ann Rheum Dis. 2005;64:1410-3.
5. 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.
6. Van der Leeden M, Steultjens MPM, Ursum J, Dahmen R, Roorda LD, van Schaardenburg D, et al. Prevalence and course of forefoot impairments and walking disability in the first eight years of rheumatoid arthritis. Arthritis Rheum. 2008;59:1596-602.
7. Fuchs HA, Brooks RH, Callahan LF, Pincus T. A simplified twenty-eight-joint quantitative articular index in rheumatoid arthritis. Arthritis Rheum. 1989;32:531.
8. van Tuyl LHD, Britsemmer K, Wells GA, Smolen JS, Zhang B, Funovits J, et al. Remission in early rheumatoid arthritis defined by 28 joint counts: limited consequences of residual disease activity in the forefeet on outcome. Ann Rheum Dis. 2012;71(1):33-7.
9. Thompson PW, Kirwan JR. Joint count: A review of old and new articular indices of joint inflammation. Br J Rheumatol. 1995;34:1003-8.
10. Fuchs HA, Pincus T. Reduced joint counts in controlled clinical trials in rheumatoid arthritis. Arthritis Rheum. 1994;37(4):470-5.
11. Prevoo MLL, van Riel PLCM, van 't Hof MA, van Rijswijk MH, van Leeuwen MA, Kuper HH, et al. Validity and reliability of joint indices. A longitudinal study in patients with recent onset rheumatoid arthritis. Br J Rheumatol. 1993;32(7):589-94.
12. Vermeer M, Kuper HH, Hoekstra M, Haagsma CJ, Posthumus MD, Brus HLM, et al. Implementation of a treat-to-target strategy in very early rheumatoid arthritis. Arthritis Rheum. 2011;63(10):2865-72.

13. Felson DT, Smolen JS, Wells G, Zhang B, van Tuyl LHD, Funovits J, et al. American College of Rheumatology/European League Against Rheumatism provisional definition of remission in rheumatoid arthritis for clinical trials. Arthritis Rheum. 2011;63(3):573-86.
14. Muraki E. A generalized partial credit model: Application of an EM algorithm. Applied Psychological Measurement. 1992;16:159-76.
15. Glas CAW. Modification indices for the 2-pl and the nominal response model. Psychometrika. 1999;64:273-94.
16. van Groen MM, ten Klooster PM, Taal E, van de Laar MAFJ, Glas CAW. Application of the health assessment questionnaire disability index to various rheumatic diseases. Qual Life Res. 2010:1-9.
17. Siemons L, ten Klooster PM, Taal E, Kuper IH, van Riel PLCM, van de Laar MAFJ, et al. Validating the 28-tender joint count using item response theory. J Rheumatol. 2011;38(12):2557-64.
18. Scheerens J, Glas CAW, Thomas SM. Educational evaluation, assessment, and monitoring. A systematic approach. Lisse: Swets & Zeitlinger; 2003.
19. Tennant A, Conaghan PG. The rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a rasch paper? Arthritis Rheum. 2007;57:1358-62.
20. Reeve BB, Fayers P. Applying item response theory modelling for evaluating questionnaire item and scale properties. In: Fayers PM, Hays RD, editors. Assessing Quality of Life in Clinical Trials: Methods and Practice. Oxford: Oxford University Press; 2005. p. 55 - 73.
21. Glas CAW. Multidimensional Item Response Theory. 2010 [cited; Available from: http://www.utwente.nl/gw/omd/afdeling/Glas/
22. Kapral T, Dernoschnig F, Machold KP, Stamm T, Schoels M, Smolen JS, et al. Remission by composite scores in rheumatoid arthritis: are ankles and feet important? Arthritis Res Ther. 2007;9(4):R72.
23. Van Riel PLCM, Fransen J, Scott DL. Eular handbook of clinical assessments in rheumatoid arthritis. Alphen aan den Rijn: van Zuiden Communications; 2004.
24. Aletaha D, Neogi T, Silman AJ, Funovits J, Felson DT, Bingham COI, et al. 2010 Rheumatoid arthritis classification criteria: An American College of Rheumatology/European League Against Rheumatism collaborative initiative. Arthritis Rheum. 2010;62(9):2569-81.

Chapter 5

Interchangeability of 28-joint disease activity scores using the erythrocyte sedimentation rate or the C-reactive protein as inflammatory marker

L. Siemons
H.E. Vonkeman
P.M. ten Klooster
P.L.C.M. van Riel
M.A.F.J. van de Laar

Clin Rheumatol. (in press)

Abstract

This paper aims to examine the interchangeability of the disease activity score in 28 joints (DAS28)-erythrocyte sedimentation rate (ESR) and DAS28-CRP scores in a diverse sample of rheumatoid arthritis (RA) patients and to evaluate generalizability over gender, age, and disease duration. A sample of 682 patients was drawn from the DREAM registry. Agreement between the two DAS28 scores was analyzed using the intraclass correlation coefficient (ICC), Bland Altman plots, and a matrix of classification agreement over DAS28 disease activity categories. Despite a strong linear correlation between the DAS28 scores and a high ICC value of 0.931, a considerable lack of individual agreement could be observed, with Bland-Altman 95% limits of agreement ranging between -0.85 and +1.25 points. On average, DAS28-CRP scores were 0.20 points lower than DAS28-ESR scores, and data stratification on age and gender showed that this systematic bias was most severe in older women (0.39 points). The overall classification agreement across DAS28 categories was 76.69%, with the agreement being lowest (35.37%) in the low disease activity group. Patients were more easily classified as being in remission when using the DAS28-CRP measure. DAS28-ESR and DAS28-CRP scores are not interchangeable within individuals. The DAS28-CRP tends to yield lower values of disease activity than the DAS28-ESR, resulting in substantial classification differences.

Introduction

The disease activity score in 28 joints (DAS28) is a widely used outcome measure for assessing disease activity in rheumatoid arthritis (RA) patients [1]. It combines information on joint tenderness and joint swelling with a marker of inflammation and a patient-reported measure of general health. The DAS28 is not only widely used in clinical trials but is also often embedded within treatment protocols to monitor patients in daily clinical practice [2, 3]. Furthermore, its use is recommended by the European League Against Rheumatism (EULAR) [4].

Although the DAS28 was originally developed with the erythrocyte sedimentation rate (ESR) as inflammatory marker (i.e. the DAS28-ESR), it has since been suggested that C-reactive protein (CRP) may be used as an equivalent, which led to the development of a separate scoring algorithm of the DAS28, the DAS28-CRP [5]. However, several previous studies demonstrated lower disease activity scores and better responses in patients assessed with the DAS28-CRP instead of the DAS28-ESR [6-9], with a possible relationship to the patient's gender, age, and disease duration [6, 7, 10-14]. These score discrepancies might lead to different interpretations of a patient's level of disease activity and, consequently, to the undesirable situation that treatment decisions depend on the chosen DAS28 algorithm.

This paper aims to examine the interchangeability of the DAS28-ESR and DAS28-CRP scores in a diverse sample of Dutch rheumatoid arthritis (RA) patients. Additionally, sub-analyses will be performed to evaluate generalizability over gender, age, and disease duration.

Methods

Patients

The DREAM registry collects data in multiple centers and cohorts throughout The Netherlands while monitoring the disease of clinically diagnosed RA patients undergoing a variety of treatment strategies. Data was drawn from two different IRB approved studies within this registry, including all patients who had a valid measure of both the DAS28-ESR and the DAS28-CRP. This resulted in a heterogeneous group of males and females of various ages (all 18 years or older), with either early RA or longstanding RA. Informed consent was obtained from each patient.

Measures of disease activity

The DAS28 scores were calculated during each hospital visit using to the following formulas [4]:

i) DAS28-ESR = $0.56 * \sqrt{\mathrm{TJC28}} + 0.28 * \sqrt{\mathrm{SJC28}} + 0.70 * \mathrm{Ln(ESR)} + 0.014 * \mathrm{GH}$

ii) DAS28-CRP = $0.56 * \sqrt{\mathrm{TJC28}} + 0.28 * \sqrt{\mathrm{SJC28}} + 0.36 * \mathrm{Ln(CRP + 1)} + 0.014 * \mathrm{GH} + 0.96$

where the TJC28 = tender joint count in 28 joints, SJC28 = swollen joint count in 28 joints, and GH = a patient-reported visual analog score of general health (on a scale of 0-100) [15]. CRP and ESR measures were determined on site, according to local standard practice.

Patients were classified into groups according to their current level of disease activity, i.e. remission if DAS28 <2.6, low disease activity if 2.6 ≤ DAS28 ≤ 3.2, moderate disease activity if 3.2 < DAS28 ≤ 5.1, and high disease activity if DAS28 >5.1 [4].

Statistical analysis

Agreement between the DAS28-ESR and DAS28-CRP was first examined with a scatter plot and the two-way random, absolute agreement, single measures intraclass correlation coefficient (ICC). Next, Bland-Altman plots [16] were made to gain more insight into the size of individual differences over the total range of DAS28 scores. A Bland-Altman plot graphs the differences between the two DAS28 scores against their mean scores [16]. The plot reflects the average degree of bias (i.e. the mean difference), together with the 95% limits of agreements (i.e. the mean score ± 1.96 * standard deviation). Besides Bland-Altman analyses on the total patient sample, sub-analyses were performed based on disease duration (< 1 year vs. ≥ 1 year), age (< 60 years vs. ≥ 60 years), and gender. Finally, classification agreement of the DAS28-ESR and DAS28-CRP across DAS28 categories (i.e. remission/low/moderate/high disease activity) was determined, as well as category-specific agreement. All analyses were performed using SPSS version 21.0.

Results

Patient characteristics at baseline

Data was collected from a sample of 682 rheumatoid arthritis patients, predominantly female (62.8%), with a mean age slightly below 60 years (57.69), and a mean disease duration of 1.51 years. Most patients did experience pain and swelling in their joints, had a diminished degree of general health and physical functioning, and showed a moderately active disease with a DAS28-ESR score of 3.88 and a DAS28-CRP score of 3.68 (Table 1).

Table 1 – Patient characteristics at baseline.

Characteristic	**Score range of measure**	**Mean (SD) Or Median (range)***
Gender (female)	-	428/682 (62.8%)
Age (years)	-	57.69 (13.85)
Body mass index (kg/m^2)	-	26.47 (4.62)
Disease duration (years)	-	0 (0-51)
DAS28-ESR	0-10	3.88 (1.61)
DAS28-CRP	0-10	3.68 (1.45)
28-Tender joint count	0-28	2 (0-28)
28-Swollen joint count	0-28	3 (0-28)
General health	0-100	41.11 (26.60)
ESR (mm/hour)	0-140	18 (1-120)
CRP (mg/l)	0-999	5 (1-158)
Pain	0-100	40.74 (26.81)
SF36 – physical health	0-100	38.20 (9.37)
SF36 – mental health	0-100	48.68 (11.46)
HAQ	0-3	0.88 (0-3)

** The values for gender are the number of patients/number of patients assessed. DAS28 = disease activity score for 28 joints, ESR = erythrocyte sedimentation rate, CRP = C-Reactive Protein, SF36 = Short Form Health Survey with 36 items, HAQ = Health Assessment Questionnaire.*

Agreement

Results showed a high ICC value of 0.931 and a strong, linear correlation between the DAS28-ESR and DAS28-CRP with a Pearson correlation coefficient of 0.945 (Figure 1). Despite this high correlation, the Bland-Altman plot showed a considerable lack of agreement between the DAS28-ESR and DAS28-CRP, with 95% limits of agreement ranging between -0.85 and +1.25 points (Figure 2). On average, DAS28-CRP scores were 0.20 points lower than DAS28-ESR scores.

The amount of bias was dependent on the mean of the two DAS28 scores (r=0.31, p<0.01), with larger discrepancies for higher levels of disease activity (i.e. mean DAS28 values > 4.0). Contrarily, for very low levels of disease activity, the DAS28-ESR tended to yield lower values than the DAS28-CRP (Figure 3).

Sub-analyses on disease duration, age and gender resulted in comparable Bland-Altman plots (Figure 4). Bias was most pronounced in RA patients with a disease duration <1 year (0.21 points compared to 0.12 points in patients with a longer disease duration) and in older women (0.39 points vs. 0.16 for younger women, -0.03 for younger men, and 0.21 for older men).

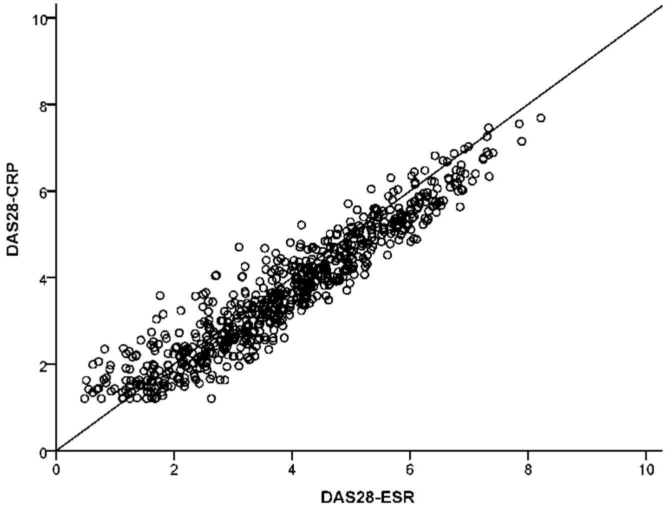


Figure 1 – The DAS28-ESR scores (x-axis) plotted against DAS28-CRP scores (y-axis). Each point corresponds to a single patient. The solid line indicates perfect agreement between the two DAS28-scores.

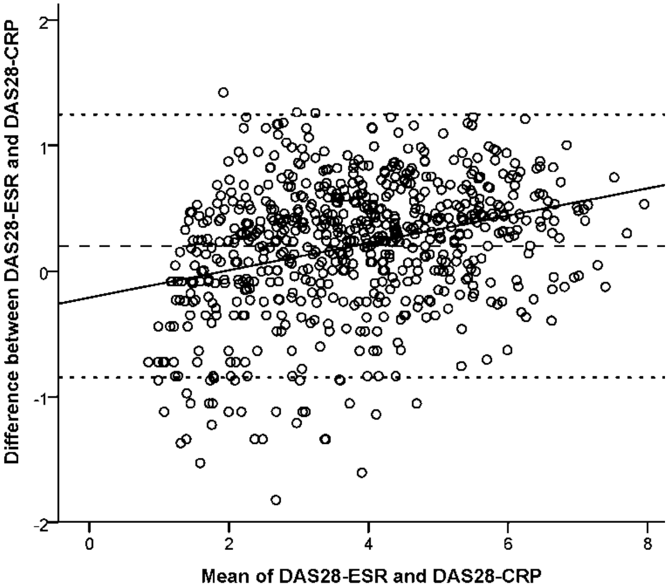


Figure 2 – Bland Altman plot of the DAS28-ESR and DAS28-CRP scores. The dashed line in the middle indicates the mean differences between both measures and the upper and lower dotted lines represent the 95% limits of agreement. The solid line shows the regression line of the average difference.

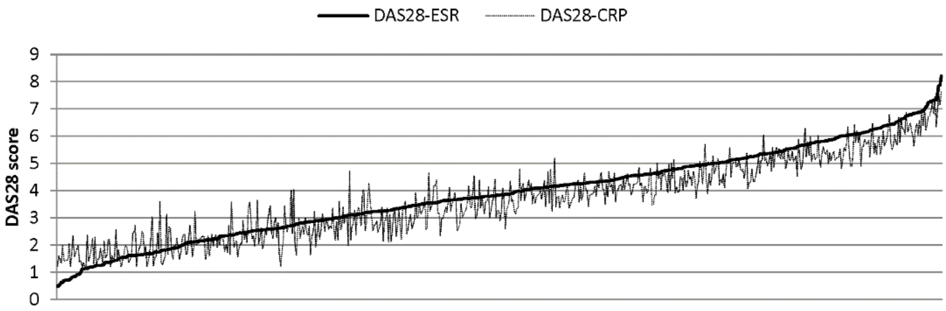


Figure 3 – Corresponding DAS28-CRP scores (thin line) to increasing DAS28-ESR scores (thick line).

Distribution disease activity groups

Overall, there was a 76.69% classification agreement across DAS28 categories. In case of disagreement, the DAS28-CRP more often yielded a lower DAS28 classification than the DAS28-ESR (120 (75.47%) vs. 39 (24.53%) times, respectively). Category-specific agreement was generally high (>79%), except for the low disease activity group, where it was only 35.37%. Patients were more easily classified as being in remission when using the DAS28-CRP (Table 2).

Table 2 – Comparison of disease activity according to the DAS28-ESR vs. DAS28-CRP.

CRP ESR	Remission (DAS28 <2.6)	Low disease activity (2.6 ≤ DAS28 ≤ 3.2)	Moderate disease activity (3.2 < DAS28 ≤ 5.1)	High disease activity (DAS28 >5.1)	Total
Remission	148 (21.70)	13 (1.91)	8 (1.17)	0 (-)	169 (24.78)
Low disease activity	31 (4.55)	26 (3.81)	13 (1.91)	0 (-)	70 (10.26)
Moderate disease activity	14 (2.05)	38 (5.57)	222 (32.55)	5 (0.73)	279 (40.91)
High disease activity	0 (-)	0 (-)	37 (5.43)	127 (18.62)	164 (24.05)
Total	193 (28.30)	77 (11.29)	280 (41.06)	132 (19.35)	682 (100)

The values correspond to the number of people in that category (%). A patient reaches remission if DAS28 <2.6, low disease activity if 2.6 ≤ DAS28 ≤ 3.2, moderate disease activity if 3.2 < DAS28 ≤ 5.1, and high disease activity if DAS28 >5.1. ESR = erythrocyte sedimentation rate, CRP = C-reactive protein.

