The gestalt of spondyloarthritis: From early recognition to long-term imaging outcomes

Sepriano, A.R.

Chapter 12

Summary and conclusions

SUMMARY AND CONCLUSIONS

With the work presented in this thesis we aimed at contributing to a better understanding of the concept of spondyloarthritis (SpA) as well as to elucidate how imaging of the axial skeleton can, more efficiently, be used to monitor and predict disease progression over time. Our main contributions to the field were as follows: First, we have addressed the issue of misclassification by the Assessment of SpondyloArthritis international Society (ASAS) classification criteria for axial SpA (axSpA) and peripheral SpA (pSpA) by evaluating their longitudinal validity against the rheumatologist's perception of the *Gestalt* of the disease. Second, we have, for the first time, shed light on the 'latent' phenotypes underlying the *Gestalt* of axSpA determined independently of the rheumatologist's opinion. This unprecedented approach has allowed us to better understand whether circularity played a role in the development of the criteria and how well the modern perception of axSpA, and the criteria developed in light of such a perception, truly overlap with the 'true Gestalt'. Third, we have proposed analytical approaches to improve our ability to reliably detect change of imaging outcomes, as well as predictive factors thereof, by limiting underlying assumptions and by giving more credit to measurement error. Fourth, we have used these approaches to provide further insights to the link between inflammation and damage in axSpA as well as to determine which outcomes should be prioritized for the monitoring of patients in clinical practice and in subsequent observational or interventional studies in early axSpA.

The studies presented in this thesis were performed in three independent cohorts: The ASAS cohort,[1, 2] the Spondyloarthritis Caught Early (SPACE) cohort,[3] and the Devenir des Spondyloarthropathies Indifférenciées Récentes (DESIR) cohort.[4] The ASAS cohort is a multicentre, prospective study in which patients had to fulfil one of two criteria to be included: i) chronic (>3 months) back pain of unknown origin (no definite diagnosis) with an age of onset below 45 years, with or without peripheral symptoms; ii) peripheral arthritis and/or enthesitis and/or dactylitis in the absence of current back pain with suspicion of SpA but no definitive diagnosis. SPACE is an ongoing multinational cohort in which consecutive patients aged ≥ 16 years with chronic back pain (CBP; ≥3 months, ≤2 years and onset <45 years) are included. DESIR is a longitudinal prospective cohort that includes adults aged over 18 and less than 50 years from 25 regional centres in France. At inclusion, patients have inflammatory back pain (IBP) with more than 3 months and less than 3 years and symptoms suggestive of SpA according to the opinion of the local investigator (level of confidence >5, scale 0-10).

In this final chapter we will summarize the main findings of the studies presented in this thesis and we will also discuss future perspectives as well as a research agenda for the topics that we have studied.

Classification and *Gestalt* of spondyloarthritis

In this thesis, we addressed the issue of misclassification by the ASAS SpA classification criteria. We started in **Chapter 2,** by determining, in the original ASAS cohort, what is the likelihood for a patient who classifies as positive to receive a clinical diagnosis of SpA after follow-up (positive predictive value; PPV). We have compared the baseline classification status according to the ASAS axSpA (also 'imaging arm' and 'clinical arm' separately), pSpA and SpA (both combined) to the clinical diagnosis ('external reference') made by ASAS experts after a mean follow-up of 4.4 years. Several important conclusions could be drawn from this exercise which, among others, argue against misclassification: first, we found that the large majority of the patients who fulfilled either the axSpA or pSpA criteria at baseline were in fact diagnosed as SpA at follow-up (PPV: 92%), which adds to the validity of the ASAS SpA criteria as a whole. Second, the pSpA criteria discriminated well between a clinical diagnosis of pSpA and no-SpA (PPV: 90%), even with similar proportions of peripheral arthritis in both groups (91%). There was, however, a significant difference in the proportion of enthesitis (pSpA: 60% vs no pSpA: 26%), which highlights the central role of enthesitis in the disease. It also indicates that the allowance of enthesitis as an entry feature does not lead to mislabelling, as previously suggested. Third, the PPV was equally high for the 'imaging arm only' (86%) and the 'clinical arm only' (88%) separately, which argues against misclassification by the 'clinical arm' and supports the view that the 'clinical arm' comprises a group of patients that belong to the SpA *Gestalt* as much as those fulfilling the 'imaging arm'. Third, almost all patients who had sacroiliitis on imaging classified positive for axSpA (98%) with many of those in the 'imaging arm' (irrespective of the 'clinical arm') having only sacroiliitis on MRI (62%). Since most of these were indeed diagnosed as axSpA at follow-up (PPV: 95%), our data reflect the dominant place that sacroiliitis on MRI holds in the ASAS axSpA criteria and testify to the high diagnostic value attributed to this feature by the rheumatologists. Even though the study presented in chapter 2 has a number of noteworthy limitations (e.g. losses to follow-up, missing data), sensitivity analyses taught us that these had little impact in our PPV calculations.

