Summary and general discussion

The development of outcome measures in the field of rheumatology has shown a rapid and constructive advance in the last decades. It has been achieved to obtain more uniformity in the evaluation of inflammatory rheumatic disorders, especially concerning long-term outcomes and disease activity assessments. In the field of RA, clinical experts and researchers have done a lot of work under the umbrella of OMERACT (Outcome Measures in Rheumatology Clinical Trials) [1]. In the field of AS, major advances in the outcome and therapeutic advances in the last years have been possible because of projects piloted within ASAS (the Assessment of Spondyloarthritis international Society).

The analyses and studies presented in this thesis were part of this process, with most of the work performed in international working groups under auspices of either OMERACT or ASAS or both.

ASSESSING RADIOGRAPHIC PROGRESSION IN RA IN CLINICAL PRACTICE

In **chapter 1**, we have assessed the metrological properties of a scoring method for radiographic damage in RA, the SENS method, which aimed at making the objective and numerical evaluation of structural damage in a given patient feasible for clinical rheumatologists without specific training [2]. Only the recognition of an erosive lesion or of any joint space narrowing in the joints of interest is required, as no further assessment of the degree of destruction is necessary to complete the scoring: While comprehensive scoring methods, in particular the Sharp and Larsen methods and their modifications, attribute a numerical value to each of the scored joint according to the level of damage as appreciated by the reader, SENS only acknowledges the presence (1 point) or absence (zero point) of erosions and joint space narrowing separately [3]. Baseline- and 52 weeks radiographies of 680 RA patients included in the TEMPO trial (Trial of Etanercept and Methotrexate with Radiographic Patient Outcome) were scored according to the Sharp-van der Heijde (SHS) method by 2 independent readers, and SENS was derived from their results [4, 5]. Reliability of status scores was found high for both methods, with intraclass correlation coefficients (ICC) for SHS of 0.81 and 0.77 at baseline and 52 weeks and 0.91 and 0.89 for SENS respectively. ICCs of change scores were somewhat lower. Both scores performed equally well with regard to sensitivity to change, and the optimal cut-off levels for progression vs. no progression based on smallest detectable change were, as determined by ROC-curve analysis, at 2 units for SHS and 1 unit for SENS. With regard to their discriminatory ability, both

methods also performed very well. Additionally, it was suggested by further analysis that the SENS does not suffer a lot from ceiling effect. In conclusion it was proven that SENS has a good reliability, sensitivity to change and discriminatory ability, and can be considered for use in clinical practice also in patients with established disease.

AUTOMATED MEASUREMENT OF JOINT SPACE WIDTH IN RHEUMATOID ARTHRITIS

In **chapter 2**, five (semi-)automated methods to measure joint space width and its change over time in joints of patients with RA have been compared. One set of radiographs from 107 patients included in the COBRA (*Combinatietherapie Bij Reumatoide Artritis*) trial was used, and different aspects of the 5 methods, such as feasibility, efficiency, reliability and discriminatory ability were tested [6]. Most methods showed efficiency problems: With a 50% success rate (i.e. at least half of the evaluated joints successfully measured both at baseline- and follow-up visits) as a benchmark, 4 of the 5 methods passed the efficiency test. But when higher requirements were asked, efficiency of most methods fell short: Only 3 methods passed the requirement of at least 75% of joints successfully measured. Repeatability, on the other hand, as assessed by ICC, was very acceptable (ICC>0.80). The interpretation of discriminatory ability of the different methods was more challenging, and direct comparison across methods is even unreasonable, because different joints are assessed by the various systems, missing (i.e. unsuccessfully assessed) joints differed across methods. The same films scored by the Sharp van der Heijde method served as external standard. The manual scoring of the joint space narrowing score did not accurately distinguish the treatment arms in the COBRA trial, while the difference was highly significant for the erosion score. However, some of the (semi)automated measurements were able to pick up a difference between the treatment arms, even with the limited number of joints included and the limited success rate. It was concluded that (semi-)automated measurement of joint space width in RA may help in better discriminating between treatment arms in the context of a clinical trial, provided that their efficiency is improved and that they are combined with the assessment of radiographic erosions.

RADIOGRAPHIC PROGRESSION IN RHEUMATOID ARTHRITIS, PLAUSIBILITY OF REPAIR

The research described in **Chapters 3 and 4** was aiming at getting a better and more detailed insight in radiographic progression in clinical trials of RA. In particular, the occurrence of so-called negative change scores, which suggest improvement, was investigated. An important finding was that negative scores were not only due to reading error but partly to true improvement (repair). Somewhat to our surprise we found that multiple and independent scoring of the same radiographs under blinded time sequence led to seemingly different results on a per joint basis: From the 7255 individual joints with 4 change scores available, pertaining to 178 patients in TEMPO trial, only a minority (1.3 to 5.8% across reads) showed change over 1 year. Most surprisingly, the absolute agreement across readers about the direction of change (either positive, i.e. worsening of erosive damage, or negative, i.e. apparent improvement of lesions) was very low: In only 12 joints a consistently positive or negative change pattern was found in all 4 readings. Discrepant change patterns, on the other hand, were also very rare. We explained the paradox of reliable patient total change scores made up by unreliable single joint scores by introducing the concept of conservative vs. sensitive readers: In this concept, the direction of the score is always determined by the sensitive reader. The conservative reader scores zero change in case of doubt and does change the magnitude of the signal but not the direction. These subtleties underscore the importance of perfect blinding of time order in readings of RCTs in RA.