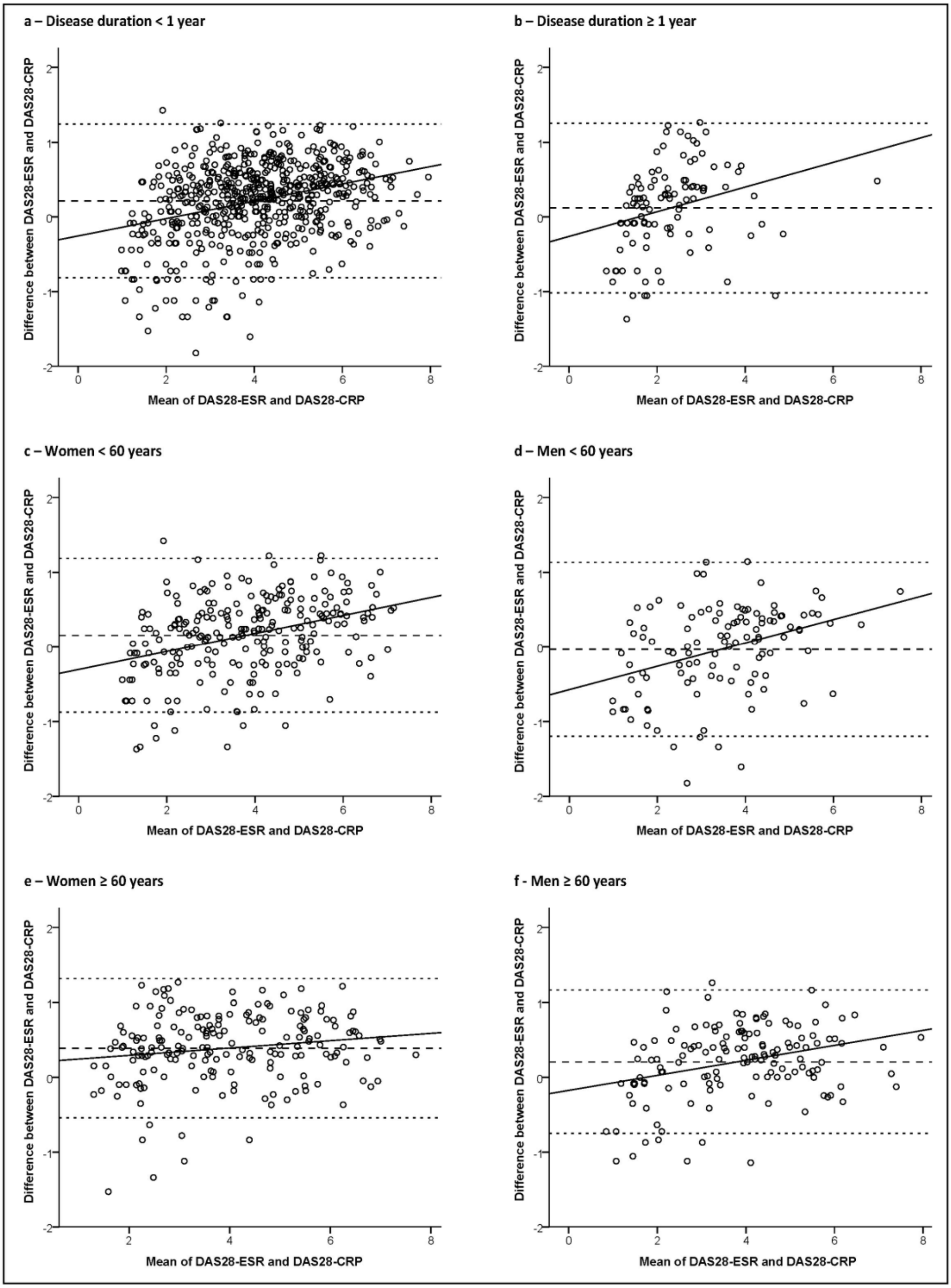


Figure 4 - Bland Altman plot analyses of the DAS28-ESR and DAS28-CRP scores, divided into groups of different disease duration (a and b) and groups of different age and gender (c-f). Within each plot, the dashed line in the middle indicates the mean differences between both measures and the upper and lower dotted lines represent the 95% limits of agreement. The solid line shows the regression line of the average difference.

Discussion

Despite the high correlation between both DAS28 scores and the “reasonably high” percentages of classification agreement over DAS28-categories, the DAS28-CRP tended to yield lower scores than the DAS28-ESR. These findings are consistent with previous studies [6-9] and emphasize the need for awareness of the score discrepancies between these two measures in order to improve standardization to make scores not only comparable within patients, but also between patients [7].

If the DAS28-CRP truly underestimates disease activity, this might preclude its use in treat-to-target strategies aimed at reaching sustained remission [8]. On the other hand, one might argue that the DAS28-ESR overestimates disease activity, whereby patients receive unnecessary medication if judgments are based solely on this score. Either way, the use of either one of these measures might lead to different interpretations of a patient’s level of disease activity and, as a result, might lead to different treatment decisions. Nevertheless, one should keep in mind that the DAS28 scores are primarily statistical representations of a patient’s disease activity and not necessarily clinical representations. They may serve as a guide but in clinical practice, rheumatologists may still observe disease activity while the DAS28 points towards a state of remission.

Category-specific agreement was especially poor within the low disease activity group. This is in accordance with the findings of Hensor et al. [9] and might (partly) be due to the lower number of patients in this category compared to the other categories; however, it does demonstrate the main area of concern when both DAS28 measures are assumed to be interchangeable. When using the DAS28-CRP, patients might too easily be categorized as being in remission.

Inconsistent instrument performances were also found over age, gender and disease duration. Consistent with findings from Matsui et al. [6], the differences in the mean values between the DAS28-ESR and DAS28-CRP were larger for females than for males and increased with age. However, Matsui et al. [6] also found larger differences as disease duration increased, whereas we found that differences were largest in the RA group with disease duration < 1 year. This might be due to the composition of our groups. By splitting the group at a disease duration of 1 year, both groups will contain patients with relatively short disease duration. However, this low cut-off point of 1 year was chosen because of the large number of patients with a disease duration of less than 1 year (N=579) vs. patients with a disease duration ≥ 1 year (N=103). Consequently, effects of longer disease duration could not be adequately evaluated within this study.

Since the contribution of the tender joint count, swollen joint count, and general health measure are equal within the DAS28-ESR and DAS28-CRP (i.e. they have the same weighing in the algorithm), score deviations are completely attributable to differences in the ESR and CRP values. Although it is beyond the scope of this study to provide a discussion about which inflammatory marker to prefer as a marker of disease activity, it is recommended to look into this more thoroughly in future studies. Both of these markers are measuring slightly different aspects of the disease process [17]; it is assumed that ESR values tend to reflect the patient’s disease activity over the past few weeks, whereas CRP values are a better reflection of short-term changes in disease activity [4-7, 17]. ESR values are believed to be affected by age and gender, whereas CRP values are not [6, 7, 10-13]. Underlying biological mechanisms might also explain (part of) these differences [13]. For instance, it has been shown that anaemia or abnormally shaped or sized red blood cells might influence ESR levels [18, 19]. Unfortunately, these effects could not be evaluated in this study because this kind of data was not available.

The score deviations cannot simply be solved by adding a constant to the DAS28-CRP (or by subtracting a constant from the DAS28-ESR), since score deviations were found to depend on the degree of disease activity. Therefore, if a rheumatologist wishes to use both scores interchangeably, future studies should focus on finding a robust way to handle the discrepancies in such a way that the transformation is generalizable across distinct patient groups. However, as pointed out by Wells et al. [7] this will not be easy. Specifying distinct disease activity thresholds for the DAS28-ESR and the DAS28-CRP might help in making them comparable. Another solution, as discussed by Hensor et al. [9], might be to incorporate age and gender as variables in the formula.

In conclusion, DAS28-ESR and DAS28-CRP scores are not interchangeable within individuals. The DAS28-CRP tends to yield lower values of disease activity than the DAS28-ESR, resulting in substantial classification differences.

References

1. Prevoo ML, van ’t Hof MA, Kuper HH, van Leeuwen MA, van de Putte LB, van Riel PL. Modified disease activity scores that include twenty-eight–joint counts: development and validation in a prospective longitudinal study of patients with rheumatoid arthritis. . Arthritis Rheum. 1995;38(1):44-8.
2. Vermeer M, Kuper HH, Hoekstra M, Haagsma CJ, Posthumus MD, Brus HL, et al. Implementation of a treat-to-target strategy in very early rheumatoid arthritis: results of the Dutch Rheumatoid Arthritis Monitoring remission induction cohort study. Arthritis Rheum. 2011;63(10):2865-72.
3. Emery P, Van Vollenhoven R, Ostergaard M, Choy E, Combe B, Graninger W, et al. Guidelines for initiation of anti-tumour necrosis factor therapy in rheumatoid arthritis: similarities and differences across Europe. Ann Rheum Dis. 2009;68(4):456-9.
4. Van Riel PL, Fransen J, Scott DL. EULAR handbook of clinical assessments in rheumatoid arthritis. Alphen aan den Rijn: van Zuiden Communications; 2004.

5. Fransen J, Welsing PMJ, de Keijzer RMH, van Riel PLCM. Disease activity scores using C-reactive protein: CRP may replace ESR in the assessment of RA disease activity. Ann Rheum Dis. 2003;62 (Suppl. 1):151.
6. Matsui T, Kuga Y, Kaneko A, Nishino J, Eto Y, Chiba N, et al. Disease Activity Score 28 (DAS28) using C-reactive protein underestimates disease activity and overestimates EULAR response criteria compared with DAS28 using erythrocyte sedimentation rate in a large observational cohort of rheumatoid arthritis patients in Japan. Ann Rheum Dis. 2007;66(9):1221-6.
7. Wells G, Becker JC, Teng J, Dougados M, Schiff M, Smolen J, et al. Validation of the 28-joint Disease Activity Score (DAS28) and European League Against Rheumatism response criteria based on C-reactive protein against disease progression in patients with rheumatoid arthritis, and comparison with the DAS28 based on erythrocyte sedimentation rate. Ann Rheum Dis. 2009;68(8):954-60.
8. Inoue E, Yamanaka H, Hara M, Tomatsu T, Kamatani N. Comparison of Disease Activity Score (DAS)28-erythrocyte sedimentation rate and DAS28- C-reactive protein threshold values. Ann Rheum Dis. 2007;66(3):407-9.
9. Hensor EM, Emery P, Bingham SJ, Conaghan PG. Discrepancies in categorizing rheumatoid arthritis patients by DAS-28(ESR) and DAS-28(CRP): can they be reduced? Rheumatology. 2010;49(8):1521-9.
10. Skogh T, Gustafsson D, Kjellberg M, Husberg M. Twenty eight joint count disease activity score in recent onset rheumatoid arthritis using C reactive protein instead of erythrocyte sedimentation rate. Ann Rheum Dis. 2003;62(7):681-2.
11. Paulus HE, Ramos B, Wong WK, Ahmed A, Bulpitt K, Park G, et al. Equivalence of the acute phase reactants C-reactive protein, plasma viscosity, and Westergren erythrocyte sedimentation rate when used to calculate American College of Rheumatology 20% improvement criteria or the Disease Activity Score in patients with early rheumatoid arthritis. J Rheumatol. 1999;26(11):2324-31.
12. Crowson CS, Rahman MU, Matteson EL. Which measure of inflammation to use? A comparison of erythrocyte sedimentation rate and C-reactive protein measurements from randomized clinical trials of golimumab in rheumatoid arthritis. J Rheumatol. 2009;36(8):1606-10.
13. Radovits BJ, Fransen J, van Riel PLCM, Laan RFJM. Influence of age and gender on the 28-joint Disease Activity Score (DAS28) in rheumatoid arthritis. Ann Rheum Dis. 2008;67(8):1127-31.
14. Leeb BF, Haindl PM, Maktari A, Nothnagl T, Rintelen B. Disease activity score-28 values differ considerably depending on patient's pain perception and sex. J Rheumatol. 2007;34(12):2382-7.
15. DAS28. [cited 2013 August 28]; Available from: http://www.das-score.nl/das28/en/
16. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986;1(8476):307-10.
17. Wolfe F. Comparative usefulness of C-reactive protein and erythrocyte sedimentation rate in patients with rheumatoid arthritis. J Rheumatol. 1997;24(8):1477-85.
18. Kushner I. C-reactive protein in rheumatology. Arthritis Rheum. 1991;34(8):1065-68.
19. Firestein GS, Budd RC, Harris Jr ED, McInnes IB, Ruddy S, Sergent JS. Kelly's textbook of rheumatology. Philadelphia: Saunders Elsevier; 2009.

Chapter 6

How age and sex affect the erythrocyte sedimentation rate and C-reactive protein in early rheumatoid arthritis

Liseth Siemons
P.M. ten Klooster
H.E. Vonkeman
P.L.C.M. van Riel
C.A.W. Glas
M.A.F.J. van de Laar

Revision submitted to BMC Musculoskeletal Disorders

Abstract

Objective: To determine which inflammatory marker, the erythrocyte sedimentation rate (ESR) or C-reactive protein (CRP), is least affected by age, sex, and BMI in patients with early rheumatoid arthritis (RA).

Methods: Baseline data from 589 patients from the DREAM registry were used for analyses. Associations between the inflammatory markers and age, sex, and BMI were evaluated first using univariate linear regression analyses. Next, it was tested whether these associations were independent of a patient's current disease activity (as measured with the other core components of the 28-joint Disease Activity Score) as well as of each other using multiple linear regression analyses with backward elimination. The strengths of the associations were compared using standardized beta (β) coefficients.

Results: Both ESR and CRP were univariately associated with age, sex, and BMI, although the association with BMI disappeared in multivariate analyses. ESR and CRP levels significantly increased with age (β-ESR=0.017, *p*<0.001 and β-CRP=0.009, *p*=0.006), independent of the number of tender and swollen joints, patient-reported general health, and sex. For each decade of aging, ESR and CRP levels became 1.19 and 1.09 times higher, respectively. Furthermore, women demonstrated average ESR levels that were 1.22 times higher than that of men (β=0.198, *p*=0.007), whereas men had 1.20 times higher CRP levels (β=-0.182, *p*=0.048). The effects were shown to be strongest on the ESR.

Conclusions: Age and sex are independently associated with the levels of both acute phase reactants in early RA, emphasizing the need to take these external factors into account when interpreting disease activity measures.

Introduction

Acute phase reactants are commonly used as a measure of inflammation in rheumatoid arthritis (RA) and are part of the provisional definition of RA remission as defined by the ACR/EULAR [1], the ACR preliminary core set of RA disease activity measures [2], and the 28-joint Disease Activity Score (DAS28) [3]. Traditionally, the erythrocyte sedimentation rate (ESR) has been the most widely used marker of inflammation in RA. The ESR is an indirect measure of inflammation, which reflects the level of acute-phase plasma proteins in the blood (e.g. fibrinogen) because these cause the red blood cells to settle more rapidly [4]. However, a number of limitations of this inflammatory marker have become apparent over the years. Although the test is relatively easy and inexpensive to perform, ESR levels respond slowly to inflammatory stimuli and, thus, to changes in disease activity. Also, because the ESR is a non-specific acute phase reactant of systemic inflammation, elevated levels are not necessarily (solely) due to the inflammation of the rheumatic disease. It has been shown that ESR levels can be greatly influenced by, for instance, infections, malignancies, abnormally shaped or sized red blood cells or serum protein concentrations [5]. They also tend to be higher in females than in males [6-11] and appear to increase with age [6-13] and with body mass index (BMI) [4, 9, 14]. Thus, even though the ESR is still widely used in clinical research and practice because of its familiarity, its simplicity, and the attention it received over the years [5], these limitations may complicate the use of ESR in assessing RA disease activity.

In an attempt to overcome some of these limitations, the C-reactive protein (CRP) has been suggested as an alternative inflammatory marker of disease activity in RA [15]. The CRP is a protein that is produced in the liver as a reaction to certain biologic ligands that appear when inflammation develops [4]. Many studies tend to favor CRP over ESR in assessing RA inflammation [16], as it is believed to give a better reflection of current disease activity than ESR because of its more rapid response to increases or decreases in inflammatory stimuli [4, 5]. Another commonly supposed advantage of CRP is its lower susceptibility to external confounding factors like age and sex, compared to the ESR [4]. However, an extensive number of studies have suggested that CRP may exhibit similar dependencies on age [12, 17-19], sex [18-20], and BMI [4, 14, 17-19, 21-23].

As current RA treatment guidelines strongly emphasize early and aggressive treatment aiming at fast remission [24, 25], optimal measurement of inflammation in this patient group is becoming increasingly important. To date, however, it remains unclear which inflammatory marker is affected most strongly by the external effects of age, sex, and BMI in patients with early RA.

Methods

Patients

Data were used from the remission induction cohort of the Dutch Rheumatoid Arthritis Monitoring (DREAM) registry [26]. This observational multicenter cohort included newly, clinically diagnosed patients with RA who experienced symptoms for less than a year. Only patients who had a valid baseline measure of both the ESR and the CRP were included for analyses. All patients were 18 years or older and had not used DMARDs or prednisolone before. Patient recruitment took place from 2006 to 2012 and data collection of included patients is still ongoing. Patients followed a treat-to-target treatment protocol that aimed at a fast remission (DAS28 <2.6). This protocol has been described elsewhere [26, 27]. Informed consent was obtained from each patient. No ethical approval was required, as evaluated by the ethics committees of the participating hospitals and in accordance with Dutch law, since all data collected are part of daily clinical practice.

Measures

Baseline data was collected on several demographic, clinical, and patient-reported measures. The patient’s age, sex, and rheumatoid factor were registered, BMI was calculated as the ratio of weight in kilograms and the square of the height in meters (kg/m^2), physical functioning was assessed with the Health Assessment Questionnaire (HAQ; score range 0-3, where higher values reflect worse outcomes) [28], and physical and mental health status with the Short Form Health Survey with 36 items (SF-36; score range 0-100, where lower values reflect worse outcomes) [29]. At each visit, disease activity was assessed by a trained rheumatologist or nurse practitioner following a standardized procedure using either the DAS28-ESR or the DAS28-CRP [3, 15]. The DAS28 is an index measure including a 28-joint count of both tenderness and swelling (TJC28 and SJC28, score range 0-28), a patient reported visual analog scale of general health (GH, score range 0-100), and an inflammatory marker, which can be either the ESR or the CRP.

Statistical analysis

BMI was divided into 3 categories (i.e. normal weight [<25 kg/m^2], overweight [25-30 kg/m^2], and obese [>30 kg/m^2]) and the values of the ESR and CRP were natural-log-transformed to normalize their distributions. The associations between the inflammatory markers and age, sex, and BMI were evaluated using univariate linear regression analyses first (using p<0.10 for significance). Next, it was tested whether these associations were independent of a patient’s current disease activity (as measured with the other core components of the DAS28) as well as of each other by performing multiple linear regression analyses with backward elimination. Possible interaction terms of the remaining variables were tested for significance. Standardized beta coefficients were used to compare the strength of the relationships in the final models. All analyses were performed using SPSS, version 21.0. Statistical significance in the multivariate analyses was defined as *p*<0.05.

Results

Patients

A total sample of 589 early RA patients was included for analyses. At inclusion, patients had active disease as characterized by a DAS28 score > 2.6 (i.e. the cut-off point for not being in remission), several tender and swollen joints, and a diminished feeling of general and physical health (Table 1). Patients were on average 58 years old, the majority being female, and most had some degree of overweight (60.7% had a BMI ≥ 25 kg/m^2).