In Chapter 2 we tested validity of the ASAS SpA criteria against the expert opinion in the ASAS cohort. In addition, the ASAS classification criteria have been further challenged around the world in different cohorts. Some of these cohorts differ in several aspects from the ASAS cohort, thus yielding unique insights into the criteria performance and applicability in a broad population of patients. In **Chapter 3** we performed a systematic literature review (SLR) of the published data pertaining to the performance of the ASAS classification criteria tested against the rheumatologist's diagnosis. In total, data from eight independent cohorts including more than 5,500 patients, was evaluated. In addition to the original studies for the development of the ASAS axSpA and pSpA criteria,[1, 2] 5 studies assessing the ASAS axSpA criteria,[3, 5-8] one study the pSpA criteria,[9] and one study the combined SpA criteria (providing separate data also for the axSpA and pSpA criteria) were also included.[10]

The pooled analysis revealed an excellent sensitivity and specificity of the ASAS SpA (axSpA and pSpA combined) criteria (73%; 88%; respectively). Good performance was also noted for the axSpA criteria (sens: 82%; spec: 87%), which was robust to variations in the setting (hospital vs community), in symptom duration (<2 years vs ≥2 years) and type of population ('restricted' vs 'original ASAS population') (sens range: 78-85%; spec range: 90-93%). Of note, splitting the

axSpA criteria into 'imaging arm only' and 'clinical arm only' compromised sensitivity (26% and 23%, respectively), but retained very high specificity (97%; 94%). This finding is aligned with Chapter 2 and further argues in favour of the combined use of both 'arms' to avoid missing axSpA patients. The finding of a higher positive likelihood ratio (LR+) for the 'imaging arm' (13.6) compared to the 'clinical arm' (6.0) again highlights the rheumatologist's reliance on positive imaging findings for making an axSpA diagnosis. Similar to the ASAS axSpA criteria, the pooled specificity of the pSpA criteria was excellent (87%). However, sensitivity was much lower (62%). The low pooled sensitivity of the pSpA criteria was driven by the two studies including patients only based on the presence of peripheral arthritis, which once again highlights the relevance of enthesitis, and dactylitis, and adds to the credibility of the ASAS pSpA criteria, that include these clinical presentations.

Both in the development and validation of the ASAS SpA classification criteria, expert opinion was used as an external 'anchor' in the absence of a 'true' gold-standard. This approach, however, entails one fundamental limitation which might compromise the criteria content validity: circularity.[11] However, circularity is not necessarily detrimental, provided the rheumatologist's perception is a good reflection of the 'true *Gestalt*'. The only way to verify this premise is to exclude the opinion of the rheumatologist from the analysis. In **Chapter 4**, we have used latent class analysis (LCA), to reveal the 'latent' (i.e. unobserved) *Gestalt* of axSpA (independent of expert judgement) by splitting patients from the SPACE and DESIR cohorts (analysed separately) into mutually exclusive classes (or phenotypes) based on the covariance of observed SpA features. SpA features were selected by us *a priori*, without any assumption on their relative value to the *Gestalt* of axSpA.

