The handle http://hdl.handle.net/1887/36591 holds various files of this Leiden University dissertation.

Author: Navarro Compan, Maria Victoria
Title: Aspects of methodology in assessing inflammation and damage in rheumatoid arthritis and axial spondyloarthritis
Issue Date: 2015-12-03
Summary and Conclusions
SUMMARY AND CONCLUSIONS

The studies of this thesis cover outstanding aspects of research methodology in the assessment of inflammation and damage in patients with rheumatoid arthritis (RA) and axial spondyloarthritis (axSpA). The studies pertaining to part I focus on RA and may help to better understand the relationship between disease activity, radiographic damage and disability in patients with RA. More specifically, they evaluate which disease activity measure is best associated with radiographic damage and which of the structural lesions detectable on radiographs is most associated with disability. In addition, these studies also address methodological issues related to the optimal assessment of radiographic progression in clinical trials. In part II of this thesis, the studies have axSpA as the main topic. In detail, they evaluate the yield of tests reflecting inflammation in general. The purpose of these tests is to help classifying patients or to monitor disease activity. Among others, the studies described in this thesis provide data for a better usage of these tests in clinical practice.

In this chapter, we present a summary of the main findings of these studies. Thereafter, we will come to a synthesis of these findings with a focus on principles of assessment and their relevance to determining outcome in chronic inflammatory rheumatic diseases. Simultaneously, we will shape potential future research questions in the field of assessing outcome.

RHEUMATOID ARTHRITIS

Relationship between disease activity measures and radiographic damage

At the start of the studies described in this thesis, disease activity indices (DAIs) were considered better than single variable instruments, including patient-reported outcomes (PROs), for monitoring disease activity in patients with RA. This prioritization was largely based on expert opinion but was supported by limited data only. We have performed a systematic literature review to explore the existing knowledge about the relationship between the DAIs and their individual components on the one hand, and radiographic progression on the other hand. We considered this type of information essential to decide which tool should be employed to monitor disease activity (chapter 2). In total, 57 studies were included in this review and it was shown that all DAIs that include a joint count are related to radiographic progression. In addition, among the single variable instruments, only measures reflecting inflammation, such as swollen joint count (SJC) and erythrocyte sedimentation rate (ESR), were related to radiographic progression. Importantly, PROs did not show such an association. Accordingly, we have therefore recommended the use of one of the DAIs including a SJC to monitor disease activity in patients with RA in clinical practice.
Relationship between types of radiographic damage and disability

Radiographic damage assessed in patients with RA usually combines three types of radiological lesions reflecting different kinds of structural damage: ‘true’ joint space narrowing (JSN), erosion and subluxation or complete luxation -(sub)luxation- \(^3\). In previous studies the level of radiographic damage was shown to be associated with disability in patients with RA, but the contribution of each subtype of lesions to several aspects of disability was not known \(^4\)\(^5\).

In the study described in chapter 3 we have performed a longitudinal analysis on the 10-year follow-up data from patients included in the Norwegian arm of the European Research on Incapacitating Diseases and Social Support (EURIDISS) cohort \(^6\). The topic of interest was the relationship between each subtype of lesion and disability. All 3 subtypes observed on radiographs (true JSN, erosion and (sub)luxation) were related to grip strength but only JSN, especially in the wrist, appeared to be independently associated: we found that an increment of 11 true JSN units in both hands will on average lead to a decrease of 1 kg in mean grip strength of both sides over a period of 10 years. Most reported studies have established a clear relationship between absolute radiographic damage (i.e. the sum of the three type of lesions scores) and overall disability, measured by the health assessment questionnaire (HAQ)-score. In our study, none of the three subtypes of lesions was independently associated with HAQ-score. Based on this, we have concluded that true JSN in the hands may contribute more to explaining variation in hand function than erosion and (sub)luxation do, while at the patient level all 3 types of radiographic damage in an aggregated fashion (as total score) contribute to explaining variation in overall disability.

Methodological aspects of assessment of radiographic damage

Because of the association with (changes in) disease activity and functional disability, inhibition (or slowing) of radiographic progression has been established by regulatory authorities as one of the claims that could be granted for new treatments in RA \(^7\). In order to appropriately investigate such a claim, a number of typical methodological issues were outstanding.