Table 1 – Characteristics of the study population at inclusion.

Characteristic	**Mean (SD) or Median (range)***
Sex (female)	362/589 (61.5%)
Age (years)	57.34 (14.20)
Body mass index (kg/m^2)	26.59 (4.44)
Normal (N=198)	22.64 (1.89)
Overweight (N=209)	27.09 (1.35)
Obese (N=97)	33.54 (3.23)
ESR (mm/hour)	21.00 (1-131)
CRP (mg/l)	7.00 (1-238)
28-Tender joint count	3.00 (0-28)
28-Swollen joint count	5.00 (0-28)
General health	44.55 (26.39)
DAS28-ESR	4.23 (1.51)
DAS28-CRP	4.00 (1.35)
SF36 – physical health	37.41 (9.32)
SF36 – mental health	48.13 (11.59)
HAQ	0.99 (0.72)
Rheumatoid factor positive	285/520 (54.8%)

** The values for sex are the number of patients/number of patients assessed. DAS28 = disease activity score for 28 joints, ESR = erythrocyte sedimentation rate, CRP = C-Reactive Protein, SF36 = Short Form Health Survey with 36 items, HAQ = Health Assessment Questionnaire.*

Univariate analyses

Without controlling for possible confounding factors, both ESR and CRP levels increased with age and tended to increase with BMI (Table 2). Furthermore, women tended to have higher ESR levels than men, whereas men tended to have higher CRP levels.

Table 2 – Univariate associations of the inflammatory markers (ESR and CRP) with age, sex, and BMI.

Univariate association	Unstandardized coefficient		Standardized coefficient	*p*-value
	Beta	Standard error	Beta	
ESR				
Age	0.019	0.003	0.296	<0.001
Female sex	0.136	0.078	0.072	0.083
Overweight	0.048	0.088	0.027	0.587
Obese	0.190	0.110	0.084	0.086
CRP				
Age	0.013	0.003	0.164	<0.001
Female sex	-0.183	0.096	-0.078	0.057
Overweight	0.248	0.109	0.111	0.023
Obese	0.250	0.136	0.089	0.067

** Analyses were performed on the naturally log-transformed ESR and CRP values to normalize their distributions. ESR = erythrocyte sedimentation rate, CRP = C-reactive protein. The males and normal weight patients were used as reference categories.*

Multivariate analyses

After adjusting for the other core components of disease activity that are included in the DAS28 (i.e. the TJC28, SJC28, and GH), sex and age were still independently associated with both inflammatory markers, while BMI was no longer related to a patient's ESR or CRP level (Table 3). Although the effects of age were slightly attenuated, the effects of sex were not and became even stronger in the ESR model.

Thus, obese patients did not show significantly higher levels of ESR or CRP (mean ± SD ESR: 29.34 ± 21.87, CRP: 17.30 ± 22.06) than overweight (ESR: 26.28 ± 18.76, CRP: 18.14 ± 19.84) or normal weight patients (ESR: 26.70 ± 22.25, CRP: 16.97 ± 24.21). On the other hand, ESR and CRP levels did significantly increase with age (β-ESR=0.017, *p*<0.001 and β-CRP=0.009, *p*=0.006), independent of the number of tender and swollen joints, the patient-reported degree of general health, and the patient's sex. The unstandardized betas (0.017 vs. 0.009) represent the change of the natural-log-transformed ESR and CRP levels, respectively, for each year of aging. Transformed back to their normal values, this means that for each decade of aging the ESR and CRP levels become 1.19 and 1.09 times higher, respectively. Furthermore, women demonstrated average ESR levels that were 1.22 times higher than those of the men (β=0.198, *p*=0.007), whereas men had 1.20 times higher CRP levels (β=-0.182, *p*=0.048). The effects of age and sex were strongest in the ESR model, indicating that the ESR is more sensitive to the effects of these external factors. Although there were no significant interactions between age and sex (interaction ESR model: *p*=0.222 and interaction CRP model: *p*=0.665), differences in ESR and CRP between male and female patients appeared to gradually decrease with age (Table 4).

Table 3 – Multivariate associations of the inflammatory markers (ESR and CRP) with age and sex.

Variables*	**Unstandardized coefficient**		**Standardized coefficient**	***p*-value**
	Beta	**Standard error**	**Beta**	
ESR model				
TJC28	-0.012	0.008	-0.074	0.115
SJC28	0.044	0.008	0.258	<0.001
General health	0.006	0.001	0.171	<0.001
Female sex	0.198	0.073	0.104	0.007
Age	0.017	0.003	0.264	<0.001
CRP model				
TJC28	-0.001	0.010	-0.006	0.908
SJC28	0.058	0.010	0.280	<0.001
General health	0.008	0.002	0.180	<0.001
Female sex	-0.182	0.092	-0.078	0.048
Age	0.009	0.003	0.108	0.006

** ESR = erythrocyte sedimentation rate, CRP = C-reactive protein, TJC28 = tender joint count in 28-joints, SJC28 = swollen joint count in 28-joints*

Table 4 – ESR and CRP levels in men and women per age group.

Men (N=227)		**Women (N=362)**	
Variable*	**Mean (sd)**	**Variable***	**Mean (sd)**
ESR		**ESR**	
Age group		Age group	
<55 years (n=70)	18.96 (19.43)	<55 years (n=161)	22.14 (20.90)
55 - 65 years (n=80)	24.34 (22.72)	55 - 65 years (n=95)	28.84 (22.93)
>65 years (n=77)	31.87 (18.73)	>65 years (n=106)	31.65 (19.98)
CRP		**CRP**	
Age group		Age group	
<55 years (n=70)	16.13 (20.43)	<55 years (n=161)	12.62 (18.57)
55 - 65 years (n=80)	19.64 (25.81)	55 - 65 years (n=95)	17.96 (30.63)
>65 years (n=77)	19.40 (22.55)	>65 years (n=106)	18.93 (22.82)

** ESR = erythrocyte sedimentation rate, CRP = C-reactive protein*

Discussion

Age and sex were independently associated with the levels of both acute phase reactants in early RA, although the effects appeared to be strongest on the ESR. These results emphasize the need to take these external factors into account when interpreting disease activity in patients with early RA. Because the acute phase reactants tend to increase with age, independent of other core measures of disease activity, the disease activity of older-aged patients might be overestimated. Also, ESR values are more likely to be elevated in women than in men, whereas the opposite appears to be the case for CRP.

Although no significant interactions were found between age and sex, the gradually decreasing differences in ESR and CRP between male and female patients with age are consistent with findings from Radovits et al. [11]. Furthermore, the finding that the ESR tends to be more elevated in women than in men is consistent with the results of previous studies [7-11]. However, inconclusive results have been reported on sex differences in CRP levels. Where Lee et al. [19] found CRP values to be higher in males than in females, other studies have reported the opposite [18, 20]. Ethnic differences across population samples (including genetic variations, diet, and lifestyle) [9, 19] and the use of different measurement methods for determining the CRP levels may possibly explain these diverse results. Since the patient population of this present study is probably more similar to the France and US populations of Piéroni et al. [18] and Lakoski et al. [20], respectively, than the Korean population examined by Lee et al. [19] it was expected to find higher CRP levels in females as well. Nevertheless, the exact underlying mechanisms behind these dependencies remain unknown and all these previous studies were conducted in healthy community populations instead of in samples with active disease. Perhaps the underlying processes of early RA, certain lifestyle differences, the higher prevalence of obesity in the women, hormonal factors, or differences in metabolic risk factor [10, 11, 19] might (partly) explain the higher baseline CRP levels in men in this study. However, this warrants further research.

Although one recent study in general RA patients concluded that age was unrelated to CRP [11], Ranganath et al. [12] did find significant associations of age with both inflammatory markers in early RA. Likewise, the increasing inflammatory marker levels with age are consistent with a wide variety of other studies [7-10, 12, 13, 17-19]. Like sex differences, the increasing ESR and CRP levels with age might be explained by certain immunological or hematological changes and, for instance, hormonal changes in women who reach the menopause [10, 11, 13]. Furthermore, it has been suggested that RA in older aged people might be more severe than in the young [11]. Supplementary post-hoc analyses did indeed show significantly more swollen joints, higher inflammatory values, and a more active disease according to the DAS28-ESR score in the two highest age groups (55-65 and >65 years old) compared to the youngest group (<55 years old).

In contrast to previous studies, no independent association between BMI and acute phase reactants was found. It has been suggested that higher BMI levels (i.e. overweight or obese) promote a higher secretion of interleukin-6 (IL-6), a pro-inflammatory cytokine, regulating the secretion of the acute phase reactants [17-19, 22]. The proportion of obese patients in this study may have been too small to detect these effects (15.2% of the men and 21.7% of the women) or perhaps BMI does not yet have a significant influence this early on in the disease. To evaluate whether the associations changed over time, the multivariate analyses were repeated on all available sample data at 1 year (data not reported) and, interestingly, BMI did indeed become significantly associated with both the ESR and the CRP at this point in time, indicating its possible importance at later stages of the disease. Age continued to be significantly associated at 1 year, although this relationship was no longer linear but increased exponentially. Sex, on the other hand, was still significantly associated with ESR levels (showing higher levels in females), but was no longer significantly associated with CRP.

A possible limitation of the current study might be the exclusion of other potential confounding variables, such as certain lifestyle factors (e.g. smoking or physical activity), dietary patterns, or medication use (e.g. estrogen, steroids, or NSAIDs) [17-19, 21, 30]. Follow up analyses including a patient's rheumatoid factor and physical functioning (as measured with the HAQ) were carried out and they did indeed point to other possible dependencies as well. In case of the ESR model, both sex and age remained significantly related but, in addition, physical functioning was significantly related as well (p=0.001) and rheumatoid factor approached significance (p=0.053). This indicates that rheumatoid factor positivity as well as a deteriorating physical functioning are both related to higher levels of the ESR. In the CRP model, on the other hand, sex lost its significance and was replaced by physical functioning (p<0.001), while rheumatoid factor did not show a significant association. Thus, both models point to an independent association between the RA patients' physical functioning and inflammatory markers. How this relationship works remains unclear. Perhaps worsening physical health coincides with less physical activity which has been shown to be independently associated with higher CRP levels [19, 21]. This might be related to potential positive effects of physical activity on the body mass index. Vigorous physical activity might prevent the accumulation of abdominal fat which, consequently, may result in a reduced IL-6 production and lower levels of the acute phase reactants [19, 31]. Perhaps this also explains why the relation with sex dissipates in the CRP model. If men were more active than women, this might have caused their CRP levels to drop significantly more, reducing any sex effects [31].

Conclusions

In conclusion, these results suggest that caution should be taken when assessing disease activity in early RA because the ESR and CRP levels are both influenced by non-inflammatory factors like age and sex. Consequently, disease activity might be either overestimated or underestimated. Since the CRP appeared to be less sensitive to external factors this might be designated as the preferred measure, which is consistent with results from Radovits et al. [11] and Crowson et al. [16]. However, one should keep in mind that there might be other confounding factors that have an effect on the inflammatory markers but which were not included in this study. It could also be argued to develop modified DAS28 scores, including an adaptation for age and sex because these are two risk factors which cannot be modified. Miller et al. [7] already proposed a simple formula for calculating age-adjusted ESR values. However, this was several decades ago, so supplementary research on the possible incorporation of sex and age adjustments in the DAS28 formula is recommended. Another solution might be to specify age and sex specific thresholds of current disease activity scores in order to make them comparable across patient subgroups of different age and sex. Finally, some studies have suggested to exclude the inflammatory markers altogether from disease activity measurement, as is done with the Clinical Disease Activity Index (CDAI) [16, 32]. However, since the inflammatory markers are two of the most reliable components of disease activity measures as the DAS28 [33], further research is recommended.

References

1. Felson DT, Smolen JS, Wells G, Zhang B, van Tuyl LHD, Funovits J, et al. American College of Rheumatology/European League Against Rheumatism provisional definition of remission in rheumatoid arthritis for clinical trials. Arthritis Rheum. 2011;63(3):573-86.
2. Felson DT, Anderson JJ, Boers M, Bombardier C, Chernoff M, Fried B, et al. The American College of Rheumatology preliminary core set of disease activity measures for rheumatoid arthritis clinical trials. The Committee on Outcome Measures in Rheumatoid Arthritis Clinical Trials. Arthritis Rheum. 1993;36(6):729-40.
3. Prevoo ML, van ’t Hof MA, Kuper HH, van Leeuwen MA, van de Putte LB, van Riel PL. Modified disease activity scores that include twenty-eight–joint counts: development and validation in a prospective longitudinal study of patients with rheumatoid arthritis. Arthritis Rheum. 1995;38(1):44-8.
4. Firestein GS, Budd RC, Harris Jr ED, McInnes IB, Ruddy S, Sergent JS. Kelly's textbook of rheumatology. Philadelphia: Saunders Elsevier; 2009.
5. Kushner I. C-reactive protein in rheumatology. Arthritis Rheum. 1991;34(8):1065-68.
6. Nestel AR. ESR changes with age - a forgotten pearl. BMJ. 2012;344:e1403.
7. Miller A, Green M, Robinson D. Simple rule for calculating normal erythrocyte sedimentation rate. Br Med J (Clin Res Ed). 1983;286(6361):266.

8. Shearn MA, Kang IY. Effect of age and sex on the erythrocyte sedimentation rate. J Rheumatol. 1986;13(2):297-8.
9. De Silva DA, Woon FP, Chen C, Chang HM, Wong MC. Serum erythrocyte sedimentation rate is higher among ethnic South Asian compared to ethnic Chinese ischemic stroke patients. Is this attributable to metabolic syndrome or central obesity? J Neurol Sci. 2009;276(1-2):126-9.
10. Böttiger LE, Svedberg CA. Normal erythrocyte sedimentation rate and age. Br Med J. 1967;2(5544):85-7.
11. Radovits BJ, Fransen J, van Riel PL, Laan RF. Influence of age and gender on the 28-joint Disease Activity Score (DAS28) in rheumatoid arthritis. Ann Rheum Dis. 2008;67(8):1127-31.
12. Ranganath VK, Elashoff DA, Khanna D, Park G, Peter JB, Paulus HE. Age adjustment corrects for apparent differences in erythrocyte sedimentation rate and C-reactive protein values at the onset of seropositive rheumatoid arthritis in younger and older patients. J Rheumatol. 2005;32(6):1040-2.
13. Hayes GS, Stinson IN. Erythrocyte sedimentation rate and age. Arch Ophthalmol. 1976;94(6):939-40.
14. Oeser A, Chung CP, Asanuma Y, Avalos I, Stein CM. Obesity is an independent contributor to functional capacity and inflammation in systemic lupus erythematosus. Arthritis Rheum. 2005;52(11):3651-9.
15. Fransen J, Welsing PMJ, de Keijzer RMH, van Riel PLCM. Disease activity scores using C-reactive protein: CRP may replace ESR in the assessment of RA disease activity. Ann Rheum Dis. 2003;62(Suppl. 1):151.
16. Crowson CS, Rahman MU, Matteson EL. Which measure of inflammation to use? A comparison of erythrocyte sedimentation rate and C-reactive protein measurements from randomized clinical trials of golimumab in rheumatoid arthritis. J Rheumatol. 2009;36(8):1606-10.
17. Kawamoto R, Kusunoki T, Abe M, Kohara K, Miki T. An association between body mass index and high-sensitivity C-reactive protein concentrations is influenced by age in community-dwelling persons. Ann Clin Biochem. 2013;50(Pt5):457-64.
18. Piéroni L, Bastard JP, Piton A, Khalil L, Hainque B, Jardel C. Interpretation of circulating C-reactive protein levels in adults: body mass index and gender are a must. Diabetes Metab. 2003;29(2 Pt 1):133-8.
19. Lee YJ, Lee JH, Shin YH, Kim JK, Lee HR, Lee DC. Gender difference and determinants of C-reactive protein level in Korean adults. Clin Chem Lab Med. 2009;47(7):863-9.
20. Lakoski SG, Cushman M, Criqui M, Rundek T, Blumenthal RS, D'Agostino Jr RB, et al. Gender and C-reactive protein: data from the Multiethnic Study of Atherosclerosis (MESA) cohort. Am Heart J. 2006;152(3):593-8.
21. Rommel J, Simpson R, Mounsey JP, Chung E, Schwartz J, Pursell I, et al. Effect of body mass index, physical activity, depression, and educational attainment on high-sensitivity C-reactive protein in patients with atrial fibrillation. Am J Cardiol. 2013;111(2):208-12.
22. Kao TW, Lu IS, Liao KC, Lai HY, Loh CH, Kuo HK. Associations between body mass index and serum levels of C-reactive protein. S Afr Med J. 2009;99(5):326-30.
23. Rawson ES, Freedson PS, Osganian SK, Matthews CE, Reed G, Ockene IS. Body mass index, but not physical activity, is associated with C-reactive protein. Med Sci Sports Exerc. 2003;35(7):1160-6.
24. Combe B, Landewe R, Lukas C, Bolosiu HD, Breedveld F, Dougados M, et al. EULAR recommendations for the management of early arthritis: report of a task force of the European Standing Committee for International Clinical Studies Including Therapeutics (ESCISIT). Ann Rheum Dis. 2007;66(1):34-45.

25. Smolen JS, Aletaha D, Bijlsma JW, Breedveld FC, Boumpas D, Burmester G, et al. Treating rheumatoid arthritis to target: recommendations of an international task force. Ann Rheum Dis. 2010;69(4):631-7.
26. Vermeer M, Kuper HH, Hoekstra M, Haagsma CJ, Posthumus MD, Brus HL, et al. Implementation of a treat-to-target strategy in very early rheumatoid arthritis: results of the Dutch Rheumatoid Arthritis Monitoring remission induction cohort study. Arthritis Rheum. 2011;63(10):2865-72.
27. Siemons L, ten Klooster PM, Vonkeman HE, Glas CAW, van de Laar MAFJ. Distinct trajectories of disease activity over the first year in early rheumatoid arthritis patients following a treat-to-target strategy. Arthritis Care Res. In press.
28. Fries JF, Spitz P, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. Arthritis Rheum. 1980;23:137-45.
29. Ware JE, Kosinski M, Dewey JE. How to Score Version Two of the SF-36 Health Survey. Lincoln, RI: QualityMetric, Incorporated; 2000.
30. Husain TM, D.H. K. C-reactive protein and erythrocyte sedimentation rate in orthopaedics. The University of Pennsylvania Orthopaedic Journal. 2002;15:13-6.
31. Albert MA, Glynn RJ, Ridker PM. Effect of physical activity on serum C-reactive protein. Am J Cardiol. 2004;93(2):221-5.
32. Aletaha D, Nell VP, Stamm T, Uffmann M, Pflugbeil S, Machold K, et al. Acute phase reactants add little to composite disease activity indices for rheumatoid arthritis: validation of a clinical activity score. Arthritis Res Ther. 2005;7(4):R796-806.
33. Siemons L, ten Klooster PM, Vonkeman HE, van de Laar MAFJ, Glas CAW. The reliability of 28-joint Disease Activity Scores unraveled and optimized. BMC Med Res Methodol. Submitted.