We identified three separate clinical entities, together forming the *Gestalt* of axSpA, in both cohorts. We labelled these 'Pure axial SpA' ('Axial'), 'Axial SpA with peripheral signs' ('IBP+Peripheral') and 'Axial SpA at risk' ('At Risk'). The 'Axial' class is characterised by a high likelihood of axial imaging abnormalities (e.g. 74% and 84% likelihood of inflammation on MRI-SIJ in SPACE and DESIR, respectively), HLA-B27 positivity and male dominance. This phenotype closely resembles the rheumatologist's conventional clinical picture of axSpA. Thus, it is not surprising that the ASAS axSpA classification criteria (developed by experts) captured almost entirely the 'Axial' class (98% in SPACE and 93% in DESIR). The 'IBP+Peripheral' phenotype is defined by the presence of IBP (100%) in conjunction with peripheral signs and symptoms. These axSpA patients (mostly female) had back pain but were unlikely to be positive for sacroiliitis on imaging and HLA-B27. Thus, these patients rather fulfilled the pSpA (SPACE: 70%; DESIR: 82%) than the axSpA classification criteria (SPACE: 45%; DESIR: 58%). The 'At Risk' class is an entity characterised by the presence of presumed risk factors for axSpA (i.e. positive family history and HLA-B27) in association with IBP but only sporadically other SpA features. These patients often fulfil the ASAS axSpA classification criteria (SPACE: 60%; DESIR: 54%). It was remarkable that five-year transitions across classes were very unlikely.

Further discussion and future perspectives

The studies included in this thesis further testify to the good performance of the ASAS SpA classification criteria. Provided the criteria are applied in patients already with a clinical

diagnosis of SpA, as they are supposed to, the risk of misclassification is not higher than the risk with other diseases. For instance, the 2010 American College of Rheumatology (ACR)/European League Against Rheumatism (EULAR) classification criteria for rheumatoid arthritis (RA) were also developed to capture patients early in the disease course.[12] A meta-analysis evaluating their performance against several reference standards revealed an overall sensitivity of 82% and specificity of 61%.[13] The same compromise in specificity has not been observed in SpA, where we found a pooled sensitivity of 82% and specificity of 87% for the axSpA criteria. The fear that the 'clinical arm' would be responsible for misclassification does not find support in our data either. Rheumatologists from all over the world recognised patients fulfilling the 'clinical arm' within their *Gestalt* of SpA as much as those with evidence of sacroiliitis on imaging ('imaging arm').

Our data strengthen the view that non-radiographic axSpA (nr-axSpA) and radiographic axSpA (r-axSpA) are not separate entities as previously claimed, but rather part of the same disease continuum. The in-depth analysis into the *Gestalt* of axSpA in Chapter 4 clearly identified an 'Axial' phenotype without any distinction between nr-axSpA and r-axSpA. This expert judgement-free observation is in line with extensive evidence suggesting that a split of axSpA into nr-axSpA and r-axSpA is artificial.[14-16] More recently, a meta-analysis further demonstrated that both groups have mostly a similar disease presentation, as well as burden of disease.[17] Differences were also noted, but these should be interpreted thoughtfully. Patients with r-axSpA are most likely male, often smokers, have more often elevation of CRP, longer symptom duration and higher impairment of spinal mobility than patients with nr-axSpA. Available literature tells us that these translate into prognostic rather than diagnostic dissimilarities. In fact, these features are well known to associate with damage accrual in axSpA, thus obviously dominate the r-axSpA phenotype.[18-20] Similar poor prognostic 'funnelling' is seen in RA patients with erosive disease as compared to those without erosions.[12] However, it has never been argued that these two phenotypes in fact are two forms of the same disease.

In the introduction of this thesis we have used a Venn diagram (Figure 1) to illustrate the theoretical relationships of the concepts of the 'true *Gestalt*' of SpA ('C'), the rheumatologist's perception (diagnosis) of the *Gestalt* ('A') and the classification criteria ('B').This framework allows a critical discussion of the overlapping circles (misclassification and misdiagnosis) and is helpful for elucidating why inflammation on magnetic resonance imaging of the sacroiliac joints (MRI-SIJ) might have been awarded a too dominant place in the ASAS axSpA classification criteria due to circular reasoning.[11] Since sacroiliitis on MRI was at the basis of the nr-axSpA concept, which instigated the development of the ASAS axSpA criteria by experts, subsequent cross-validation against an expert's diagnosis may have resulted in classification criteria that reflect experts' beliefs ('A') rather than an objective presence of axSpA ('C'). Both in Chapter 2 and 3 we have indeed seen that the presence of inflammation on MRI-SIJ was almost synonymous to a clinical diagnosis in axSpA. This is in line with a recent study in the SPACE cohort in which patients with CBP received a diagnosis from a rheumatologist before and after the latter became aware of the result of imaging. Once known, a switch in diagnosis in 51% of the patients for whom the MRI-SIJ/pelvic radiograph result and the first diagnostic-judgement were incongruent was seen.[21] This is not so surprising though. Already in the original study on which the axSpA criteria were developed with 'paper patients', ASAS experts changed their diagnosis in 21% of the patients after knowing the result of the MRI-SIJ.[22]