Adjudication

The first issue pertains to ‘adjudication’. Usually, in clinical trials two readers independently provide scores to sets of images that they have to judge with unknown chronology. If an important discrepancy between the scores of these two readers occurs, a third reader (named the adjudicator) is asked to provide an adjudication score that, together with the closest score of the two initial readers, is used to obtain the ‘mean change score’. This process, which serves to constrain measurement error in the trial, is methodologically legitimate but requires a certain threshold for the difference between the two primary readers before adjudication is started. This threshold is arbitrary, but regulatory authorities have expressed concerns if -for a given clinical trial- 20% or more of the final mean change scores of patients
are resolved by adjudication. Often, using the Sharp-van der Heijde (SHS) method, differences between reader-scores of 7-15 units were operationalized as adjudication threshold, but this choice was completely arbitrary. Being aware of this methodological issue, we aimed to provide data regarding the proportions of patients to be adjudicated given a predetermined threshold for the difference in change score between two initial readers in RA trials (chapter 4). We have analysed datasets of 15 recent randomised controlled trials (RCTs) with 2-4 time points per trial in which radiographic progression had been assessed by 13 readers acting in pairs. As expected, the adjudication rate was inversely related to the threshold for the difference between the two readers. But the rate was rather low (always below 22%) even if very conservative thresholds of ≥3 units were applied. In addition, we found that particular features could influence the adjudication rates: the adjudication rate increased by increasing number of time points per trial, by a longer time gap between visits and by shorter disease duration.

Smallest detectable change

The second unresolved methodological issue pertains to the question how to decide if radiographic progression in an individual patient who has participated in a clinical trial is ‘true’ (beyond measurement error) or not. There is consensus about the preferred metric here: the smallest detectable change (SDC) \(^8,9\). But there is no consensus about how to determine this SDC if RCTs with more than two time points are analysed. At the beginning of this study, two analytical methods for estimating an SDC were available: One ‘simple’ method based on Bland & Altman (B&A) analysis; and a more complex method based on generalizability theory and involving analysis of variance (ANOVA) \(^9\). The simple B&A-method suffices for two time points trials, but for complex datasets with several time points such as in most recent RCTs only the method based on generalizability theory is available. In chapter 5, using same data employed in the analysis of the study in chapter 4, we propose a simple extension of the B&A-method: we have investigated if the mean of all interval-SDCs obtained by the simple B&A method is an appropriate surrogate for the ‘ANOVA-method’ for estimating the overall SDC of radiological progression in complex databases. For this purpose, we have evaluated the agreement between the two different methods. The mean (standard deviation) difference observed between the two methods was only \(-0.13 (0.28)\), range \((-0.48, 0.25)\) units. If we consider the minimal clinically important difference for radiographic progression (3 units of SHS method) \(^10\), this difference of 0.13 units is negligible. Accordingly, we propose to report the average of all interval SDCs as an appropriate surrogate for the ANOVA-based SDC in complex databases.

We have addressed another methodological issue of B&A-based SDC calculations: The B&A method purports to determine an upper and lower level of agreement (LoA), which are boundaries enclosing an area within which it cannot be precluded that a measured change (here: progression score) in an individual patient is due to measurement error rather than to true change. The upper and lower LoAs are statistically derived and enclose 95% of all
observed data. The choice for 95% threshold is arbitrary and not well justified in literature. The 95% LoA resemble 95%-confidence intervals but are fundamentally different. We thought that, because of this fundamental difference, 95% LoA are unnecessarily conservative in the context of radiographic progression scores, where they importantly limit the sensitivity to detect subtle changes. We have compared SDCs based on the 95% LoA with the SDCs based on the 80% LoA (chapter 4). The mean (standard deviation) SDC was 3.0 (0.7) when based on 95% LoA and 2.0 (0.4) when based on 80% LoA. If we choose 80% LoA, more patients would classify as ‘true progressors’ in both groups, which is of (modest) statistical benefit (more statistical power). But there is a more clear advantage related to this: a more liberal SDC will lead to more patients being classified as ‘progressor’. The aim of an RCT is to investigate if a new treatment reduces the number of ‘progressors’. In order to reliably demonstrate this, it is important that there is a sufficient proportion of patients with ‘progression beyond measurement error’ in the control group, and a 80% LoA SDC helps achieving this without unacceptably jeopardizing trial-results.

AXIAL SPONDYLOARTHRITIS

Issues of diagnosis

HLA-B27 and MRI-SI

Unlike many other diseases in rheumatology, axSpA is a disease in which supplementary tests and procedures play an important role to make a diagnosis or to measure disease activity. These supplementary procedures include relatively simple laboratory tests such as tests to quantify the level of acute phase reactants (APRs) and a blood test to determine human leukocyte antigen B27 (HLA-B27) carriernship, but also relatively complicated imaging procedures such as magnetic resonance imaging of the sacroiliac joints (MRI-SI) that quantifies bone marrow edema. In part II of this thesis, several aspects of these supplementary procedures for different purposes were further evaluated.