Chapter 7

Distinct trajectories of disease activity over the first year in early rheumatoid arthritis patients following a treat-to-target strategy

L. Siemons
P.M. ten Klooster
H.E. Vonkeman
C.A.W. Glas
M.A.F.J. van de Laar

Arthrit Care Res. (in press)

Abstract

Objective: Although treat-to-target (T2T) strategies are effective in early rheumatoid arthritis (RA) patients, important individual variations exist in the course toward remission. Growth mixture modeling provides more insight into this heterogeneity by identifying subgroups of patients with similar response patterns. This study aimed to identify distinct trajectories of disease activity in early RA patients following a T2T strategy during their first year.

Methods: Data on various clinical and patient-reported measures were collected from the Dutch Rheumatoid Arthritis Monitoring remission induction cohort. Growth mixture modeling was applied to examine the impact of T2T on subgroups characterized by different types of growth trajectories, as measured with the Disease Activity Score in 28 joints.

Results: Three distinct trajectories of disease activity were found. The normative trajectory contained most patients (82.6%), showing a quickly decreasing disease activity that stabilized at remission after 9 months. This group performed best on clinical and patient-reported measures over time and were more likely to be men. A smaller group (14.1%) also approached remission, but demonstrated a slower response to treatment. Finally, a minority (3.3%) showed no improvement after 1 year, despite an initial quick decrease in disease activity during the first months of treatment.

Conclusion: Disease activity in early RA patients during the first year of a T2T strategy does not follow a linear pattern, nor is a single developmental trajectory applicable to all patients. Future studies should attempt to identify more specific risk factors for poor outcome to enable early identification of patients in need of alternative therapeutic approaches.

Introduction

Since rheumatoid arthritis (RA) cannot yet be cured, the purpose of treatment is to reduce disease activity as quickly and as much as possible. Disease activity is a multifactorial concept and various components should be evaluated in order to obtain insight into the progress of the disease. The Disease Activity Score in 28 joints (DAS28) is one of the index measures developed to get an overall impression of a patient's disease activity [1]. It combines a 28 tender joint count (TJC28), a 28 swollen joint count (SJC28), a laboratory measure of inflammation, and a patient-reported rating of general health into a single measure of disease activity.

The DAS28 is an accepted and widely used instrument, is embedded in treatment protocols, and is often used as the criterion measure when deciding on the most appropriate treatment for a specific patient and for evaluating treatment effectiveness over time [2, 3]. Treat-to-target (T2T) strategies have been shown to be effective in quickly reaching remission in early RA patients in daily clinical practice [3]. However, one size may not fit all. Even though T2T strategies are effective at a group level, this does not mean that a single T2T strategy will be the best treatment option for all patients. Consequently, treating a heterogeneous population of RA patients according to a relatively strictly defined treatment protocol might not be equally advantageous for all patients.

Whereas previous studies already described the average development of disease activity within early RA patients using a single trajectory [3-5], this study focused on the deviations from this trajectory. By identifying distinct developmental trajectories of the DAS28, patient groups with similar response patterns can be identified [6, 7] that might need different treatment strategies in order to control their disease. The value of trajectory analyses has already been recognized in previous studies. Distinct health-related quality of life trajectories have been found within patients undergoing coronary artery bypass graft surgery [8], distinct trajectories of pain were found within adolescents in a general population [9] and in patients with hip osteoarthritis [10], and distinct trajectories of psychological distress were found in RA patients [11]. However, no study has yet examined the trajectories of RA disease activity.

The aim of this study was to identify distinct trajectories of disease activity in the first year of treatment within an early RA patient group following a T2T strategy and to examine the patient characteristics of each identified developmental trajectory. Identifying distinct trajectories can improve our understanding of normative and non-normative response trajectories over time, and a description of the group characteristics might provide insight into the reasons why certain patients respond differently than others. Identifying these distinct trajectories in an early stage of the disease may help in refining future treatment protocols.

Patients and methods

Patients

The present study used data from the ongoing Dutch Rheumatoid Arthritis Monitoring remission induction cohort, an observational, multicenter cohort that was established in 2006 to evaluate the effect of a protocoled T2T strategy aimed at reaching a (sustained) state of remission in early RA patients [2]. Subjects were included in this cohort as soon as they were clinically diagnosed with early RA, were ages ≥18 years, had no history of taking disease-modifying antirheumatic drugs (DMARDs) or prednisolone, and had a maximum symptom duration of 1 year. The study protocol was evaluated by the ethics committees of all participating hospitals and they determined that no ethical approval was required because the study data was gathered during daily clinical practice, which is in accordance with the Dutch Law. Nonetheless, informed consent was obtained from each patient.

Treatment

All patients were following a T2T treatment strategy aimed at reaching remission (DAS28 <2.6). Treatment was intensified according to protocol (Table 1) when treatment effects were insufficient (i.e. if DAS28 was not <2.6 when treated with DMARDs and if the DAS28 was not <3.2 when treated with anti-tumor necrosis factor α [anti-TNFα] agents).

Table 1 – Treat-to-target treatment protocol*.

Week	Medication and doses
Week 0	MTX 15 mg/week
Week 8	MTX 25 mg/week
Week 12	MTX 25 mg/week + SSZ 2 x 1000 mg/day
Week 20	MTX 25 mg/week + SSZ 3 x 1000 mg/day
Week 24	MTX 25 mg/week + adalimumab 40 mg/2 weeks
Week 36	MTX 25 mg/week + adalimumab 40 mg/week
Week 48-52	MTX 25 mg/week + etanercept 50 mg 1x/week
1 year, 3 months	MTX 25 mg/week + infliximab 3mg/kg/8 weeks (after a loading dose at weeks 0, 2 and 6)
1 year, 6 months	MTX 25 mg/week + infliximab 3mg/kg/4 weeks

** MTX = methotrexate, SSZ = sulfasalazine*

As soon as remission was reached, medication was held constant. If remission continued for at least 6 months, medication doses could be reduced, starting with the medication that was added last, and medication might even be discontinued completely. Specific details on medication use were not available for this study but have been described before, together with a more elaborative description of the treatment protocol [2].

Measures

Data were collected on various clinical and patient-reported measures, including measures of disease activity and physical functioning, laboratory measures, and other disease-related variables. Disease activity was assessed by a trained nurse practitioner or rheumatologist using the DAS28, which consists of a TJC28, an SJC28, a 100-mm visual analog scale (VAS) on general health (where 0 = very good and 100 = very bad), and the erythrocyte sedimentation rate (ESR) [1].

Furthermore, the patients rated their pain on a 100-mm VAS (where 0 = no pain and 100 = unbearable pain) and they completed the Short Form 36 health survey in order to get an impression of their current physical and mental health states [12]. Finally, some additional laboratory measures were collected, including C-reactive protein (CRP) level. Blood samples were tested with a time-resolved fluoroimmunoassay of the IgM rheumatoid factor (RF) [13] to determine a patient’s RF.

Statistical analysis

Data at inclusion and after 3, 6, 9, and 12 months of T2T were included in the analyses. Growth mixture modeling (GMM) was carried out to determine whether different developmental trajectories could be distinguished. As opposed to regression, factor analysis, and structural equation modeling, which are variable-centered approaches, GMM is a person-centered approach, focusing on the relationships among individual patients [7]. GMM uncovers unobserved heterogeneity in the development of the DAS28 over time by identifying distinct subgroups within the early RA patient sample, where each subgroup follows a different growth trajectory and where patients within a group are more similar to each other than patients between groups [6, 7, 14]. To determine the optimal number of classes for describing the patient population, a combination of fit indices was examined [6]. A significant ($p<0.05$) adjusted Lo-Mendell-Rubin likelihood ratio test (LRT) suggests that the model with K classes is an improvement over the model with K-1 classes. In addition, information criteria were compared among different models, including the Bayesian information criterion (BIC), the sample size-adjusted BIC, and Akaike’s information criterion (AIC), where smaller values indicate a better fit of the model to the data. Finally, the entropy value was evaluated, which shows the accuracy of the classification. Values >0.80 demonstrate good classification [6]. Both linear and quadratic models were evaluated.

After determining the best-fitting model, the groups were compared on sex compilation and RF using cross-tab analysis. Furthermore, groups comparisons were made on clinical or patient-reported outcome measures at each time point using one-way analyses of variance (ANOVAs). A Bonferroni sample size-adjusted *p*-value was used to test for significance. GMM analyses were performed in Mplus, version 7.11, and cross-tab and one-way ANOVAs were performed in SPSS, version 20.0.

Results

Patient characteristics of the total sample

An overview of the patient characteristics at baseline is shown in Table 2. Data were available from 568 patients (62.9% women), with a mean age of 57 years and a moderately active disease (mean DAS28 4.22). On average, patients had 5 tender joints, 6 swollen joints, an ESR of 26.27 mm/hour, a CRP level of 16.41 mg/l, and a rather low general health score of 43.90. Furthermore, most patients were RF positive (59.9%), experienced pain (mean score 43.51), and had a diminished physical health status (mean score 37.12).

Table 2 – Baseline characteristics of the patients (N=568)*.

Variables	Mean ± SD
Female sex, no. (%)	357 (62.9)
Age, years	57.31 ± 14.29
DAS28	4.22 ± 1.50
TJC28	4.72 ± 5.45
SJC28	5.89 ± 5.39
General health score	43.90 ± 26.30
ESR, mm/hour	26.27 ± 20.64
Pain score	43.51 ± 26.65
SF-36 physical	37.12 ± 9.14
SF-36 mental	48.25 ± 11.59
CRP, mg/l	16.41 ± 20.94
BMI, kg/m^2	26.77 ± 4.52
RF positive, no./total (%)	264/441 (59.9)

** Values are the mean ± SD unless indicated otherwise. DAS28 = Disease Activity Score in 28 joints, TJC28 = 28 tender joint count, SJC28 = 28 swollen joint count, ESR = erythrocyte sedimentation rate, SF-36 = Short Form 36 health survey, CRP = C-reactive protein, BMI = body mass index, RF = rheumatoid factor.*

Followup measurements of the DAS28 included a decreasing number of patients, since not all patients visited their rheumatologist every 3 months after diagnosis and because some patients had a followup time of <1 year: 3-month scores were available from 520 patients, 6-month scores were available from 465 patients, 9-month scores were available from 419 patients, and 12-month scores were available from 406 patients. It was not expected that missing data were related to the disease course. To support this claim, for every missing DAS28 score, it was evaluated whether the measurements both before and after the missing measurement were significant outliers in their respective distributions within the 5 time points and the identified trajectory groups. The number of outliers was not significant (i.e. 2% of the outliers were significant, which was below the nominal significance level of 5%).

GMM

Based on clinical judgment as well as the adjusted LRT, BIC, sample size-adjusted BIC, AIC, and entropy values, a quadratic 3-class model was selected as the most appropriate model for describing the development of the DAS28 scores of the early RA patients (Table 3). Although the entropy value was slightly below the cutoff value of 0.80, it was highest in this model.

Table 3 - Fit statistics for various growth mixture modeling analyses*.

	2 classes, linear	**3 classes, linear**	**2 classes, quadratic**	**3 classes, quadratic**
BIC	7,145.305	7,163.575	6,964.097	6,977.581
Sample size-adjusted BIC	7,097.687	7,100.084	6,897.431	6,888.694
AIC	7,080.174	7,076.733	6,872.912	6,856.001
Adjusted LRT (p-value)	0.0243	0.4551	0.0029	0.0294
Entropy	0.427	0.652	0.631	0.768
Group sizes, no.	96 / 472	3 / 65 / 500	108 / 460	19 / 80 / 469

** A 2-class solution means that a growth model with 2 classes was specified beforehand. Likewise, a 3-class solution retrieves 3 distinct classes from the data. BIC = Bayesian information criterion, AIC = Akaike's information criterion, LRT = Lo-Mendell-Rubin likelihood ratio test*

The first subgroup consisted of 469 patients (82.6%) showing a quick decrease in the DAS28 score in the early treatment stages, which slowly stabilized at the remission level after 9 months (Figure 1). This "fast response" trajectory represents the normative course. The second subgroup consisted of 80 patients (14.1%) who did not react to the treatment as quickly as the first group but who did show a steady decrease in disease activity, approaching remission after 1 year of treatment. This subgroup was defined as the "slow response" group. The third and final group was much smaller (n=19 [3.3%]) and consisted of patients who showed no improvement after 1 year of treatment (the "poor outcome" group). Although their disease activity quickly decreased over the first 6 months, it flared after that, and after 12 months they were back at their initial level of disease activity.

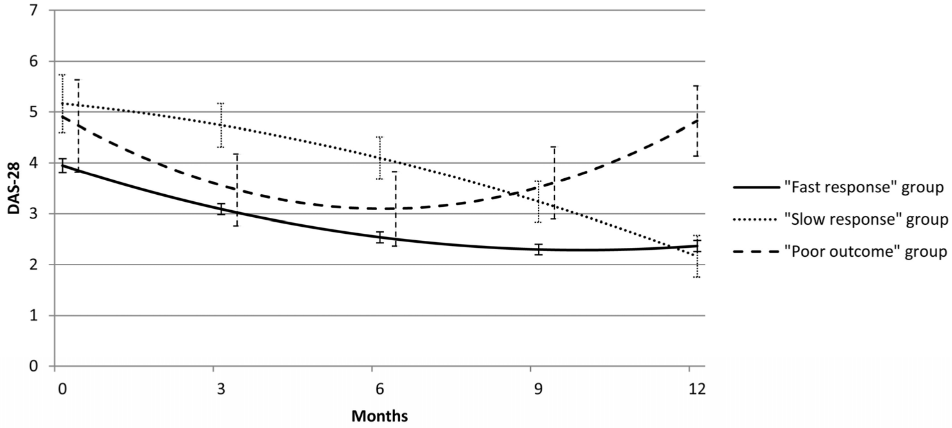


Figure 1 – Three distinct trajectories of the Disease Activity Score in 28 joints (DAS28) were identified within the early rheumatoid arthritis cohort: the “fast response” group (n=469 [82.6%]), the “slow response” group (n=80 [14.1%]), and the “poor outcome” group (n=19 [3.3%]). For clarity, the error bars of the “poor outcome” group are shifted a bit to the right.

Patient characteristics of the 3 trajectories

Group comparisons at baseline (Table 4) showed that the 3 identified groups did not significantly differ in age, sex, or RF positivity, but the fast response group did have better scores for disease activity and health status. Significantly better DAS28, TJC28, SJC28, general health, ESR, pain, and physical health status measures were found compared to the slow response group, and better DAS28 and TJC28 measures were found compared to the poor outcome group. Aside from some small deviations, comparable results were found on subsequent measurement points (results not shown). In addition, analyses showed that the poor outcome group performed significantly better than the slow response group after 3 and 6 months, but had the worst performance of all groups after 12 months.

Table 4 – Results of the one-way ANOVAs testing differences between groups at baseline*.

Variable	Fast response group (n=469)	Slow response group (n=80)	Poor outcome group (n=19)	F value	*p*
DAS28[†]	4.01 ± 1.44	5.24 ± 1.28	5.11 ± 1.72	27.949	0.000[‡]
TJC28[†]	3.96 ± 4.60	8.58 ± 7.40	7.21 ± 7.56	29.219	0.000[‡]
SJC28[§]	5.48 ± 5.18	7.75 ± 5.62	8.05 ± 7.35	7.820	0.000[‡]
General health[§]	41.79 ± 25.91	54.82 ± 25.83	50.11 ± 26.96	9.060	0.000[‡]
ESR[§]	24.62 ± 19.52	34.79 ± 24.39	31.06 ± 21.99	8.838	0.000[‡]
Pain[§]	41.54 ± 26.45	53.27 ± 23.06	51.00 ± 35.71	7.358	0.001[‡]
SF-36 physical[§]	37.97 ± 9.04	32.93 ± 7.90	33.92 ± 11.54	9.926	0.000[‡]
SF-36 mental	48.60 ± 11.33	46.26 ± 13.14	48.18 ± 10.59	1.163	0.313
CRP	15.51 ± 19.91	20.53 ± 22.25	21.56 ± 34.83	2.422	0.090
BMI	26.63 ± 4.39	27.03 ± 5.23	28.74 ± 3.91	1.942	0.145
Age	57.32 ± 14.61	57.98 ± 12.49	54.21 ± 13.48	0.533	0.587
RF positive, no./total (%)[¶]	213/357 (59.7)	42/69 (60.9)	9/15 (60)	0.035	0.983
Male sex, no. (%)[¶]	185 (39.4)	21 (26.2)	5 (26.3)	6.085	0.048

** Values are the mean ± SD unless indicated otherwise. ANOVA = analysis of variance, DAS28 = Disease Activity Score in 28 joints, TJC28 = 28 tender joint count, SJC28 = 28 swollen joint count, ESR = erythrocyte sedimentation rate, SF-36 = Short Form 36 health survey, CRP = C-reactive protein, BMI = body mass index, RF = rheumatoid factor.*

† The normative group (i.e. the fast response group) scores were significantly lower than the other 2 groups.

‡ This Bonferroni sample size–adjusted p-value is significant at 0.05/13 = 0.0038.

§ The normative group (i.e. the fast response group) scores were significantly lower than the slow response group.

¶ Group differences were evaluated using cross-tab analysis instead of one-way ANOVA. The results show Pearson's chi-square statistic with its corresponding p-value.

Discussion

Three developmental trajectories of disease activity could be distinguished within the sample of early RA patients following a T2T strategy. The normative trajectory was in accordance with the clinical experience that most patients show a steep decrease in disease activity during the first 6 months, reaching remission quickly. The second group did not improve as fast as this first group, but did show gradually decreasing levels of disease activity, approaching a state of remission after 12 months. The final group contained only a small proportion of the patients, but they had the worst disease outcome; although their disease activity quickly decreased in the beginning, this course reversed after 6 months and returned to the baseline level of disease activity.

To ensure that the results were not biased by any skewed distributions of the variables, Kruskal-Wallis analyses were also performed for between-group comparisons. Since no large deviations were found from the original results, the primary outcomes were supported.