Nonetheless, this evidence alone does not clarify whether the dominance of inflammation on MRI-SIJ is 'inappropriate' or just. In other words, it might be that such dominance according to the rheumatologist ('A') and translated into the axSpA criteria ('B') is coherent with the 'true *Gestalt*' of axSpA ('C'). The identification of the above-mentioned 'Axial' phenotype, with dominant imaging abnormalities (including inflammation on MRI-SIJ), independently of the rheumatologist's opinion, seems to suggest that. However, this phenotype, corresponds to less than 20% of the patients in SPACE and DESIR. This means that the phenotypical expansion of axSpA driven by MRI-SIJ, and reflected in the ASAS axSpA criteria, had indeed increased the 'AC' and 'BC' interactions (more 'true patients' diagnosed and classified), but at the cost of overlooking non-imaging-dominated phenotypes. The 'IBP+Peripheral', with female dominance, very low likelihood of abnormalities on axial imaging and a weak association with HLA-B27, is often seen by the expert clinician. However, this does not find reflection in the ASAS axSpA criteria, because these criteria either require positive imaging or HLA-B27-positivity to classify as positive. Thus, our data support that inappropriate circularity had indeed occurred when developing the ASAS axSpA classification criteria.

Our data help us to better appreciate the likely under-representation of the 'IBP+Peripheral' phenotype in previous studies (e.g. randomised clinical trials) in which the axSpA criteria were used for inclusion. However, a solution to the issue of over-valuing inflammation on MRI-SIJ and HLA-B27 positivity in the ASAS axSpA criteria is not straightforward. Even though the pSpA criteria perform reasonably well in capturing axSpA patients who have negative imaging and are HLA-B27 negative, they were developed to be applied in patients with exclusively peripheral manifestations, not in patients with current back pain as with the 'IBP+Peripheral' phenotype. More research is needed to better understand the overlap between axSpA and pSpA which is found to be greater than initially thought when the ASAS classification criteria were developed. One possible way forward is to better understand the 'cause' of IBP among female axSpA patients without imaging abnormalities. The lower than expected specificity of this feature needs also to be considered.[23, 24]

We live in the era of early diagnosis and early treatment. This modern paradigm, which undoubtedly brought many benefits to patients, also raises important challenges.[25] Axial SpA is difficult to diagnose and rheumatologists rely on pattern recognition for its identification, which is more than a simple sum of SpA features.[26] The SpA-pattern is less obvious in early disease when 'typical' features may still be absent which leads to uncertainty ('grey-zone'). The experienced clinician, will disentangle patients who do not have axSpA from those with the disease and appropriately handle those for whom either diagnosis is not beyond any doubt. However, others, when dealing with uncertain or difficult cases may be tempted to apply classification criteria to inform binary diagnostic judgements (e.g. axSpA vs no axSpA) that do not allow grey zones. We have shown that such clinicians are seriously in risk of 'overdiagnosing' and consequently 'overtreating' axSpA. We have labelled patients who drive this clinical conundrum, as 'At Risk', and for the first time provided a clear description of this entity. Presumed risk factors for axSpA (i.e. positive family history and HLA-B27) are the 'anchors', which often associate with IBP but only sporadically with other SpA features. It is easy to see why these patients often fulfil the ASAS axSpA criteria, especially the 'clinical arm' which require HLA-B27 to be positive in addition to two SpA features (family history and IBP). However, family history has been shown to be redundant when HLA-B27 is known.[27] Moreover, IBP is less

specific than initially thought which may contribute to overcalling axSpA when the ASAS axSpA criteria are wrongly used for diagnostic purposes. Of note, this is not a problem if the criteria are appropriately used only after a clinical diagnosis has been made, thus continuing efforts for education are key to avoid 'overdiagnosis' and 'overtreatment'.[28] Future studies should give resolution on the long-term outcomes of this and the previously described phenotypes of the *Gestalt* of axSpA.