First, cost-, availability- and feasibility issues make that HLA-B27 testing and especially MRI-SI cannot be performed in all patients presenting with chronic back pain (CBP), as this is a very prevalent symptom in clinical practice and still the most common presenting symptom in patients with axSpA.

In chapter 6, we have addressed the principles of sequential testing by investigating if particular SpA features point to a higher likelihood of a subsequent positive MRI-SI or a positive HLA-B27 test in patients with CBP referred to rheumatologists participating in the ESPeranza Programme. Such SpA features, that can be obtained by history taking or simple physical examination, could potentially contribute to an efficient ‘test-sequence’ to
be applied in patients presenting with CBP. The prevalence of a positive MRI-SI or a positive HLA-B27 test in patients in ESPeranza was 41% and 40%, respectively. Unfortunately, we did not find any of the SpA features to increase the likelihood of a subsequent positive HLA-B27 test. But interestingly, we found that the presence of the symptom ‘awakening at second half of night’ together with inflammatory back pain (IBP) according to the Calin definition or the presence of the rather uncommon symptom ‘alternating buttock pain’ together with IBP according to the Assessment of SpondyloArthritis international Society (ASAS)/Calin criteria increased the likelihood on a subsequent positive MRI-SI from 40 to 79-80%. In addition, we found that the findings ‘dactylitis’ and ‘presence of inflammatory bowel disease’ increased this likelihood from 41 to 73% and 81%, respectively. Based on these results, and in case of limited resources, the presence of any of these four characteristics may be valuable in helping rheumatologists to improve the efficiency of ordering MRI-SI in patients with suspected axSpA.

*High sensitivity CRP*

Concerning how to diagnose patients with early axSpA, the new classification ASAS criteria developed in 2009 have shown to outperform the older criteria. ‘Elevated C-reactive protein –CRP.’ is one of the SpA features in the ASAS criteria. At the time of development of these criteria, CRP was measured with the conventional CRP-methods. Recently, high sensitivity CRP (hsCRP)-methods, which are more sensitive than conventional CRP-methods, have been developed and have replaced the conventional CRP methods in many hospitals. However, hsCRP-methods have not been validated for application in the ASAS criteria. We have evaluated in patients included in the Devenir des Spondylarthopathies Indifférenciées Récentes (DESIR) cohort to what extent the replacement of the conventional-CRP test by the hsCRP test would (maybe inappropriately) increase the sensitivity of the ASAS criteria for classifying patients with axSpA (*chapter 7*). In the subgroup of DESIR-patients with normal conventional CRP, we have observed higher hsCRP mean serum levels in the patients with axSpA as compared to patients without axSpA. But importantly, when elevated conventional CRP was substituted by elevated hsCRP, not one single patient could additionally be classified as axSpA. This means that –for the purpose of classifying patients with axSpA— conventional CRP-methods can be replaced by hsCRP-methods without further consideration.

*Issues of monitoring disease activity*

*High sensitivity CRP*

Changing assessment-methods, such as the methods to measure CRP, may not only affect the performance of criteria for axSpA but also the instruments to measure disease activity of axSpA. The Ankylosing Spondylitis Disease Activity Score (ASDAS), an index that was recently developed, integrates several PROs as well as an APR reflecting inflammation. At the development of the ASDAS, only conventional CRP (but not hsCRP) was used. Arbitrarily,
and awaiting further data driven guidance, it was suggested to use 50% of the threshold value of CRP in the ASDAS formula if conventional CRP was below the detection limit. This suggestion was not supported by data. In the study described in chapter 8, we have first selected patients from DESIR with a conventional CRP below the limit of detection, then have calculated the ASDAS using hsCRP values and finally compared these ASDAS values with the values of ASDAS obtained by imputing CRP with 11 different artificial threshold-values (range 0-5 mg/L, at 0.5 mg/L intervals). The best agreement with truly measured hsCRP (the external standard) was found for imputed values of 1.5 and 2 mg/L, with only a minimal discrepancy between them. Based on the performance in clinically relevant disease activity states a value of 2 mg/L was finally proposed as the best option for imputation: if the conventionally measured CRP level yields a result reported as ‘below the limit of detection’ or if the hsCRP is reported as ‘below 2 mg/L’, a value of 2 mg/L should be imputed for CRP in the ASDAS formula. In addition, we also took the opportunity to evaluate if the ASDAS employing ESR (instead of CRP) was in agreement with the ASDAS using hsCRP. Our findings confirmed that the agreement between the two formulae is good.