This study shows that improvement (or deterioration) in RA disease activity is not a linear process. Also, one overall clinical image of the development of disease activity over time does not reflect the true situation, since not all patients follow the same developmental trajectory of disease activity, a finding that supports experiences from clinical practice. It is important that rheumatologists are aware of this heterogeneity within their patient populations and of the fact that a T2T treatment strategy does not have the same impact on all patients.

Although no longer significant after Bonferroni correction, the normative trajectory appeared to contain more men than the other 2 trajectories. This is consistent with previous findings that men generally have a less severe disease and show higher remission rates than women [15-18]. Interestingly, RF positivity did not differ significantly between the 3 groups. Therefore, even though RF is being used as a diagnostic factor and RF positivity is often associated with poorer disease outcomes [19], it does not appear to be predictive of poorer outcome in this population of early RA patients.

Aside from the overall finding that the poor outcome group performs worst on most outcome variables and the fast response group performs best, it is important to note that this study only examined the trajectories of disease activity during the first year of treatment. It is recommended to investigate a longer time period in order to determine whether the poor outcome group continues to have worse outcomes over time or whether they eventually also reach a state of remission.

Two other issues need to be addressed as well. First, it is not clear why some patients did not improve over time. The analyses showed no clear cause for the differences in treatment responses. Supplementary case history examinations that were carried out in the poor outcome group pointed to 2 main factors that might be related to their loss of treatment response: patients were tapering their initial dose of prednisolone and patients stopped taking their DMARDs, mostly because of side effects. These are interesting findings, emphasizing the need to further investigate the role of prednisolone in the development of a patient’s disease activity.

Second, more research is needed to validate the existence of the 3 trajectories in other early RA cohorts and to determine which factors predict the trajectory that a patient is most likely going to follow (i.e. to predict group membership). If possible, such studies should not only focus on a wide variety of clinical factors, but also on genetic markers. Knowledge on the predicted disease course group can be of significant importance in clinical practice, since it offers a rheumatologist the opportunity to tailor a single T2T strategy more specifically to the needs of the distinct patient subgroups. For instance, it can be argued that the fast response group probably does not need any extra attention, since they respond well to the current treatment, reaching remission quickly. The slow response group, however, might need a different treatment strategy, since their treatment effect is much less pronounced. On average, they need several months longer than the fast response group to reach a state of remission. Perhaps these patients should be given additional steroids or they should initially start their treatment with TNF inhibitor drugs instead of DMARDs. Finally, the poor outcome group clearly needs close monitoring, especially when these patients stop taking a certain drug or when they are tapering their doses. Future studies should try to identify the most effective treatment strategy for each group.

In conclusion, the present results confirm the effectiveness of T2T strategies in reaching remission in early RA patients in daily clinical practice. However, the disease activity of early RA patients following a T2T strategy does not follow a linear pattern, nor is a single developmental trajectory of disease activity applicable to all patients. Although most patients do show a favorable disease course, future studies should attempt to identify more specific risk factors for poor outcome to enable early identification of patients who might benefit from an alternative therapeutic approach.

Acknowledgements

The authors would like to thank all of the patients for their participation in this study and the rheumatologists from the Medisch Spectrum Twente (Enschede), Ziekenhuisgroep Twente (Almelo, Hengelo), Universitair Medisch Centrum Groningen (Groningen), Isala Klinieken (Zwolle), TweeSteden Ziekenhuis (Tilburg), and UMC St Radboud (Nijmegen) for patient recruitment and data collection.

References

1. Prevoo MLL, van ‘t Hof MA, Kuper HH, van Leeuwen MA, van de Putte LBA, van Riel PLCM. Modified disease activity scores that include twenty-eight-joint counts. Arthritis Rheum. 1995;38:44-8.
2. Vermeer M, Kuper HH, Hoekstra M, Haagsma CJ, Posthumus MD, Brus HLM, et al. Implementation of a treat-to-target strategy in very early rheumatoid arthritis. Arthritis Rheum. 2011;63(10):2865-72.
3. Schipper LG, Vermeer M, Kuper HH, Hoekstra MO, Haagsma CJ, Den Broeder AA, et al. A tight control treatment strategy aiming for remission in early rheumatoid arthritis is more effective than usual care treatment in daily clinical practice: a study of two cohorts in the Dutch Rheumatoid Arthritis Monitoring registry. Ann Rheum Dis. 2012;71(6):845-50.
4. Vermeer M, Kuper HH, Moens HJ, Drossaers-Bakker KW, van der Bijl AE, van Riel PLCM, et al. Sustained beneficial effects of a protocolized treat-to-target strategy in very early rheumatoid arthritis: Three-year results of the Dutch Rheumatoid Arthritis Monitoring remission induction cohort. Arthritis Care Res. 2013;65(8):1219-26.
5. Bakker MF, Jacobs JW, Welsing PM, Verstappen SM, Tekstra J, Ton E, et al. Low-dose prednisone inclusion in a methotrexate-based, tight control strategy for early rheumatoid arthritis: a randomized trial. Ann Intern Med. 2012;156(5):329-39.

6. Wang M, Bodner TE. Growth mixture modeling: Identifying and predicting unobserved subpopulations with logitudinal data. Organizational Research Methods. 2007;10(4):635-56.
7. Jung T, Wickrama KAS. An introduction to latent class growth analysis and growth mixture modeling. Social and Personality Psychology Compass. 2008;2(1):302-17.
8. Le Grande MR, Elliott PC, Murphy BM, Worcester MUC, Higgins RO, Ernest CS, et al. Health related quality of life trajectories and predictors following coronary artery bypass surgery. Health and Quality of Life Outcomes. 2006;4:49.
9. Dunn KM, Jordan KP, Mancl L, Drangsholt MT, Le Resche L. Trajectories of pain in adolescents: A prospective cohort study. Pain. 2011;152:66-73.
10. Verkleij SPJ, Hoekstra T, Rozendaal RM, Waarsing JH, Koes BW, Luijsterburg PAJ, et al. Defining discriminative pain trajectories in hip osteoarthritis over a 2-year time period. Ann Rheum Dis. 2012;71(9):1517-23.
11. Norton S, Sacker A, Young A, Done J. Distinct psychological distress trajectories in rheumatoid arthritis: Findings from an inception cohort. Journal of Psychosomatic Research. 2011;71:290–5.
12. Ware JE, Kosinski M, Dewey JE. How to Score Version Two of the SF-36 Health Survey. Lincoln, RI: QualityMetric, Incorporated; 2000.
13. van der Sluijs Veer G, Soons JW. A time-resolved fluoroimmuno assay of the IgM-rheumatoid factor. Eur J Clin Chem Clin Biochem. 1992;30(5):301-5.
14. Muthén B, Muthén LK. Integrating person-centered and variable-centered analyses: Growth mixture modeling with latent trajectory classes. Clinical and Experimental Research. 2000;24(6):882-91.
15. Forslind K, Hafström I, Ahlmén M, Svensson B. Sex: a major predictor of remission in early rheumatoid arthritis? Ann Rheum Dis. 2007;66:46-52.
16. Iikuni N, Sato E, Hoshi M, Inoue E, Taniguchi A, Hara M, et al. The influence of sex on patients with rheumatoid arthritis in a large observational cohort. J Rheumatol. 2009;36(3):508-11.
17. Sokka T, Toloza S, Cutolo M, Kautiainen H, Makinen H, Gogus F, et al. Women, men, and rheumatoid arthritis: analyses of disease activity, disease characteristics, and treatments in the QUEST-RA study. Arthritis Res Ther. 2009;11(1):R7.
18. Tengstrand B, Ahlmén M, Hafström I. The influence of sex on rheumatoid arthritis: A prospective study of onset and outcome after 2 years. J Rheumatol. 2004;31(2):214-22.
19. Aletaha D, Alasti F, Smolen JS. Rheumatoid factor determines structural progression of rheumatoid arthritis dependent and independent of disease activity. Ann Rheum Dis. 2013;72(6):875-80.

Chapter 8

Further optimization of the reliability of the 28-joint Disease Activity Score in patients with early rheumatoid arthritis

L. Siemons
P.M. ten Klooster
H.E. Vonkeman
M.A.F.J. van de Laar
C.A.W. Glas

PLOS ONE (in press)

Abstract

Background: The 28-joint Disease Activity Score (DAS28) combines scores on a 28-tender and swollen joint count (TJC28 and SJC28), a patient-reported measure for general health (GH), and an inflammatory marker (either the erythrocyte sedimentation rate [ESR] or the C-reactive protein [CRP]) into a composite measure of disease activity in rheumatoid arthritis (RA). This study examined the reliability of the DAS28 in patients with early RA using principles from generalizability theory and evaluated whether it could be increased by adjusting individual DAS28 component weights.

Methods: Patients were drawn from the DREAM registry and classified into a "fast response" group (N=466) and "slow response" group (N=80), depending on their pace of reaching remission. Composite reliabilities of the DAS28-ESR and DAS28-CRP were determined with the individual components' reliability, weights, variances, error variances, correlations and covariances. Weight optimization was performed by minimizing the error variance of the index.

Results: Composite reliabilities of 0.85 and 0.86 were found for the DAS28-ESR and DAS28-CRP, respectively, and were approximately equal across patients groups. Component reliabilities, however, varied widely both within and between sub-groups, ranging from 0.614 for GH ("slow response" group) to 0.912 for ESR ("fast response" group). Weight optimization increased composite reliability even further. In the total and "fast response" groups, this was achieved mostly by decreasing the weight of the TJC28 and GH. In the "slow response" group, though, the weights of the TJC28 and SJC28 were increased, while those of the inflammatory markers and GH were substantially decreased.

Conclusions: The DAS28-ESR and the DAS28-CRP are reliable instruments for assessing disease activity in early RA and reliability can be increased even further by adjusting component weights. Given the low reliability and weightings of the general health component across subgroups it is recommended to explore alternative patient-reported outcome measures for inclusion in the DAS28.

Introduction

If a concept or condition is too complex to measure with a single instrument, multiple measurements are often combined into a linear composite score (i.e. an index measure). For rheumatoid arthritis (RA) the 28-joint Disease Activity Score (DAS28) is such an index measure, widely used for determining a patient's degree of disease activity [1]. It consists of 4 different individual components: a 28-tender joint count, a 28-swollen joint count, a patient-reported rating of general health, and a non-specific acute phase reactant of systemic inflammation which can be either the erythrocyte sedimentation rate (ESR) or the C-reactive protein (CRP). Each component has its own specific weight in the composite score, based on canonical discriminant functions for classifying high and low disease activity.

The DAS28 has received much attention over the years and has been shown to be a valid measure [1, 2]. However, since reliability is a prerequisite for validity [3], the index should be reliable as well. While several studies have already determined the reliability of the DAS28 using Cronbach's Alpha [2, 4, 5], it is not appropriate to use this internal consistency measure with an index measure, as opposed to scales. Where a scale consists of correlated items which all measure the same construct, an index consists of items which are not necessarily highly correlated but which are considered indicators because they themselves define the construct [6]. As a result, the components might measure completely different aspects of disease activity, which is also the case with the DAS28. This poses significant methodological challenges for reliability testing. To overcome these challenges, generalizability theory [7, 8] can be used to estimate the reliability of an index score by disentangling different sources of error.

As such, the first aim of this study was to determine the reliability of the DAS28 using generalizability theory. Since reliability is a concept defined relative to a specific population of patients [9], the second aim of this study was to examine whether the reliability of the index is acceptably high in relevant subpopulations and whether this reliability can be increased by adjusting the weightings of the individual component scores within the DAS28, which was shown to be the case.

Methods

Ethics Statement

As evaluated by the ethics committees of the participating hospitals, and in accordance with Dutch law, no ethical approval was required because data collection took place in daily clinical practice. Nevertheless, informed consent was obtained from each patient.

Patients

Patients were drawn from the remission induction cohort of the Dutch Rheumatoid Arthritis Monitoring (DREAM) registry [10]. This observational multicenter cohort started in 2006 and, although patient recruitment for this cohort has stopped in 2012, data collection of included patients is still ongoing. For this study, all data available at 0, 3, 6, 9, and 12 months were accessible for analyses. To be eligible for inclusion in the cohort, patients were DMARD and prednisolone naïve, they were allowed to have a maximum symptom duration of 1 year, and they needed to be 18 years or older. Additionally, they were not in remission, as measured with the DAS28. Early RA classification was based on a clinical diagnosis by the rheumatologist. For the present study, patients were classified into 2 groups as identified in the study by Siemons et al. [11]: a “fast response” group of patients quickly reaching remission and a “slow response” group of patients reaching remission at a slower pace. Because reliability calculations are sample dependent slight differences can be expected between the two response groups.

Measures

The primary measures of interest were the 28-tender joint count (TJC28), the 28-swollen joint count (SJC28), a 100 millimeter visual analog scale on general health (GH: where 0=very good and 100=very bad), the erythrocyte sedimentation rate (ESR), and the C-reactive protein (CRP). Using these variables, the DAS28-ESR and DAS28-CRP were calculated as follows [12]:

$$\text{DAS28-ESR} = 0.56 * \sqrt{\mathrm{TJC28}} + 0.28 * \sqrt{\mathrm{SJC28}} + 0.70 * \mathrm{Ln(ESR)} + 0.014 * \mathrm{GH} \quad [1]$$

$$\text{DAS28-CRP} = 0.56 * \sqrt{\mathrm{TJC28}} + 0.28 * \sqrt{\mathrm{SJC28}} + 0.36 * \mathrm{Ln(CRP + 1)} + 0.014 * \mathrm{GH} + 0.96 \quad [2]$$

Patients additionally completed the Health Assessment Questionnaire (HAQ) which measures physical functioning [13] and the 36-item Short Form Health Survey (SF-36) which assesses physical and mental health status [14].

Statistical analyses

Baseline between-group comparisons were made using independent t-tests for normally distributed continuous variables, Kruskal Wallis tests for variables with skewed distributions, and Chi-Square tests for dichotomous variables.

Although a DAS28 score is assumed to represent a patient’s disease activity, in reality this score consists of two parts: 1) the actual (true) score of that patient’s disease activity, and 2) random measurement errors [3]. Errors give rise to an under- or overestimation of the true score. This might be due to, among others, certain distractions during test administration, the patient's mood while filling out the test, or a misreading of the items. Reliability is a representation of measurement consistency; it is the ratio between true score variance and observed total score variance, where the latter consists of a true score part and an error part [3]. Higher error variance leads to lower reliability. However, although these basic principles do apply to the individual components of the DAS28, the composite reliability of an index also depends on the interrelationships of its components. Composite reliability is a function of the reliability of the individual components, the weights that are assigned to the components as reflected in the DAS28 formulas, the variances and error variances of the component scores, and the correlations and covariances between the different components. All this can be combined into the following formula [8, 15]:

$$r = 1 - \frac{\sum_{i=1}^{n}(w_i^2 \sigma_{e,Xi}^2)}{\sum_{i=1}^{n}(w_i^2 \sigma_{Xi}^2) + \sum_{i=1}^{n}\sum_{j(\neq i)=1}^{n}(w_i w_j \sigma_{Xi,Xj})} \quad [3]$$

Where:

I. w_i is the weight of component i, as defined in the DAS28 formulas 1 and 2 described above;

II. σ_{Xi}^2 is the observed variance of component *i*;

III. $\sigma_{e,Xi}^2$ is the error variance of component *i*, which is a function of the component reliability and the observed variance: $\sigma_{e,Xi}^2$ = (1 - reliability of component i) * σ_{Xi}^2 ;

IV. $\sigma_{Xi,Xj}$ is the covariance between the two components, which can be rewritten as: *Correlation* $_{Xi,Xj}$ * *Standard deviation* $_{Xi}$ * *Standard deviation* $_{Xj}$

Note that the nominator of the ratio in formula 3 is the error variance of the index, while the denominator is the total variance.

In this study, the error variances of joint count components were calculated from their observed variance and reliability levels, whereas the error variances of GH and both inflammatory measures were obtained from univariate linear regression analyses. However, two significant problems arose during component reliability computations.

First, it was not appropriate to calculate the reliabilities of the joint count components with Cronbach's Alpha, since Cronbach's Alpha assumes the total score to be a linear combination of all items (i.e. each item is regarded as a parallel test of the other component items) whereas the joint count components are square root transformed in the DAS28. Consequently, split-half reliabilities were determined instead. As demonstrated by Siemons et al. [16], RA is characterized by a definite left-right symmetry of joint involvement. Consequently, the square root sum scores of the left and right joints were considered to represent two parallel tests and their correlation was used as an estimate of reliability.

Second, given the structure of the cohort, with time frames of several weeks or even months between consecutive measurements, it was not possible to perform proper test-retest reliabilities for the single-item components (i.e. ESR, CRP, and GH). Therefore, a generalizability theory principle was used to determine reliability. After running a univariate general linear model analysis on a longitudinal dataset (including all available data at 0, 3, 6, 9, and 12 months) the person variance could be separated from the time variance and reliability could be calculated as the ratio between person variance and total variance.

All computations of reliabilities were performed on the total patient group as well as on the two identified subgroups using SPSS version 21.0.

After calculating the composite reliabilities, it was investigated whether the reliability in both the total as well as the specific patient groups could be optimized by adjusting the component’s weights. Optimal weights were computed by minimizing the error variance of the index, that is, the nominator of the ratio in formula 3, subject to the constraint that the total variance of the index does not change. The resulting quadratic optimization problem under quadratic constraints was solved using a procedure developed by Albers, Critchley, and Gower [17]. For more applications, refer to Albers, Critchley, and Gower [18].

Results

Patients

A total of 565 patients were included for analysis; 466 from the “fast response” group and 80 from the “slow response” group. Overall, they had a mean age of approximately 58 years, the majority were women (63%) and rheumatoid factor positive (54%), and their baseline disease was characterized by several tender and swollen joints as well as by a diminished physical health status (Table 1).

No significant differences were found in age, mental health, or rheumatoid factor positivity between the fast and slow response group, but the slow response group did contain significantly more women. Additionally, the slow response group had a significantly worse disease condition, characterized by higher DAS28 scores, a poorer state of general and physical health, higher levels of inflammatory markers, and a higher number of tender and swollen joints, as compared to the fast response group.

Table 1 – Patient characteristics and group comparisons at inclusion.