Assessment of radiographic progression at the sacroiliac joints

Definite damage seen on the radiographs of the SIJ (X-SIJ) is defined according to the modified New York (mNY) grading system, as the presence of bilateral grade 2 or unilateral grade 3 or 4 'sacroiliitis' ('mNY-positive'), which is a key feature in the classification of r-axSpA.[29] However radiographic 'sacroiliitis' has been shown to be an unreliable finding, especially when assessed by untrained local readers.[30, 31] Determining the irreversible progression from mNY-negative (nr-axSpA) to mNY-positive (r-axSpA) is, arguably, even more ambiguous. Previous studies have focused only on positive change,[16] however, from a methodological perspective, bi-directional change, if present, cannot be ignored.

In **Chapter 5**, we compared two pelvic radiographs (X-SIJ) read several years apart (4.4 years on average) in patients with suspected SpA from the ASAS cohort in order to assess positive and negative change according to the mNY criteria. In total, 357 had paired X-SIJ available (at baseline and follow-up) read by the local observer. Of the 357 included patients, 17% (62/357) were mNY-positive at baseline. At follow-up this proportion increased to 22% (80/357). However, more than half (36/62) of those considered mNY-positive at baseline were assessed mNY-negative at follow-up. Assuming that structural damage in the SIJ is an inherently irreversible feature, and knowing that readers were aware of the correct time-order, these 'improvements' are very difficult to understand. The sobering truth is that progression, 'regression' and measurement error are not easy to disentangle in this setting. Thus, the question remained on what is the 'real' rate of radiographic progression at the SIJ level and how to handle measurement error on its calculation.

In previous studies, researchers have largely ignored or overlooked measurement error when reporting binary scores of progression, such as the change from mNY-negative to mNY-positive.[18, 19, 32-35] In **Chapter 6**, we undertook an analytical exercise that testifies to the truth of this statement and we made a plea for measurement error (or 'noise') not to be ignored when interpreting imaging studies.[36] We exposed the false assumptions underlying commonly used binary definitions of progression and proposed an analytical approach that we argue will best handle error. We evaluated the change between mNY-negative and mNY-positive after 5 years in the DESIR cohort, in which, contrary to the ASAS cohort, readers were blinded to time-order. In this setting, 'improvements' (i.e. change from mNY-positive to mNY-negative) should be judged as measurement error ('noise'). Each reader reported a binary score (mNY-positive vs mNY-negative) and the final status score was defined by the agreement of at least 2 of the 3 readers. The cross-tabulation between the baseline and 5-year reading, resulted in 3 possible change scores (mNY-positive to mNY-negative, no change, mNY-negative to mNY-positive).

At baseline, 62 (15%) of the 416 included patients were mNY-positive. Of the 354 mNY-negative patients at baseline, 24 (6.8%) changed into mNY-positive after 5 years. We labelled this figure as 'Crude progression', the simplest and most often used method to measure pelvic radiographic progression, which refers only to the rate of positive change. However, this method is spurious since it implies that the baseline reading is free of error and that a change in the opposite direction (here: 3/62: 4.8%) can be ignored. More recently, the method of 'Conditional net progression' has been proposed which gives credit to the rate of negative change by subtracting it from the rate of positive change (6.8%-4.8%: 2%).[18, 19] However, this method implicitly assumes that 'worsening' can only happen in patients who are mNY-negative at baseline and 'improvement' only in mNY-positive patients. Since readers are not aware which film is the baseline film, this assumption does not hold.

We therefore proposed a third method that we called (assumption-free) 'net progression' with which both 'positive change' and 'negative changes' are 'allowed' and scores of individual patients are not interpreted as 'true progression' or 'noise'. 'Net progression' is expressed as a percentage calculated as follows: number of positive changes minus number of negative changes divided by all patients [(24-3)/416=5%]. This calculation follows the same reasoning of the area under the curve (AUC) of probability plots (positive area minus negative area) that provides the mean continuous change score taking measurement error into account.[37] Thus, this ´net progression' yields the least biased estimates since it gives most credit to measurement error. That is, always includes error without a prior assumption on the imaging modality ability to reliably capture change.

Further discussion and future perspectives

Our data argue that, in a clinical practice setting, the arbitrary distinction between a mNY-negative and mNY-positive X-SIJ, is of little prognostic (but also diagnostic) utility. The same conclusion does not necessarily apply to its use in clinical research, where strategies to reduce measurement error can be implemented. Having films read by calibrated and trained central readers and the final scores determined by an 'agreement algorithm' (e.g. 2 out of 3) are some examples that reduce the 'noise'. However, even with such strategies, measurement error cannot be fully eliminated as shown by the occurrence of the unexpected improvements in Chapter 6. Thoughtful analytical approaches can help, as the one we propose (the so-called 'net progression'), to be used in clinical research to handle measurement error and to best estimate true progression of binary change-scores. Obviously, decreasing bias carries many benefits such as the better detection of treatment effects in randomised trials. Even though, we have used radiographic progression at the SIJ in axSpA to describe this method, its application extends to all examples where imaging scores on structural damage are obtained under blinded conditions.