Further exploration of radiographic results at the individual joint level in **chapter 4** showed that the negative change scores can be regarded as a surrogate of repair, namely as a decrease in the size of erosive lesions. The finding that a negative change score was statistically more likely to occur in a joint of a patient treated with a TNF blocker, if that joint also showed improvement or resolution of clinical swelling, adds to the belief that repair is a real event and not just an assessment artifact.

THERAPEUTIC APPROACH OF EARLY INFLAMMATORY ARTHRITIS: BEHAVIORS AND CONSEQUENCES

Chapter 5 reported and analyzed the therapeutic approach of French rheumatologists facing a patient with early inflammatory arthritis. It was demonstrated that both the time-to-treatment-start and the choice of the drug were hetero-

geneous across geographical regions. The tendency of initiating a DMARD in a patient with high disease activity, abnormal acute phase reactants, polyarthritis (i.e. more than 3 joint groups) and the presence of anti-CCP antibodies, however, was paramount. Interestingly, a significant interaction between the study center and the result of the anti-CCP test on one hand, and swollen joint count on the other hand, pointed to differences in the interpretation of clinical data in light of the therapeutic decision: Some rheumatologists find the presence of bad prognostic markers less important than others.

In **chapter 6**, the therapeutic behavior of rheumatologists in early inflammatory arthritis and its consequences with regard to radiographic progression were further investigated. Because the ESPOIR cohort is observational and does not dictate a treatment protocol, time to initiate a DMARD and choice of the drug were left to the discretion of the treating rheumatologist. Consequently, as was previously described in chapter 5, the time elapsed from the arthritis onset to the start of a DMARD was variable across patients, and could thus be modeled as a factor of its efficacy [7, 8]. The propensity analysis approach was chosen in these circumstances, because in observational studies, the "treatment"-groups (here early treatment versus delayed treatment) by default show imbalances on prognostic factors [9]. These prognostic imbalances may confound the relation between the chosen DMARD treatments and the outcome (radiographic progression): It is difficult to attribute differences in responses to the treatment itself because the patient- and disease characteristics are also believed to influence the response. The propensity score method aims to reduce the confounding effects of such covariates, and allows differences of responses to be attributed to differences in therapeutic strategies (early versus delayed). In the case of the ESPOIR study, we were able to demonstrate that in daily clinical practice the very early initiation of a DMARD, which means within the first 3 months after the first occurrence of joint swelling, had a significant impact on the radiographic progression after 12 months of follow up. This demonstration adds to the credibility of the current recommendations regarding the management of patients with early arthritis, as well as to the appreciation that both a fast referral to a rheumatologist and an early treatment start are required to improve the long term prognosis of the disease [10-13].

MAGNETIC RESONANCE IMAGING IN AS: OUTCOME MEASURES

Chapter 7 was a multi-reader experiment to compare feasibility, discriminatory ability, responsiveness and reliability of 3 different methods developed to measure inflammation on MRI in the spine of AS patients. The Ankylosing Spondylitis spine MRI score for activity (ASspiMRI-a), the Berlin method (a modification of the ASspiMRI-a), and the Spondyloarthritis Research Consortium of Canada Magnetic Resonance Imaging Index for Assessment of Spinal Inflammation in AS (SPARCC) were compared using a set of MRIs from 30 patients participating in a clinical trial that compared a TNF-blocking drug with placebo [14-16]. The MRI sets were presented in randomized order to 9 readers who scored them by all three methods. Repeatability and inter-reader reliability for every reader pair were highly heterogeneous across methods, although consistently higher ICCs (relative agreement) were found for the SPARCC method. Absolute agreement, however, was worst for the SPARCC method. This apparent contradiction can be explained by the fact that the scoring methods are based on different possible score ranges [17]. With regard to sensitivity to change and discrimination between active drug and placebo, all methods were appropriate. In conclusion, a definite preference for one of the 3 methods could not be made. While all methods would probably show similar performances in the context of a clinical trial, where sensitivity to change and discriminatory ability are of major interest, assessment of MRIs of AS patients included and followed up in a longitudinal cohort would make the SPARCC method less useful, because only 6 discovertebral units are taken into account (the most severely affected).

ASSESSING DISEASE ACTIVITY IN AS: THE ASDAS

Chapter 8 describes the results of an analytical process to derive a composite score with optimal metrological performance in assessing disease activity in AS patients, the ASDAS (Ankylosing Spondylitis Disease Activity Score). This work resulted in 4 draft indices that combine patient-reported assessments of disease activity and acute phase reactants. All these 4 indexes showed high discriminatory ability, both in the cohort they were derived from (ISASS: *International Study on Starting tumour necrosis factor-blocking agents in Ankylosing Spondylitis*) and in an independent cohort that was chosen: the OASIS (*Outcome in Ankylosing Spondylitis International Study*) cohort. They showed a consistently good performance in distinguishing patients with high versus low

disease activity, regardless of the external reference that was used. Most importantly, this discriminatory ability was systematically higher in comparison with the current reference measure, the BASDAI (Bath Spondylitis Disease Activity Index), which is often criticized because it is a strictly patient-reported outcome measure [18]. Follow-up research, conducted by others, has since advanced the concept: The ASDAS has further been validated in relevant subgroups, confirming a stable performance in patients with AS with peripheral arthritis or in patients with normal levels of acute phase reactants [19]. Definition of disease activity levels have also been derived, as well as cutoff values to determine whether a change in ASDAS observed after therapeutic adjustment can be regarded as clinically relevant: To separate inactive disease, moderate, high and very high disease activity, cutoff values were derived: 1.3, 2.1 and 3.5 units. A change ≥1.1 unit was considered clinically important improvement and a change ≥2.0 units a major improvement [20].