Variable*	**Mean (SD) or Median (range)* Total group**	**Mean (SD) or Median (range)* "Fast response" group**	**Mean (SD) or Median (range)* "Slow response" group**	**Significance (*p*) group comparisons**
Gender (female)	342/546 (62.6%)	283/466 (60.7%)	59/80 (73.8%)	0.026~
Age (years)	57.96 (14.31)	57.85 (14.62)	58.59 (12.41)	0.672¶
DAS28-ESR	4.28 (1.45)	4.11 (1.42)	5.22 (1.29)	<0.001¶
DAS28-CRP	4.02 (1.31)	3.88 (1.28)	4.79 (1.24)	<0.001¶
28-Tender joint count	3 (0-28)	3 (0-28)	6 (0-28)	<0.001#
28-Swollen joint count	5 (0-24)	4 (0-24)	6.50 (0-24)	0.001#
GH	44.32 (26.11)	42.52 (25.76)	54.82 (25.83)	<0.001¶
ESR (mm/hour)	21.50 (1-120)	20 (1-111)	30 (7-120)	<0.001#
CRP (mg/l)	9.00 (1-158)	6 (1-158)	11 (1-115)	0.004#
Rheumatoid factor +	274/506 (54.2%)	233/431 (54.1%)	41/75 (54.7%)	0.923~
SF36 – physical health	37.17 (9.09)	37.93 (9.09)	32.91 (7.88)	<0.001¶
SF36 – mental health	48.16 (11.67)	48.55 (11.37)	45.99 (13.07)	0.094¶
HAQ	1.00 (0.72)	0.92 (0.70)	1.41 (0.66)	<0.001¶

** The values for gender and rheumatoid factor positivity are the number of patients/number of patients assessed (%). DAS28 = 28-joint Disease Activity Score, GH = general health, ESR = erythrocyte sedimentation rate, CRP = C-reactive protein, SF36 = Short Form Health Survey with 36 items, HAQ = Health Assessment Questionnaire.*

# Group comparisons performed with Kruskal Wallis tests;

~ Group comparisons performed with Chi-Square tests.

¶ Group comparisons performed with independent t-tests.

Reliability calculations

Tables 2-4 show the variances, error variances and reliabilities of the individual components together with their correlations and covariances. All correlations were small to moderate, as is not unusual in index measures. Component reliabilities varied widely both within and between sub-groups, ranging from 0.614 for GH in the "slow response" group to 0.912 for ESR in the "fast response" group. Furthermore, ESR component reliabilities were highest and the TJC28 outperformed the SJC28 across all groups.

Using formula 3 resulted in overall reliabilities of 0.85 and 0.86 for the DAS28-ESR and DAS28-CRP composites, respectively. Sub-analyses showed DAS28-ESR reliabilities of 0.85 and 0.82 and DAS28-CRP reliabilities of 0.85 and 0.84 for the "fast response" and "slow response" group, respectively. These results demonstrate that both DAS28 scores are approximately equally reliable across patient groups, that is, the differences were only small and all reliability levels were high (>0.80).

Table 2 – Variances, error variances, and reliabilities of the DAS28 components.

	Total patient group (N=546)			"Fast response" group (N=466)			"Slow response" group (N=80)		
Component	**Variance**	**Error variance**	**Component Reliability**	**Variance**	**Error variance**	**Component Reliability**	**Variance**	**Error variance**	**Component Reliability**
$\sqrt{\text{TJC28}}$	1.756	0.400	0.772	1.580	0.409	0.741	2.085	0.325	0.844
$\sqrt{\text{SJC28}}$	1.521	0.420	0.724	1.522	0.425	0.721	1.342	0.394	0.706
GH	681.624	269.687	0.783	663.439	247.769	0.766	666.917	359.309	0.614
LN(ESR)	0.779	0.193	0.913	0.807	0.191	0.912	0.477	0.191	0.878
LN(CRP)	0.909	0.246	0.805	0.920	0.231	0.806	0.771	0.335	0.773

TJC28 = 28-tender joint count, SJC28 = 28-swollen joint count, GH = general health, ESR = erythrocyte sedimentation rate, CRP = C-reactive protein

Table 3 – Pearson's correlations and covariances of DAS28 components in total patient sample.

Covariance **Pearson's Correlation**	$\sqrt{\text{TJC28}}$	$\sqrt{\text{SJC28}}$	**GH**	**LN(ESR)**	**LN(CRP)**
$\sqrt{\text{TJC28}}$	----	0.909	14.818	0.152	0.207
$\sqrt{\text{SJC28}}$	0.555**	----	8.467	0.337	0.402
GH	0.427**	0.262**	----	4.129	5.181
LN(ESR)	0.131**	0.309**	0.179**	----	0.488
LN(CRP)	0.164**	0.343**	0.207**	0.581**	----

*** p<0.01, TJC28 = 28-tender joint count, SJC28 = 28-swollen joint count, GH = general health, ESR = erythrocyte sedimentation rate, CRP = C-reactive protein*

Table 4 – Pearson's correlations of the DAS28 components in the "fast response" and "slow response" group.

Pearson's correlation group 2 (N=80) **Pearson's correlation group 1 (N=466)**	$\sqrt{\text{TJC28}}$	$\sqrt{\text{SJC28}}$	**GH**	**LN(ESR)**	**LN(CRP)**
$\sqrt{\text{TJC28}}$	1	0.547**	0.313**	-0.09	-0.032
$\sqrt{\text{SJC28}}$	0.545**	1	0.288*	0.143	0.263*
GH	0.423**	0.239**	1	0.067	0.104
LN(ESR)	0.125**	0.312**	0.166**	1	0.578**
LN(CRP)	0.171**	0.342**	0.205**	0.574**	1

*** p<0.01, TJC28 = 28-tender joint count, SJC28 = 28-swollen joint count, GH = general health, ESR = erythrocyte sedimentation rate, CRP = C-reactive protein*

Optimizing reliability

The results of the optimization of composite reliability by adjusting the weights are shown in Table 5. The original weights are given in the third column. The next three columns give the estimated optimal weights in the total group and in the “fast response” and “slow response” groups, respectively. Further, in the rows labeled “Reliability”, the reliabilities using the original weights and the optimal weights are given.

In all groups, reliabilities increased after weight optimization. The largest gains were obtained in the total and “fast response” groups by decreasing the weight of the TJC and GH. In the smaller slow response group, on the other hand, the weights of the TJC28 and SJC28 were increased, while the weights of the inflammatory markers and GH were substantially decreased.

Table 5 – Reliabilities of the index measures and optimal weights in subpopulations.

Index measure	Component *	Original Weight	Total patient group (N=546)	“Fast response” group (N=466)	“Slow response” group (N=80)
DAS28-ESR	$\sqrt{\text{TJC28}}$	0.560 (0.742)	0.226 (0.299)	0.156 (0.196)	0.739 (1.067)
	$\sqrt{\text{SJC28}}$	0.280 (0.345)	0.271 (0.334)	0.226 (0.279)	0.355 (0.411)
	GH	0.014 (0.366)	0.004 (0.104)	0.004 (0.103)	0.004 (0.103)
	LN(ESR)	0.700 (0.618)	0.663 (0.585)	0.650 (0.584)	0.052 (0.036)
Reliability	**Original Weights**		0.854	0.848	0.821
	Optimal Weights		0.933	0.942	0.859
DAS28-CRP	$\sqrt{\text{TJC28}}$	0.560 (0.742)	0.374 (0.496)	0.247 (0.310)	0.720 (1.040)
	$\sqrt{\text{SJC28}}$	0.280 (0.345)	0.378 (0.466)	0.298 (0.368)	0.356 (0.412)
	GH	0.014 (0.366)	0.006 (0.157)	0.005 (0.129)	0.004 (0.103)
	LN(CRP)	0.360 (0.343)	0.522 (0.498)	0.571 (0.548)	0.028 (0.025)
Reliability	**Original Weights**		0.858	0.852	0.845
	Optimal Weights		0.888	0.911	0.858

The weights between brackets are the standardized values by fixing the component variances at 1.
** TJC28 = 28-tender joint count, SJC28 = 28-swollen joint count, GH = general health,*
ESR = erythrocyte sedimentation rate, CRP = C-reactive protein

Discussion

Overall, composite reliability levels of 0.85 and 0.86 were found for the DAS28-ESR and DAS28-CRP, respectively. This is sufficiently high for group use and around common thresholds considered sufficient for individual use [19, 20], justifying the use of the DAS28 in both clinical research and clinical practice. Moreover, reliability could be increased even further by optimizing the component weights.

Several findings are worth mentioning when comparing the individual component reliabilities or when optimizing the component weights. At first, it can be observed that ESR had the highest reliability of all DAS28 components. The finding that this measure of inflammation was also more reliable than the CRP measure might be explained by their different responsiveness to changes in inflammatory stimuli. While the ESR is a relatively stable measure over time, which responds slowly to changes and reflects the disease activity of the past few weeks, the CRP fluctuates more heavily due to a more rapid response to short-term changes in the inflammatory stimuli [12, 21, 22]. The ESR is also given a higher weighting than the CRP in the DAS28 formulas and that remained to be the case after weight optimization.

When looking at the joint count reliabilities, the swollen joint count had a lower reliability than the tender joint count, consistent with findings from previous studies [9, 23, 24]. Joint counts are sometimes referred to as a semi-objective clinical measures [1] and, as discussed by Pincus [25], they have been shown to be poorly reproducible. Large intra- and interobserver variability is commonly found especially in the swollen joint count [24]. This might be explained by a higher dependency of the swollen joint assessment on factors like the assessors' levels of training and experience, a lack of standardization in examination methods, unclear definitions of swelling, or the degree of joint deformity [24-26]. After weight optimization, however, the weight of the tender joint count was substantially lowered, even below the weight of the swollen joint count. This corresponds with the common clinical perspective of disease activity in RA. Joint swelling is usually considered to be a more representative measure of inflammation than joint tenderness [27] and has been shown to play a major role in the physician's assessment of disease activity [28].

Finally, it can be observed that the patient reported degree of GH in the "slow response" group had the lowest reliability of all components, even below the recommended reliability threshold for group use (r>0.70). Its weight was also substantially decreased after weight optimization. This could be a confirmation of the weakness of this component, as the inclusion of GH in the DAS28 has been often criticized. For instance, previous studies have shown elevated GH scores while none of the other DAS28 components showed any sign of an active disease, possibly due to effects beyond the clinical inflammatory processes of RA [29]. Also, GH ratings have been shown to be different across patients with similar DAS28 scores, dependent on the moment of administration, possibly caused by a response shift [30]. The GH component is also the most subjective component of the DAS28 and, therefore, more susceptible to measurement error. Although it could be argued to solely include the more objective clinical measures, the inclusion of a patient-driven component is desirable given the increased awareness of the importance of the patient perspective in assessing disease activity since the 1980s [31] as reflected by their inclusion in the provisional ACR/EULAR definition of remission in RA [32] as well as in the preliminary core set of disease activity measures [33]. Disease activity in RA is a multifactorial concept and appears to be best measured by both objective clinical measures and patient-reported outcomes as they each address a different aspect of disease activity [27]. Therefore, it would be interesting to explore other, more reliable, patient-reported outcome measures for inclusion in the DAS28. What measure would be best warrants further research. Measures of pain or fatigue appear to be promising alternatives, given the recognition of pain as one of the most important determinants of a patient's global assessment of disease activity [34-36] and the recommendation to measure fatigue in addition to the other core set measures of RA [37]. But, of course, other patient-reported measures can also be explored.

A possible limitation of the current study is the difference in sample size between the fast response and slow response group. Though this might cause the results of the "fast response" group to be more robust than the results of the "slow response" group, the importance of the tender joint count in the slow response group is consistent with the general belief that approximately 10-20% of patients with RA have secondary fibromyalgia (FM) [38, 39]. FM patients tend to have a lower pain threshold [38, 39] which might explain the higher relevance of the TJC28 rating in this patient group, increasing our confidence in the robustness of the results.

Conclusions

The DAS28-ESR and the DAS28-CRP are both reliable instruments for assessing disease activity in early RA although reliability can be increased even further by adjusting the individual component weights. Overall, the findings suggest that the largest gains in reliability can be achieved by substantially lowering the weights of the tender joint count and patient-reported general health. Future studies should explore the possibilities of including a better indicator of the patient perspective in the disease activity score.

References

1. Prevoo ML, van 't Hof MA, Kuper HH, van Leeuwen MA, van de Putte LB, van Riel PL. Modified disease activity scores that include twenty-eight–joint counts: development and validation in a prospective longitudinal study of patients with rheumatoid arthritis. Arthritis Rheum. 1995;38(1):44-8.
2. Salaffi F, Cimmino MA, Leardini G, Gasparini S, Grassi W. Disease activity assessment of rheumatoid arthritis in daily practice: validity, internal consistency, reliability and congruency of the Disease Activity Score including 28 joints (DAS28) compared with the Clinical Disease Activity Index (CDAI). Clin Exp Rheumatol. 2009;27(4):552-9.
3. Lord FM, Novick MR. Statistical theories of mental test scores. Reading, MA: Addison-Welsley Publishing Company; 1968.
4. Leeb BF, Andel I, Sautner J, Bogdan M, Maktari A, Nothnagl T, et al. Disease activity measurement of rheumatoid arthritis: Comparison of the simplified disease activity index (SDAI) and the disease activity score including 28 joints (DAS28) in daily routine. Arthritis Rheum. 2005;53(1):56-60.
5. Leeb BF, Sautner J, Mai HT, Haindl PM, Deutsch C, Rintelen B. A comparison of patient questionnaires and composite indexes in routine care of rheumatoid arthritis patients. Joint Bone Spine. 2009;76(6):658-64.
6. Crosby RA, DiClemente RJ, Salazar LF. Research methods in health promotion. San Francisco, CA: Jossey-Bass; 2006.
7. Brennan RL. Generalizability theory. Instructional Topics in Educational Measurement. 1992;11(4):225-32.
8. He Q. Estimating the reliability of composite scores. Coventry: Ofqual; 2009.
9. Lassere MN, van der Heijde D, Johnson KR, Boers M, Edmonds J. Reliability of measures of disease activity and disease damage in rheumatoid arthritis: implications for smallest detectable difference, minimal clinically important difference, and analysis of treatment effects in randomized controlled trials. J Rheumatol. 2001;28(4):892-903.
10. Vermeer M, Kuper HH, Hoekstra M, Haagsma CJ, Posthumus MD, Brus HL, et al. Implementation of a treat-to-target strategy in very early rheumatoid arthritis: results of the Dutch Rheumatoid Arthritis Monitoring remission induction cohort study. Arthritis Rheum. 2011;63(10):2865-72.
11. Siemons L, ten Klooster PM, Vonkeman HE, Glas CAW, van de Laar MAFJ. Distinct trajectories of disease activity over the first year in early rheumatoid arthritis patients following a treat-to-target strategy. Arthritis Care Res. In press.
12. Van Riel PL, Fransen J, Scott DL. EULAR handbook of clinical assessments in rheumatoid arthritis. Alphen aan den Rijn: van Zuiden Communications; 2004.
13. Fries JF, Spitz P, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. Arthritis Rheum. 1980;23:137-45.
14. Ware JE, Kosinski M, Dewey JE. How to Score Version Two of the SF-36 Health Survey. Lincoln, RI: QualityMetric, Incorporated; 2000.
15. Mosier CI. On the reliability of a weighted composite. Psychometrika 1943;8:161-8.
16. Siemons L, ten Klooster PM, Taal E, Kuper IH, van Riel PLCM, van de Laar MAFJ, et al. Validating the 28-tender joint count using item response theory. J Rheumatol. 2011;38(12):2557-64.
17. Albers CJ, Critchley F, Gower JC. Quadratic minimisation problems in statistics. Journal of Multivariate Analysis. 2011;102:698–713.

18. Albers CJ, Critchley F, Gower JC. Applications of quadratic minimisation problems in statistics. Journal of Multivariate Analysis. 2011;102(714-722).
19. Tennant A, Conaghan PG. The rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a rasch paper? Arthritis Rheum. 2007;57:1358-62.
20. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol. 2007;60(1):34-42.
21. Firestein GS, Budd RC, Harris Jr ED, McInnes IB, Ruddy S, Sergent JS. Kelly's textbook of rheumatology. Philadelphia: Saunders Elsevier; 2009.
22. Kushner I. C-reactive protein in rheumatology. Arthritis Rheum. 1991;34(8):1065-68.
23. Siemons L, Ten Klooster PM, Taal E, Kuper IH, van Riel PLCM, Glas CAW, et al. Contribution of assessing forefoot joints in early rheumatoid arthritis patients: insights from item response theory. Arthritis Care Res. 2013;65(2):212-9.
24. Cheung PP, Gossec L, Mak A, March L. Reliability of joint count assessment in rheumatoid arthritis: A systematic literature review. Semin Arthritis Rheum. In press.
25. Pincus T. Limitations of a quantitative swollen and tender joint count to assess and monitor patients with rheumatoid arthritis. Bull NYU Hosp Jt Dis. 2008;66(3):216-23.
26. Marhadour T, Jousse-Joulin S, Chalès G, Grange L, Hacquard C, Loeuille D, et al. Reproducibility of joint swelling assessments in long-lasting rheumatoid arthritis: influence on Disease Activity Score-28 values (SEA-Repro study part I). J Rheumatol. 2010;37(5):932-7.
27. Wolfe F. The prognosis of rheumatoid arthritis: Assessment of disease activity and disease severity in the clinic. Am J Med. 1997;103(6A):12S-8S.
28. Soubrier M, Zerkak D, Gossec L, Ayral X, Roux C, Dougados M. Which variables best predict change in rheumatoid arthritis therapy in daily clinical practice? J Rheumatol. 2006;33(7):1243-6.
29. Vermeer M, Kuper HH, van der Bijl AE, Baan H, Posthumus MD, Brus HLM, et al. The provisional ACR/EULAR definition of remission in RA: a comment on the patient global assessment criterion. Rheumatology (Oxford). 2012;51(6):1076-80.
30. Kievit W, Welsing PMJ, Adang EMM, Eijsbouts AM, Krabbe PFM, van Riel PLCM. Comment on the use of self-reporting instruments to assess patients with rheumatoid arthritis: the longitudinal association between the DAS28 and the VAS general health. Arthritis Rheum. 2006;55(5):745-50.
31. Fries JF, Bruce B, Cella D. The promise of PROMIS: Using item response theory to improve assessment of patient-reported outcomes. Clin Exp Rheumatol. 2005;23(39):S53-S7.
32. Felson DT, Smolen JS, Wells G, Zhang B, van Tuyl LHD, Funovits J, et al. American College of Rheumatology/European League Against Rheumatism provisional definition of remission in rheumatoid arthritis for clinical trials. Arthritis Rheum. 2011;63(3):573-86.
33. Felson DT, Anderson JJ, Boers M, Bombardier C, Chernoff M, Fried B, et al. The American College of Rheumatology preliminary core set of disease activity measures for rheumatoid arthritis clinical trials. The Committee on Outcome Measures in Rheumatoid Arthritis Clinical Trials. Arthritis Rheum. 1993;36(6):729-40.
34. Studenic P, Radner H, Smolen JS, Aletaha D. Discrepancies between patients and physicians in their perceptions of rheumatoid arthritis disease activity. Arthritis Rheum. 2012;64(9):2814-23.