Despite its merits, it should be noted, that this method implies that outcomes are irreversible (mainly structural damage), and are evaluated over short periods, as 'true repair' cannot be excluded with longer follow-up. Further studies should help us to understand the meaning of 'negative changes' in other settings than those with irreversible damage. They should also explain how 'true improvements' (i.e. repair) possibly contribute to the overall net progression. In addition, the 'assumption-free' method yields an average estimate of 'true progression' at the

group level (i.e. beyond measurement error) but does not translate to individual patients. So, it becomes impossible to declare an individual patient as a 'progressor'. Consequently, net progression is not to be used in prediction models aiming at determining factors associated with radiographic progression, such as inflammation on MRI.

Relationship between inflammation and structural damage

Patients with axSpA experience varying levels of radiographic progression, so identifying those with a higher likelihood of damage accrual is key for prognostic stratification. A significant effort has been put forward to the study of drivers of damage in axSpA, with inflammation receiving a large amount of attention by the international rheumatology community. At the time of the start of this thesis, there was already robust evidence supporting that inflammation associates with radiographic progression at the spinal level.[20, 38-42] It would be expected for the same association to be present at the SIJ level, however evidence supporting or rejecting this hypothesis was still scarce at that time.[18, 19]

In **Chapter 7**, we sought to determine whether inflammation on MRI-SIJ (ASAS definition) was associated with structural damage on X-SIJ (mNY grading) 5 years later in the DESIR cohort. As mentioned above, the estimated net progression from mNY-negative to positive was 5%. Other binary definitions of progression based on the grading of the SIJs proved to be more sensitive to change. Net progression was 13% for the change in at least one grade in at least one SIJ and 10% for change in at least one grade in at least one SIJ and a final absolute value of at least 2 in the worsened joint. Objective inflammatory markers (CRP and inflammation on MRI-SIJ) at baseline had a large impact on the likelihood of net progression especially among patients who were HLA-B27-positive. For instance, we found that patients who were HLA-B27-negative and who had a normal CRP and a negative MRI-SIJ had a likelihood of only 1% to (net) progress from mNY-negative to mNY-positive. In contrast, this likelihood was eighteen times higher (18%) if all three variables were positive.

We further tested whether inflammation on MRI-SIJ at baseline associated with subsequent radiographic progression at 5 years in multivariable models. Since the figure of net progression does not identify individual patients it could not be used in the models. Instead we defined our outcomes at 5 years irrespective of the scoring at baseline. For instance, the outcome was the mNY status (positive vs negative) at 5-years and not the change from mNY-negative to positive which would imply including in the analysis only those unreliably judged mNY-negative at baseline. This approach not only reduces bias but also increases the statistical power by including a larger number of patients. Two main conclusions could be drawn. First, we found that inflammation on MRI-SIJ was independently associated with damage at 5-years; second, that this association was modified by the baseline HLA-B27 status. That is, the effect of MRI-SIJ inflammation on mNY after 5 years was stronger in HLA-B27 positive patients [odds ratio (OR) 5.39 (95% CI: 3.25–8.94)] than in HLA-B27 negative patients [OR 2.16 (95% CI: 1.04–4.51)].

Although the results from this Chapter 7 are methodologically robust, they are also hard to translate to clinical practice where images are not read by multiple trained readers blinded to chronology. As demonstrated in Chapter 5, substantial 'noise' is expected in locally read films. Thus, it was unclear whether inflammation on MRI-SIJ as seen in clinical practice had the same

prognostic connotation as inflammation found with central reading. In **Chapter 8**, using locally read data from the ASAS and DESIR cohorts, we found that, despite all the 'noise', there was a clear prognostic value for objective inflammation: patients with a normal CRP and no inflammation on MRI-SIJ were unlikely to progress from mNY-negative to mNY-positive (ASAS: 4%; DESIR: 3%), whereas those who had both elevated CRP and inflammation on MRI-SIJ had very high probability to progress (ASAS: 33%; DESIR: 17%). In the multivariable analysis, inflammation on MRI-SIJ was found to be an independent predictor of the development of radiographic damage both in the ASAS (OR=3.2 [95% CI: 1.3; 7.9]), and DESIR (OR=7.6 [95% CI: 4.3; 13.2]) cohort. This study strongly argues in favour of the prognostic value of inflammation on MRI-SIJ as available in daily clinical practice.