**MRI-SI**

While the value of MRI-SI in diagnosing patients with axSpA is undisputed, the added value of MRI-SI in monitoring disease activity in patients with axSpA was (and still is) a matter of debate. In this regard, first it was important to investigate if signs and symptoms, quantified by simple clinical or laboratory measures, are longitudinally associated with the presence of inflammatory lesions detected by MRI-SI. Data regarding this topic were scarce and had mainly been obtained from studies including patients with established disease and employing only cross-sectional analysis \(^{19,20}\). In chapter 9, we have investigated and observed for the first time the existence of a longitudinal relationship between clinical disease activity measures and inflammatory lesions detected by MRI-SI, assessed according to the Spondyloarthritis Research Consortium of Canada (SPARCC) score \(^{14}\). Among all measures, the ASDAS was best associated to MRI-SI: on average, an increase of one unit in ASDAS is associated with an increase of 2.8 units in SPARCC score. Interestingly, we also have shown for the first time that the relationship between clinical disease activity measures and MRI-SI inflammatory lesions is different in males and females: there was a marked association between ASDAS and MRI-SI in males, while this association was absent in females. Translating these data into clinical terms, we suggest that monitoring the ASDAS in male patients with axSpA provides useful ‘subjective’ information as well as ‘objective information’ that is congruent with the pathophysiological hallmark of the disease, being this inflammation of the sacroiliac joints (on MRI). In females with axSpA the situation is far less clear: the lack of association between the ASDAS and MRI-SI in females points again to the observation that axSpA has a different expression in males than females (see below).
The central theme of this thesis is **assessment**. Reading the chapters of this thesis it becomes clear that **assessment** influences many aspects of outcome in inflammatory rheumatic diseases. The basic construct underlying the research in this thesis is the well-known appreciation that inflammation in a chronic inflammatory rheumatic disease such as RA or axSpA leads to some kind of structural damage, which in turn contributes to functional impairment, which can be considered a long-term outcome.

It is important to realize that all these levels can and should be **assessed** so that diseases can better be diagnosed, prognostic factors for outcome can better be identified, pivotal associations can be constructed, long-term functional impairment can better be explained, interventions can be designed to improve long-term outcome, etc: better **assessment** will lead to a better understanding and a better management of our chronic inflammatory diseases.

The importance of **assessment** in understanding the different levels of the here defined outcome pyramid will be briefly discussed below one-by-one:

1. **Assessment** starts at diagnosis. Inflammatory rheumatic diseases are typically diagnoses by ‘pattern recognition’: the diagnostician combines his knowledge about the pattern of a disease (the ‘Gestalt’) with information obtained from patients presenting to him with complaints and symptoms. Rheumatology is a medical discipline in which the pattern of the disease is spelled out by classification criteria, not by an unambiguous external standard! These criteria are the product of consensual deliberations among experts (including patients) who have integrated the best of their knowledge into classification algorithms, followed by validation studies. RA and axSpA are examples of diseases in which the pattern is defined by classification criteria. Many of these criteria should be quantified, **assessed**. This feature makes criteria susceptible to technical and methodological developments over time. An often overlooked consequence of this is that every change that takes place in the **assessment** of criteria must be validated to some extent in order to be able to oversee the consequences of such a change. The replacement of conventional CRP tests by hsCRP tests, as highlighted in this thesis, seems a logical move, but is in fact an example of a change with potentially bad consequences for the classification of axSpA. But obviously issues like this do not pertain solely to laboratory tests: MRI-SI is another example of a procedure with relevance for a diagnosis or a classification. Changing the content of a positive MRI-result, or the threshold, would imply a change in the **assessment** of the disease that may have important consequences for classification and diagnosis.