35. Khan NA, Spencer HJ, Abda E, Aggarwal A, Alten R, Ancuta C, et al. Determinants of discordance in patients' and physicians' rating of rheumatoid arthritis disease activity. Arthritis Care Res. 2012;64(2):206-14.
36. Markenson JA, Koenig AS, Feng JY, Chaudhari S, Zack DJ, Collier D, et al. Comparison of physician and patient global assessments over time in patients with rheumatoid arthritis: a retrospective analysis from the RADIUS cohort. J Clin Rheumatol. 2013;19(6):317-23.
37. Kirwan JR, Minnock P, Adebajo A, Bresnihan B, Choy E, de Wit M, et al. Patient perspective: fatigue as a recommended patient centered outcome measure in rheumatoid arthritis. J Rheumatol. 2007;34(5):1174-7.
38. Coury F, Rossat A, Tebib A, Letroublon MC, Gagnard A, Fantino B, et al. Rheumatoid arthritis and fibromyalgia: a frequent unrelated association complicating disease management. J Rheumatol. 2009;36(1):58-62.
39. Leeb BF, Andel I, Sautner J, Nothnagl T, Rintelen B. The DAS28 in rheumatoid arthritis and fibromyalgia patients. Rheumatology. 2004;43(12):1504-7.

Chapter 9

General discussion

Rheumatoid arthritis (RA) is a chronic disease which cannot yet be cured. In order to control the disease as much as possible, early, aggressive, protocolized treatment strategies have been proposed. They have been proven to be effective in mitigating a patient’s disease activity in clinical practice, resulting in less tender and swollen joints and an improved physical functioning and wellbeing [1, 2]. However, these disease-activity-driven treatment strategies require an optimal measurement of disease activity. This thesis has addressed some of the concerns that have been raised about the current use of the Disease Activity Score for 28 joints (DAS28) and its findings can be used to further improve the assessment of disease activity in early RA.

Joint counts

Symmetrically inflamed joints characterize RA. Consequently, joint counts belong to the core set measures of disease activity and their use is highly recommended [3-5]. However, the assessment of all joints would not only be unfeasible in clinical practice but probably redundant as well. For instance, as shown in Chapter 4, the omission of the forefoot joints did not reduce the measurement range and measurement precision of the 28-tender and 28–swollen joint count in patients with early RA [6]. Additionally, the omission of certain joints seems justified because of major assessment difficulties or because of their high sensitivity to influences that go beyond RA-specific disease activity. Of course, this does not imply that these omitted joints aren’t important in the evaluation and management of individual RA patients in daily clinical practice [7-10], but it is believed that the assessment of these joints will not necessarily provide additional quantitative clinical information that is relevant for the assessment and monitoring of the disease activity of early RA patients following a treatment protocol [7, 8, 10-12].

Yet further research is required to establish whether or not the current selection of included joints is most optimal, given the large discrepancy between the patient and joint distributions as shown in Chapter 4. The current joint counts cannot discriminate well between patients with minor joint tenderness or swelling. As such, the contribution of other joints should be evaluated as well and the consequences of these different joint selections for the accuracy of RA disease activity classifications should be taken into account. Finally, it is important to realize that the affected joints at baseline do not necessarily resemble the affected joints at later stages of RA. Consequently, joints that might not be informative at early stages of RA might be at later stages. Both should be taken into account when developing a reduced joint count that is applicable across the disease course. Nonetheless, the final selection will always require a compromise between the optimal statistical selection on the one hand and the clinical feasibility on the other hand.

Laboratory measure

Two different DAS28 indices have been developed: the DAS28-ESR and the DAS28-CRP. Although the ability to calculate DAS28 scores with two different acute phase reactants can be convenient if only one of the two laboratory measures is available, it was shown that the resulting scores cannot be used interchangeably (Chapter 5). The DAS28-CRP tends to indicate a lower disease activity than the DAS28-ESR, resulting in considerable classification differences. Furthermore, the concentrations of both the ESR and CRP are affected by non-inflammatory external influences. Hence, these external influences should be taken into account when interpreting disease activity scores (Chapter 6). The results of these studies did not unambiguously point to one of these inflammatory markers or one of these DAS28 measures as the ultimate measure to use. However, clinicians should be aware of such external influences and score discrepancies. It emphasizes the need of standardization to make scores not only comparable within patients, but also between patients. Yet how this standardization should be accomplished requires further research.

Selecting one of these acute phase reactants as the ”gold standard” might be the easiest solution. Given the rapid response of the CRP to changes in inflammatory stimuli [13-17], the possibility to freeze and store CRP sera [15-17], and its seemingly lower susceptibility to external influences compared to the ESR (Chapter 6), the CRP might be the obvious acute phase reactant to choose for measuring a patient’s current disease activity. Nonetheless, it is important to keep in mind that only a small selection of possibly confounding factors was included in Chapter 6, so no conclusions can be drawn about other potentially distorting effects. Also, clinicians are very familiar with the ESR [13] and may be reluctant to dismiss the ESR that easily. Another, more complex solution might be the preservation of both measures. However, this will require certain adjustments for which further research is required. As discussed in Chapter 5, adding a constant to the DAS28-CRP (or subtracting a constant from the DAS28-ESR), will not resolve the discrepancy problem because score deviations were found to depend on the degree of disease activity. The development of gender- and age-adjusted DAS28 formulas or the specification of age- and gender adjusted cut-off points of disease activity seem more promising, but might be difficult to generalize across RA populations [18]. Also, gender and age are just two of the potentially long list of relevant variables to correct for. This thesis does suggests that the exclusion of the inflammatory markers altogether, as is done with the Clinical Disease Activity Index [19], is not preferable, since the inflammatory markers are the most reliable components of the DAS28 (Chapter 8).

General health

Patient-reported outcome measures belong to the core set of disease activity measures [3]. Several studies have shown that physicians and patients focus on different aspects of the disease when evaluating the patient’s general health status (i.e. how they are doing considering their disease). Where clinicians place more emphasis on the objective clinical measures of disease activity like the joints counts and acute phase reactants, patients tend to focus more on the subjective measures such as pain, fatigue, or physical functioning they encounter on a daily basis [20-22]. Thus, in order to capture both the physician’s and the patient’s perspective of disease activity, it is considered important that indices of disease activity (like the DAS28) include a patient-reported outcome measure as well.

The DAS28 includes a visual analog scale of general health (GH) which asks the patient to give an overall assessment of how they are currently doing, considering all the ways in which the RA influences their lives. Yet whether this general health measure is the most suitable measure to represent the patient perspective remains debatable. Although it prevailed over patient-reported measures of pain or morning stiffness during the development of the DAS [23] and it has been reported to be sensitive to change [3], it received the lowest weighting of all components included in the DAS28 and it has frequently been criticized. GH scores are frequently elevated despite a good clinical disease state [24] and patients with equal disease activity tend to rate their general health differently depending on the moment of administration in the disease course [25]. General health scores are also very difficult to interpret [26]. If a patient is asked “considering all the ways in which the RA influences your life, how are you doing” the answer can be based on multiple dimensions. The patient might take pain or limitations due to joint damage or non-RA related comorbidities into consideration, basing their answer on mostly irreversible and non-RA related aspects, or they might focus on aspects more directly related to current disease activity like the experienced fatigue or the number of tender and swollen joints. As such, the GH rating might measure not only disease activity, but also disease severity or disease impact [27, 28]. Ratings tend to be influenced by non-inflammatory-related problems as (low back) pain, fatigue, and functional limitations [28], as well as by the patient’s mood [27]. Consequently, scores cannot be compared across patients (and maybe neither within patients). In line with all these limitations, the general health component was shown to have the lowest reliability of all DAS28 components and its weight in the DAS28 formula was reduced even further after reliability optimization (Chapter 8).

Given these limitations regarding the current measurement of GH, it would be interesting to explore alternative patient-reported outcome measures for inclusion in the

DAS28 to reflect the patient perspective on disease activity. Pain has been reported as one of the most important determinants of the patient’s global assessment of disease activity [20-22] and might be considered a suitable candidate. However, fatigue has also gained much attention over the years. It has been confirmed as an important outcome measure in RA that should be measured in addition to other core set measures of RA [29]. People with higher fatigue levels are more likely to report their health as fair or poor [30]. Pain and fatigue have also been identified as some of the most important domains for developing a core set definition of a flare [31, 32]. Of course, other patient-reported outcome measures can be explored as well, but these two examples might serve as a starting point. In addition, it would also be interesting to examine whether multi-item questionnaires are to be preferred over single item visual analog scales since the former might be more reliable [33].

Interpreting a DAS28 score

Although current protocolized treatment strategies have been proven to be effective in reducing disease activity in clinical practice [1, 2], a DAS28 score only provides a simplified reflection of a patient’s disease activity and it is possible that patients experience residual disease activity in omitted joints [34-36] or that joints counts are elevated because of non-inflammatory stimuli like a low pain threshold, fluid retention, or the wrong footwear [11, 12]. Furthermore, ESR measures may be affected by abnormally shaped or sized red blood cells, changes in plasma composition, anemia, pregnancy, or certain drugs [13-17], whereas CRP concentrations can be elevated because of other diseases, the presence of bacterial infections, or non-inflammatory influences like sleep deprivation or unhealthy diets [16]. To make things even more complicated, ESR and CRP levels may also be normal in the presence of active disease [35, 37, 38], emphasizing the need to look at multiple measures and not just one. Finally, non-inflammatory stimuli like pain, fatigue, and functional limitations can enhance GH scores [28], as well as the patient’s mood at the moment of administration [27]. Thus, DAS28 scores are difficult to interpret and a thorough investigation of its underlying disease and non-disease related factors is essential in order to promote well-reasoned treatment decisions.

Early RA population

Since current treatment strategies emphasize an aggressive interference early on in the disease, this thesis focused on early RA patients. Patients participated in the Dutch Rheumatoid Arthritis Monitoring (DREAM) remission induction cohort; an observational, multicenter cohort that was established in 2006 to evaluate the effects of a protocolized treat-to-target strategy aimed at (sustained) remission in early RA patients in daily clinical practice [2].

Although this strategy has been shown to be highly effective, several distinct disease activity trajectories could be identified within the early RA patients (Chapter 7). As consistent with clinical experience in RA, the disease activity of the large majority of patients quickly decreased towards a state of remission. A second group responded much slower to the treatment, but did approach remission after approximately 12 months. Finally, a small minority group of patients did have a positive treatment response at first but deteriorated again after 6 months, returning to their baseline levels of disease activity [39].

Although no apparent reasons could be found why some patients tend to follow one trajectory while others follow another, supplementary case history examinations on the minority group showed that their loss of treatment response might be due to a discontinuation of their DMARDs due to side effects, or because of a reduction in their initial dose of prednisolone, emphasizing the need to monitor patients closely when medication adjustments occur. Other indications of the trajectory that someone is most likely going to follow were provided by the reliability analyses on the DAS28. These analyses were performed on the two largest groups only, given the small sample size of the third group. Results showed that the disease activity of the normative group could be measured most reliable by administering the acute phase reactants and the joints counts, i.e. by using the traditional and (semi-)objective clinical measures or RA disease activity. The disease activity of the slow response group, on the other hand, was mainly determined by joint tenderness, which might indicate that this subgroup of patients might actually be suffering from more centralized pain and/or secondary fibromyalgia (FM). It is well known that patients with FM tend to have a lower pain threshold, resulting in higher tender joint counts and general health ratings [40-42]. Furthermore, it has been reported that approximately 10-20% of patients with RA have secondary FM [40, 42], which roughly corresponds with the size of this subsample (14.1%). Rheumatologists should be aware of this heterogeneity in their patient population and the fact that their protocolled treat-to-target strategy does not work equally well for all patients. Patients who’s DAS28 scores are predominantly based on the tender joint count, may be in need of alternative therapeutic interventions. Future studies should attempt to predict which trajectory someone is most likely going to follow in order to promote better patient-tailored treatment strategies.

As all studies in this thesis were performed among patients with early RA, the outcomes of this thesis may not be directly applicable to patients with established RA. Joints that may not be informative at early stages of RA may be at later stages; ESR and CRP concentrations can become dependent on other external factors; and patient-reported ratings of general health can change along the disease course when patients gather more information about their disease, learn to cope with it, or change their expectations. Therefore, more research is needed to explore the stability and generalizability of results towards patients with established RA.

Conclusion

This thesis was aimed at improving our understanding of the complexities behind the measurement of disease activity in early rheumatoid arthritis patients. Disease activity is a multidimensional concept that should be measured and monitored using multiple measures. This should include both (semi)objective clinical measures as well as patient-reported outcome measures because they address different aspects of the disease and, as such, complement each other [32, 38]. Although this thesis showed that the DAS28 is not a flawless instrument and all of its individual components can be criticized to some degree, one should realize that no single index measure will be able to produce optimal results for all patients. As expressed by Fuchs, Books, Callahan, and Pincus, p. 536 [12]:

> *"All quantitative indices of clinical status represent a compromise between comprehensiveness and feasibility."*

Add the statistical aspect to this expression and is reflects the actual compromise even better. The DAS28 is no exception to this rule, and it is important that rheumatologists are aware of its limitations when using it. This thesis provides several possible starting points for the further improvement of RA disease activity measurement. However, it is important to keep in mind that no quantitative measure can fully replace a thorough inquiry of the clinical and patient-reported symptoms experienced by individual patients.

References

1. Schipper LG, Vermeer M, Kuper HH, Hoekstra MO, Haagsma CJ, Den Broeder AA, et al. A tight control treatment strategy aiming for remission in early rheumatoid arthritis is more effective than usual care treatment in daily clinical practice: a study of two cohorts in the Dutch Rheumatoid Arthritis Monitoring registry. Ann Rheum Dis. 2012;71(6):845-50.
2. Vermeer M, Kuper HH, Hoekstra M, Haagsma CJ, Posthumus MD, Brus HLM, et al. Implementation of a treat-to-target strategy in very early rheumatoid arthritis. Arthritis Rheum. 2011;63(10):2865-72.
3. Felson DT, Anderson JJ, Boers M, Bombardier C, Chernoff M, Fried B, et al. The American College of Rheumatology preliminary core set of disease activity measures for rheumatoid arthritis clinical trials. The Committee on Outcome Measures in Rheumatoid Arthritis Clinical Trials. Arthritis Rheum. 1993;36(6):729-40.
4. Felson DT, Smolen JS, Wells G, Zhang B, van Tuyl LHD, Funovits J, et al. American College of Rheumatology/European League Against Rheumatism provisional definition of remission in rheumatoid arthritis for clinical trials. Arthritis Rheum. 2011;63(3):573-86.
5. Van Riel PLCM, Fransen J, Scott DL. Eular handbook of clinical assessments in rheumatoid arthritis. Alphen aan den Rijn: van Zuiden Communications; 2004.
6. Siemons L, ten Klooster PM, Taal E, Kuper IH, van Riel PLCM, Glas CAW, et al. Contribution of assessing forefoot joints in early rheumatoid arthritis patients: insights from item response theory. Arthritis Care Res. 2013;65(2):212-9.
7. 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(1):38-43.
8. Fuchs HA, Pincus T. Reduced joint counts in controlled clinical trials in rheumatoid arthritis. Arthritis Rheum. 1994;37(4):470-5.
9. van Tuyl LHD, Britsemmer K, Wells GA, Smolen JS, Zhang B, Funovits J, et al. Remission in early rheumatoid arthritis defined by 28 joint counts: limited consequences of residual disease activity in the forefeet on outcome. Ann Rheum Dis. 2011;71(1):33-7.
10. Kapral T, Dernoschnig F, Machold KP, Stamm T, Schoels M, Smolen JS, et al. Remission by composite scores in rheumatoid arthritis: are ankles and feet important? Arthritis Res Ther. 2007;9(4):R72.
11. Thompson PW, Kirwan JR. Joint count: A review of old and new articular indices of joint inflammation. Br J Rheumatol. 1995;34:1003-8.
12. Fuchs HA, Brooks RH, Callahan LF, Pincus T. A simplified twenty-eight-joint quantitative articular index in rheumatoid arthritis. Arthritis Rheum. 1989;32:531.
13. Kushner I. C-reactive protein in rheumatology. Arthritis Rheum. 1991;34(8):1065-68.
14. Husain TM, Kim DH. C-reactive protein and erythrocyte sedimentation rate in orthopaedics. UPOJ. 2002;15:13-6.
15. Paulus H, E., Brahn E. Is erythrocyte sedimentation rate the preferable measure of the acute phase response in rheumatoid arthritis? J Rheumatol. 2004;31(5):838-40.
16. Firestein GS, Budd RC, Harris Jr ED, McInnes IB, Ruddy S, Sergent JS. Kelly's textbook of rheumatology. Philadelphia: Saunders Elsevier; 2009.
17. Wolfe F. Comparative usefulness of C-reactive protein and erythrocyte sedimentation rate in patients with rheumatoid arthritis. J Rheumatol. 1997;24(8):1477-85.
18. Wells G, Becker JC, Teng J, Dougados M, Schiff M, Smolen J, et al. Validation of the 28-joint Disease Activity Score (DAS28) and European League Against Rheumatism response criteria based on C-

reactive protein against disease progression in patients with rheumatoid arthritis, and comparison with the DAS28 based on erythrocyte sedimentation rate. Ann Rheum Dis. 2009;68(8):954-60.