Recently, there has been an increasing interest in the use of MRI not only to measure inflammation but also structural damage. Definitions of individual lesions (e.g. fatty lesions, erosions) have been proposed and composite scores validated.[43-46] In **Chapter 9** we tested, for the first time, the effect of inflammation on MRI of the SIJ and spine on the subsequent development of structural damage also measured on MRI over five years in the DESIR cohort. The presence of BME on MRI-SIJ at baseline was predictive of structural damage on MRI-SIJ 5 years later according to several binary definitions [range OR: 4.1-5.6]. Testing the association of interest on MRI-spine was challenged by low numbers of lesions, resulting in lower precision. Only the association between inflammation and ≥3 fatty lesions was statistically significant. In addition to the baseline models, we have shown that axial inflammation detected on MRI is longitudinally associated with subsequent development of structural damage also on MRI over 5 years (longitudinal models) both at the SIJ and at the spinal level. This study adds to the existing evidence by showing that the association between axial inflammation and structural damage can also be measured with MRI in patients with early axSpA.

Further discussion and future perspectives

Our findings add to the literature by showing that an association between inflammation and damage is seen at the SIJ level similar to what was previously shown for the spine. If fact, no matter how we look into it, an unequivocal positive association between these two types of lesions is always found: In early disease (Chapter 7 and 9) and in more established disease (Chapter 8); with local readings (Chapter 8) and with central readings (Chapter 7 and 9); with damage measured in conventional radiographs (Chapter 7 and 8) and in MRI (Chapter 9). Of note, we did not only find a predictive association between baseline inflammation and follow-up damage; We also found that having inflammation on MRI in one visit increased the likelihood of having structural damage in the subsequent visit up to 5 years of follow-up adjusting for the presence of damage at the first visit. These, so-called, time-lagged and 'autoregressive' models allow a more causal interpretation and add credibility to our findings.

It should be noted that all analyses were performed at the patient level. For instance, inflammation was said to be present on the SIJs according to the ASAS definition and damage if the mNY criteria for 'sacroiliitis' were met. Another interesting question would be to evaluate whether inflammation in one specific SIJ quadrant leads to subsequent damage at the same quadrant. Such an analysis likely yields further insights into the complex pathophysiology of

axSpA. For instance, it has been shown that inflammation at the vertebral unit level increased the likelihood of the formation of a new syndesmophyte in the same location 2 years later, but most new syndesmophytes appeared in vertebral units without signs of inflammation.[42] This remarkable finding suggests that unknown pathophysiological mechanisms may play a role in structural progression in axSpA. Some have argued that local injury and muscle dysfunction may be responsible for driving inflammation-independent damage.[47, 48] However, such mechanisms are not yet fully understood.

In Chapter 9 we found that inflammation on MRI of the SIJ and spine was associated with the subsequent formation of both fatty lesions and erosions. However, the interpretation of such association is not straightforward, because the underlying cause of these imaging findings remains to be clarified. One hypothesis defends that inflammation in axSpA fluctuates and that bone proliferation (e.g. formation of syndesmophytes in the spine and ankylosis of the SIJs) is a repair process that ignites only once inflammation subsides and is mediated by the formation of fatty lesions.[39, 42] On the other hand, if inflammation is persistent, repair is not possible and catabolic bone changes dominate, which in turn leads to bone destruction (e.g. erosions). Understanding the complex interplay between inflammation, bone formation and bone destruction in axSpA potentially has important therapeutical consequences. However, axSpA is a slowly progressive disease, thus long-term studies are needed to better understand the complex relationship between these abnormalities, including their sequence, frequency and rate of change over time. These studies pose some methodological challenges that we address in the following section.