In this thesis we have also addressed a slightly different aspect of diagnostic **assessment**: we have explored sequential testing in axSpA; we have investigated if simple clinical
tests may increase the likelihood of a certain result obtained by another diagnostic test (MRI-SI or HLA-B27). Sequential testing is a technique that is implicitly applied by experienced clinicians who make a diagnosis. Essentially, sequential testing means that the ordering of a second (often more costly or incriminating) test is made on the basis of the result of the first (often simple) test. Sequential testing is clinical reasoning, which is the opposite of ‘checkbox medicine’ that implies ticking checkboxes in a set of classification criteria. Sequential testing is also congruent with Bayesian reasoning in that ‘prior knowledge’ is taken into consideration when deciding if additional testing in an individual patient should be carried out. We have given an example of a study here that makes transparent that sequential testing (in axSpA) may increase the yield of subsequent more delicate diagnostic procedures (e.g. MRI-SI). It is our conviction that Bayesian reasoning in Rheumatology will lead to more accurate diagnoses, is less expensive, and more satisfactory to patient and physician, than ‘cookbook medicine’ by protocol based on sets of classification criteria that nowadays and unfortunately is propagated in many countries as a means to control health care costs.

2. The next levels of our outcome construct pertain to the assessment of disease activity and structural damage. While we do not dispute anymore that in inflammatory rheumatic diseases inflammation (operationalized as disease activity) leads to damage, we may ignore too often that the strength of this relationship is subjected to methodological principles of assessment underlying disease activity and damage. In other words: if we change the way in which we measure disease activity or damage, or change the components of a disease activity measure or a damage measure, we may also change the strength of the association, and consequently the impact that this association may have on our perception of the disease. In this thesis we have investigated several aspects of assessment of disease activity and damage. Not only have we substantiated once again that the association between disease activity and damage is best served by choosing DAIls rather than by single components, but we have also worked out ‘open ends’ in the assessment of structural damage by addressing topics like ‘adjudication’ and ‘smallest detectable change’. At first glance, these topics may seem trivial, until one realizes that—in the context of clinical trials— adjudication influences the precision of a scoring result, and the determination of the SDC directly influences the number of patients with progression of damage in a trial. In summary, these seemingly trivial aspects of assessment may each have its influence on the strength of the association between disease activity and structural damage, and are therefore relevant to mention. Along similar lines, one may argue that relatively subtle changes in the method to assess disease activity, such as the replacement of conventional CRP by hsCRP in the formula of the ASDAS may or may not have repercussions on the performance of the ASDAS, and consequently on the association of ASDAS and structural damage in axSpA.

3. The last level of our outcome pyramid pertains to disability. Physical function or the impairment thereof, has many components in itself: in RA impaired grip-strength implies impairment of the hand-function, but an increased HAQ-score far more refers to general
disability. Needless to say that if one aims at investigating the association between damage and physical function in RA, it is crucial to define how to assess damage and how to assess disability. Far too often, these associations have been investigated at a generic level by combining a total damage score (such as the total SHS in RA or the modified Stoke Ankylosing Spondylitis Score in axSpA) with a generic measure for disability (such as the HAQ-score in RA or the Bath Ankylosing Spondylitis Functional Index in axSpA). By exploring the example of RA, we have disentangled in this thesis the subcomponents of the total SHS (JSN, (sub)luxation and erosion), and demonstrated that the strength of the associations between damage and function is dependent on the choice of the damage assessment and the choice of the disability assessment. Future investigators should better realize that these associations are not ‘a given’ but are dependent on the context, and that choices regarding the assessment of damage and disability matter with regard to the interpretation of the results.

As described above, principles of assessment are of pivotal importance for the understanding of our chronic inflammatory diseases, and we have argued how seemingly trivial aspects of assessment may have a far-stretching impact on our perception of the diseases in all their facets in this thesis.

Assessment may also influence our pathogenic thinking about disease, and we have given an example of this in the last chapter of this thesis, where we have investigated the longitudinal relationship between disease activity assessed by the ASDAS and inflammation assessed by MRI-SI: not only have we formally established a longitudinal association between ASDAS and MRI-SI, but more importantly have we demonstrated that this association was fundamentally different in males as compared to females. This observation, which cannot be explained by methodological artefacts, points to the importance of gender in explaining phenotypical differences in axSpA, and fits neatly with previous observations in ankylosing spondylitis that disease activity (assessed by ASDAS) and syndesmophyte formation are associated in males but not in females. Future researchers can take this example of dissociation of symptoms and signs of inflammation by gender as a starting point for further (translational and clinical) research.

**IN CONCLUSION**

In this thesis we have argued that methodological principles of assessment are as important as issues related to content of measurement instruments for measuring outcome in patients with inflammatory rheumatic conditions and for better understanding these diseases.

If improving the outcome of these diseases is a major goal of research in the field, sufficient attention for principles of assessment and proper implementation of them into clinical practice are of pivotal importance.
REFERENCES