19. Aletaha D, Nell VP, Stamm T, Uffmann M, Pflugbeil S, Machold K, et al. Acute phase reactants add little to composite disease activity indices for rheumatoid arthritis: validation of a clinical activity score. Arthritis Res Ther. 2005;7(4):R796-806.
20. Studenic P, Radner H, Smolen JS, Aletaha D. Discrepancies between patients and physicians in their perceptions of rheumatoid arthritis disease activity. Arthritis Rheum. 2012;64(9):2814-23.
21. Khan NA, Spencer HJ, Abda E, Aggarwal A, Alten R, Ancuta C, et al. Determinants of discordance in patients' and physicians' rating of rheumatoid arthritis disease activity. Arthritis Care Res. 2012;64(2):206-14.
22. Markenson JA, Koenig AS, Feng JY, Chaudhari S, Zack DJ, Collier D, et al. Comparison of physician and patient global assessments over time in patients with rheumatoid arthritis: a retrospective analysis from the RADIUS cohort. J Clin Rheumatol. 2013;19(6):317-23.
23. Van der Heijde DMFM, van ’t Hof MA, van Riel PLCM, Theunisse LAM, Lubberts EW, van Leeuwen MA, et al. Judging disease activity in clinical practice in rheumatoid arthritis: First step in the development of a disease activity score. Annals of the Rheumatic Diseases. 1990;49:916-20.
24. Vermeer M, Kuper HH, van der Bijl AE, Baan H, Posthumus MD, Brus HLM, et al. The provisional ACR/EULAR definition of remission in RA: a comment on the patient global assessment criterion. Rheumatology (Oxford). 2012;51(6):1076-80.
25. Kievit W, Welsing PMJ, Adang EMM, Eijsbouts AM, Krabbe PFM, van Riel PLCM. Comment on the use of self-reporting instruments to assess patients with rheumatoid arthritis: the longitudinal association between the DAS28 and the VAS general health. Arthritis Rheum. 2006;55(5):745-50.
26. van Tuyl LH, Boers M. Patient's global assessment of disease activity: what are we measuring? Arthritis Rheum. 2012;64(9):2811-3.
27. van Tuyl LH, Hewlett S, Sadlonova M, Davis B, Flurey C, Hoogland W, et al. The patient perspective on remission in rheumatoid arthritis: 'You've got limits, but you're back to being you again'. Ann Rheum Dis. In press.
28. Masri KR, Shaver TS, Shahouri SH, Wang S, Anderson JD, Busch RE, et al. Validity and reliability problems with patient global as a component of the ACR/EULAR remission criteria as used in clinical practice. J Rheumatol. 2012;39(6):1139-45.
29. Kirwan JR, Minnock P, Adebajo A, Bresnihan B, Choy E, de Wit M, et al. Patient perspective: fatigue as a recommended patient centered outcome measure in rheumatoid arthritis. J Rheumatol. 2007;34(5):1174-7.
30. Wolfe F, Hawley DJ, Wilson K. The prevalence and meaning of fatigue in rheumatic disease. J Rheumatol. 1996;23(8):1407-17.
31. Bingham III CO, Alten R, Bartlett SJ, Bykerk VP, Brooks PM, Choy E, et al. Identifying preliminary domains to detect and measure rheumatoid arthritis flares: report of the OMERACT 10 RA Flare Workshop. J Rheumatol. 2011;38(8):1751-8.
32. Berthelot JM, De Bandt M, Morel J, Benatig F, Constantin A, Gaudin P, et al. A tool to identify recent or present rheumatoid arthritis flare from both patient and physician perspectives: the 'FLARE' instrument. Ann Rheum Dis. 2012;71(7):1110-6.
33. Lassere MN, van der Heijde D, Johnson KR, Boers M, Edmonds J. Reliability of measures of disease activity and disease damage in rheumatoid arthritis: implications for smallest detectable difference,

minimal clinically important difference, and analysis of treatment effects in randomized controlled trials. J Rheumatol. 2001;28(4):892-903.

34. Landewé R, van der Heijde D, van der Linden S, Boers M. Twenty-eight-joint counts invalidate the DAS28 remission definition owing to the omission of the lower extremity joints: A comparison with the original DAS remission. Ann Rheum Dis. 2006;65:637-41.
35. Mäkinen H, Kautiainen H, Hannonen P, Sokka T. Is DAS28 an appropriate tool to assess remission in rheumatoid arthritis? Ann Rheum Dis. 2005;64:1410-3.
36. van der Leeden M, Steultjens MP, van Schaardenburg D, Dekker J. Forefoot disease activity in rheumatoid arthritis patients in remission: results of a cohort study. Arthritis Res Ther. 2010;12(1):R3.
37. Crowson CS, Rahman MU, Matteson EL. Which measure of inflammation to use? A comparison of erythrocyte sedimentation rate and C-reactive protein measurements from randomized clinical trials of golimumab in rheumatoid arthritis. J Rheumatol. 2009;36(8):1606-10.
38. Wolfe F. The prognosis of rheumatoid arthritis: assessment of disease activity and disease severity in the clinic. Am J Med. 1997;103(6A):12S-8S.
39. Siemons L, ten Klooster PM, Vonkeman HE, Glas CAW, van de Laar MAFJ. Distinct trajectories of disease activity over the first year in early rheumatoid arthritis patients following a treat-to-target strategy. Arthritis Care Res. In press.
40. Coury F, Rossat A, Tebib A, Letroublon MC, Gagnard A, Fantino B, et al. Rheumatoid arthritis and fibromyalgia: a frequent unrelated association complicating disease management. J Rheumatol. 2009;36(1):58-62.
41. Toms J, Soukup T, Bradna P, Hrncir Z. Disease activity composite indices in patients with rheumatoid arthritis and concomitant fibromyalgia. J Rheumatol. 2010;37(2):468.
42. Leeb BF, Andel I, Sautner J, Nothnagl T, Rintelen B. The DAS28 in rheumatoid arthritis and fibromyalgia patients. Rheumatology. 2004;43(12):1504-7.

List of publications

Articles related to this thesis

Siemons L, ten Klooster PM, Taal E, Glas CAW, van de Laar MAFJ. Modern psychometrics applied in rheumatology - A systematic review. BMC Musculoskelet Disord. 2012;13:216.

Siemons L, ten Klooster PM, Taal E, Kuper IH, van Riel PLCM, van de Laar MAFJ, Glas CAW. Validating the 28-tender joint count using item response theory. J Rheumatol. 2011;38(12):2557-64.

Siemons L, ten Klooster PM, Taal E, Kuper IH, van Riel PLCM, Glas CAW, van de Laar MAFJ. The contribution of assessing forefoot joints in early rheumatoid arhtirtis patients: insights from item response theory. Arthrit Care Res. 2013;65(2):212-219.

Siemons L, Vonkeman HE, ten Klooster PM, van Riel PLCM, van de Laar MAFJ. Interchangeability of 28-joint disease activity scores using the erythrocyte sedimentation rate or the C-reactive protein as inflammatory marker. Clin Rheumatol. (in press)

Siemons L, ten Klooster PM, Vonkeman HE, Glas CAW, van de Laar MAFJ. Distinct trajectories of disease activity over the first year in early rheumatoid arthritis patients following a treat-to target strategy. Arthrit Care Res. (in press)

Siemons L, ten Klooster PM, Vonkeman HE, van de Laar MAFJ, Glas CAW. Further optimization of the reliability of the 28-joint Disease Activity Score in patients with early rheumatoid arthritis. PLOS ONE. (in press)

Other articles

Ten Klooster PM, Taal E, Siemons L, Oostveen JC, Harmsen EJ, Tugwell PS, Rader T, Lyddiatt A, van de Laar MA. Translation and validation of the Dutch version of the Effective Consumer Scale (EC-17). Qual Life Res. 2013;22(2):423-429.

ten Klooster PM, Vonkeman HE, Taal E, Siemons L, Hendriks L, de Jong AJL, Dutmer EAJ, van Riel PLCM, van de Laar MAFJ. Performance of the Dutch SF-36 version 2 as a measure of health related quality of life in patients with rheumatoid arthritis. Health Qual Life Outcomes. 2013;11:77.

Siemons L, ten Klooster PM, van de Laar MA, van den Ende CH, Hoogeboom TJ. Validity of summing painful joint sites to assess joint-pain comorbidity in hip or knee osteoarthritis. BMC Musculoskelet Disord. 2013;14:234.

Fries JF, Lingala B, Siemons L, Glas CA, Cella D, Hussain YN, Bruce B, Krishnan E. Extending the floor and ceiling for assessment of physical function. Arthritis Rheum. (in press)

Siemons L, Krishnan E. A short tutorial on item response theory in rheumatology. Clin Exp Rheumatol. (in press)

Summary

Since rheumatoid arthritis (RA) cannot yet be cured, current treatment strategies emphasize an aggressive interference at an early stage of the disease in order to suppress the patient's disease activity as quickly, as completely, and as long as possible. Although these so-called treat-to-target strategies have been shown to be effective in early RA patients in daily clinical practice, they require valid and reliable measurements of disease activity. This thesis addressed a number of concerns have been expressed about the frequently used Disease Activity Score for 28 joints (DAS28) for this purpose. The DAS28 is an index measure that combines a 28-tender joint count, a 28-swollen joint count, a laboratory measure of inflammation (either the erythrocyte sedimentation rate [ESR] or the C-reactive protein [CRP]), and a patient-reported feeling of general health into a single measure of disease activity. It is often used as a criterion variable for treatment decisions or treatment effectiveness. However, several studies have raised questions about the following three points 1) the relevance of residual disease activity in omitted joints, 2) the equivalent use of two very different acute phase reactants, and 3) the inclusion of a patient-reported outcome measure.

1) Because RA is characterized by symmetric inflammations of the peripheral joints, joints are often referred to as the major "organ" involved in RA. Consequently, joint counts are considered essential for assessing RA disease activity. However, because it is often considered unfeasible to assess all joints, reduced 28 joint counts have been proposed, including only the joints of the hands, wrists, elbows, shoulders, and knees. Although we demonstrated that these 28 joints adequately reflect the left-right symmetry of joint involvement that characterizes RA, information about a patient's disease activity was predominantly provided by the smaller joints (i.e. the metacarpophalangeal and proximal interphalangeal joints of the hand). The inclusion of the forefoot joints did not significantly improve the measurement range or measurement precision of the joints counts in patients with early RA. However, the forefoot joints were shown to be frequently affected, so the assessment of omitted joints might be important when monitoring the disease trajectory of individual patients in daily clinical practice.

2) Acute phase reactants are commonly used to quantify the severity of inflammation in RA. Although the ability to calculate DAS28 scores with two different acute phase reactants might seem convenient, and the DAS28-ESR and DAS28-CRP were both shown to be reliable for assessing disease activity in early RA, their scores cannot be used interchangeably in clinical practice. The DAS28-CRP tends to yield lower scores than the DAS28-ESR, resulting in substantial classification differences. In addition, it

was shown that elevated concentrations of the acute phase reactants were not solely due to the inflammation of the rheumatic disease but were also influenced by non-inflammatory external factors like a patient’s age and gender. Hence, these external influences should be taken into account when interpreting the meaning of an elevated ESR and CRP concentration. A modification of the DAS28 formula might be required, for instance by specifying gender and age specific cut-off points of disease activity, or by including gender and age as variables into the model. Abandoning the measurement of the inflammatory markers altogether is not preferable as they were shown to be the most reliable components included in the DAS28.

3) It is believed that clinical and patient-reported outcome measures (PROs) address different aspects of the disease and that both should be administered to evaluate a patient’s disease activity. As such, PROs also belong to the core set measures of disease activity and the DAS28 includes a PRO as well, namely the visual analog scale of general health. However, the scores on this measure are difficult to interpret because general health may reflect multiple dimensions. Health is not merely the absence of disease, but touches upon physical, mental, as well as social aspects of life. As such, a patient’s general health rating can also be influenced by non-inflammatory or personal factors and, consequently, its inclusion has often been criticized. Our analyses appear to support the weakness of this component, as shown by its poor reliability and its low weighting within the DAS28 indices (a weighting that was decreased even further after weight optimization). Therefore, it would be interesting to explore alternative, more reliable, patient-reported outcome measures that could reflect the patient perspective on disease activity within the DAS28.

In conclusion, this thesis confirms that disease activity is a complex, multidimensional concept that should be measured and monitored using both (semi)objective clinical measures as well as patient-reported outcome measures. Although this thesis showed that the DAS28 is a reliable measure, partly justifying its use in both clinical research and clinical practice, it is certainly not flawless and all of its individual components can be criticized to some degree. On a population level the DAS28 score can give a good estimation of disease activity in early RA patients, but in individual RA patients inconsistencies can occur. Scores should always be interpreted within their broader context, including both disease related and non-disease related factors. The DAS28 can guide the treatment process in clinical practice, but a thorough inquiry of the clinical and patient-reported symptoms experienced by individual patients remains important as well. This thesis provides several clues for further improvements in the assessment of disease activity in early RA.

Samenvatting

Aangezien reumatoïde artritis (RA) nog niet kan worden genezen, richten huidige behandelingsstrategieën zich op een agressieve interferentie in een vroeg stadium van de ziekte om de ziekteactiviteit van de patiënt zo snel mogelijk, compleet mogelijk en lang mogelijk te onderdrukken. Deze zogeheten treat-to-target strategieën zijn effectief gebleken bij vroege RA patiënten in de klinische praktijk, maar ze vereisen wel dat ziekteactiviteit valide en betrouwbaar gemeten wordt. Dit proefschrift behandelt een aantal mogelijke beperkingen van de hiervoor veelvuldig gebruikte ziekteactiviteitscore in 28 gewrichten (DAS28). De DAS28 is een index waarin een meting van pijn in 28 gewrichten, een meting van zwelling in 28 gewrichten, een ontstekingswaarde (ofwel de bezinking [BSE] of het C-reactief proteïne [CRP]) en een patiënt-gerapporteerde maat van algeheel welbevinden worden gecombineerd tot één maat van ziekteactiviteit. Het wordt vaak gebruikt als maatstaf voor het bepalen van de behandelingsstrategie of behandelingseffectiviteit. Toch hebben verscheidene studies vragen gesteld bij de volgende drie aspecten: 1) de relevantie van residuele ziekteactiviteit in niet opgenomen gewrichten, 2) het equivalente gebruik van twee zeer verschillende acute fase reactanten en 3) de inclusie van een patiënt-gerapporteerde uitkomstmaat.

1) Doordat RA wordt gekarakteriseerd door symmetrische ontstekingen in de perifere gewrichten, worden gewrichten vaak gezien als het primaire “orgaan” in RA. Het meten van zogeheten *joint counts* wordt dan ook essentieel geacht voor het bepalen van de ziekteactiviteit in RA patiënten. Echter, omdat het meten van alle gewrichten vaak niet haalbaar wordt geacht, zijn verkorte 28 joint counts voorgesteld waarin alleen de gewrichten van de handen, polsen, ellebogen, schouders en knieën worden meegenomen. Hoewel wij hebben aangetoond dat deze 28 gewrichten de voor RA karakteriserende links-rechts symmetrie in aangedane gewrichten goed weergeven, werd informatie over de ziekteactiviteit van de patiënt voornamelijk verkregen uit de kleine gewrichten (i.e. de metacarpofalangeale en proximale interfalangeale gewrichten van de hand). Het includeren van de voorvoetgewrichten resulteerde niet in een verbetering van het meetbereik of de meetprecisie van de joint counts in de vroege RA patiënten. Wel waren de voorvoetgewrichten vaak aangedaan, wat laat zien dat het afnemen van gewrichten die niet in de joint count zijn opgenomen wel degelijk belangrijk kan zijn voor het monitoren van het ziekteproces van individuele patiënten in de klinische praktijk.

2) Acute fase reactanten worden vaak gebruikt om de ernst van de ontsteking in RA te kwantificeren. Hoewel de mogelijkheid om de DAS28 met twee verschillende acute

fase reactanten te berekenen wellicht handig kan zijn, en de DAS28-ESR en DAS28-CRP beide betrouwbaar bleken voor het bepalen van de ziekteactiviteit in vroege RA, werd eveneens duidelijk dat de twee scores niet onderling uitwisselbaar zijn in de klinische praktijk. De DAS28-CRP heeft de neiging om lagere scores te geven dan de DAS28-ESR, wat leidt tot aanzienlijke classificatie verschillen. Bovendien werd aangetoond dat verhoogde concentraties van de acute fase reactanten niet alleen waren toe te wijzen aan de reumatische ontsteking, maar ook aan niet-inflammatoire externe factoren zoals leeftijd of geslacht. Deze niet-inflammatoire factoren zouden moeten worden meegenomen bij het interpreteren van een verhoogde BSE of CRP concentratie. Wellicht is er een modificatie van de DAS28 score nodig, bijvoorbeeld door middel van het specificeren van geslachts- en leeftijdsspecifieke afkapwaarden van ziekteactiviteit of door geslacht en leeftijd als variabelen op te nemen in het model. Het geheel afstand nemen van ontstekingswaarden in de DAS28 geniet niet de voorkeur, omdat zij tot de meest betrouwbare componenten van de DAS28 behoren.

3) Er wordt verondersteld dat klinische en patiënt gerapporteerde uitkomstmaten (PROs) verschillende aspecten van de ziekte belichten en dat beide zouden moeten worden afgenomen om de ziekteactiviteit van een patiënt te evalueren. Als zodanig behoren PROs dan ook tot de kernmaten van ziekteactiviteit en de DAS28 bevat eveneens een PRO, namelijk de visuele analoge schaal van algeheel welbevinden. De scores op deze schaal zijn echter moeilijk te interpreteren, omdat algeheel welbevinden verschillende dimensies kan weerspiegelen. Gezondheid is niet alleen de afwezigheid van ziekte, maar heeft raakvlakken met de fysieke, mentale, als ook sociale aspecten van het leven. Als zodanig kan de beoordeling van algeheel welbevinden van een patiënt ook worden beïnvloed door niet-inflammatoire of persoonsgebonden factoren en derhalve is de opname deze PRO in de DAS28 vaak bekritiseerd. Onze analyses lijken de zwakheid van deze component te staven, gezien zijn lage betrouwbaarheid en lage weging binnen de DAS28 index (een weging die zelfs nog lager werd na optimalisatie). Het zou daarom interessant zijn om naar alternatieve, betrouwbaardere, patiënt-gerapporteerde uitkomstmaten te zoeken om het patiënten perspectief op ziekteactiviteit in de DAS28 mee te nemen.

Samenvattend bevestigt dit proefschrift dat ziekteactiviteit een complex, multi-dimensioneel concept is dat gemeten en gemonitord zou moeten worden met behulp van zowel (semi-)objectieve klinische maten als patiënt-gerapporteerde uitkomstmaten. Hoewel dit proefschrift laat zien dat de DAS28 een betrouwbare maat is, wat het gebruik ervan in klinisch onderzoek en in de klinische praktijk deels rechtvaardigt, is het zeker niet zonder beperkingen en kunnen alle individuele componenten tot op zekere hoogte worden bekritiseerd. Op populatieniveau kan de DAS28 score een goede schatting geven van de ziekteactiviteit in vroege RA patiënten, maar bij individuele patiënten met RA kunnen inconsistenties optreden. Scores dienen altijd te worden geïnterpreteerd binnen hun bredere context, waarbij zowel ziekte gerelateerde en andere invloedrijke factoren worden genomen. De DAS28 kan als leidraad dienen binnen het behandelingstraject in de klinische praktijk, maar een grondig onderzoek van overige klinische en patiënt-rapporteerde symptomen die door de patiënt worden ervaren blijft eveneens belangrijk. Dit proefschrift biedt verscheidene aanknopingspunten voor verdere verbeteringen in het beoordelen van ziekteactiviteit in patiënten met vroege RA .