Multilevel analysis of imaging data

Researchers designing long-term cohort studies, usually do not want to wait several years before their data can be analysed. A common practice is to 'split' the cohort into parts after a certain period of data collection.[36] For instance, patients included in the DESIR cohort are planned to be followed up to 10 years,[4] but it was already possible to use the available 5-year data to address several relevant research questions. In this setting imaging-data is usually read in several 'reading-waves'. In DESIR imaging data were, thus far, collected at baseline, 1 year, 2 years and 5 years and read by trained central readers in 3 'reading waves': in wave 1, baseline images were scored by 2 readers and 1 adjudicator (in case of disagreement); in wave 2, images from baseline, 1 and 2 years were also scored by 2 readers and one adjudicator; in wave 3, images from baseline, 2 and 5 years were scored by 3 central readers.

In previous chapters, we evaluated whether inflammation on MRI at baseline predicted subsequent structural lesions after 5 years in the DESIR cohort (baseline prediction models). Therefore, we have used data from the only 'reading-wave' that contains 5-year data: 'wave 3'. Although logical, this choice it is not without underlying assumptions that are often not fully appreciated. For instance, to be included in our baseline prediction models, patients had to have complete 5-year data ('completers analysis'), meaning that all those who had only follow-up imaging data at 2 years in wave 3 or images scored only in the other 2 waves were excluded (right censoring). In addition, yet another analytical choice had to be made. In wave 3 each of the 3 readers reported a score per patient, and these scores had to be somehow combined to

define the inflammation and damage variables. We have decided for the rule of 2 out of 3 for binary variables (e.g. inflammation present /absent according to the ASAS definition; or mNY positive/negative;) and the average of 3 for continuous variables (e.g. the SPARCC score; or the mNY grading). These combination algorithms are practical but are also not assumption-free and lead to loss of information.

We have already partially addressed these analytical problems in previous chapters. When we modelled baseline inflammation against the 5-year damage (both with combined scores), a 'traditional' logistic regression model would suffice. However, we have pursued two additional approaches. First, we have modelled the baseline inflammation against the 5-year structural outcomes using data from each individual reader separately. Since scores per reader are not independent, the assumption of independency of observations of logistic regression models does not hold. Therefore, we used generalised estimating equations (GEE), an analytical technique that takes correlated data into account. Our baseline predictive GEE models had one level of correlation (the reader), thus we called them 1-level GEE models. Still, these models leave out the 2-year visit data. So, we have used a so-called time-lagged longitudinal model that take all visits from wave 3 into account. That is, we have tested the association between baseline MRI-SIJ inflammation and 2-year damage and between 2-year inflammation and 5-year damage in a '2-level longitudinal GEE model'. The first level being the patient (repeated scores over time) and the second level the reader.

Despite the merits of the 1-level and 2-level GEE models, we are still disregarding a large number of scores yielded by the central readers and adjudicators from wave 1 and wave 2. In theory, including all data without aggregation-algorithms protects against bias, since the analyst does not need to intervene in data selection and computation. The 'trade off' is adding some variability ('noise') to the estimates, which may lead to a lower precision (i.e. wider 95% CI). Combining all available imaging-information has been previously shown to be a robust approach to analyse long-term imaging data in patients with RA using all available information.[49] In **Chapter 10** we investigated if an 'integrated analysis' affects the precision of estimates of change of imaging outcomes in patients with axSpA, with a conventional completers analysis as reference standard. To achieve that we had to consider one additional level of correlated data (the 'reading-wave' level). Thus, our data has 3 levels of correlation: Each visit is clustered within patient, each patient is clustered within reader, and each reader is clustered within the 'reading-wave'. To estimate the change over time while considering the various levels of correlation, 3-level GEE models were used. In these models, time was our independent variable of interest. The 'exchangeable' working correlation structure was found to best fit the data when taking the repeated scores over time per patient. The two 'higher' levels of correlation (reader and wave) were added as covariates to the model, an approach that has been previously proposed.[20]

We have challenged the 'integrated analysis' with a large number of continuous and binary outcomes reflecting: i. inflammation at the SIJ and spinal level (e.g. SPARCC score and the ASAS definition of inflammation at MRI-SIJ and MRI-spine);[50-54] ii. damage in spine radiographs (e.g. mSASSS and the presence of ≥ 1 syndesmophyte);[55] iii. damage on pelvic radiographs (e.g. mNY continuous grade and mNY-positivity);[29] iv. damage on MRI-SIJ (e.g. ≥ 3 fatty lesions);[45, 56] and v. damage on MRI-spine (e.g. ≥ 5 fatty lesions).[57] Each outcome was tested in a separate model. We did not focus on the point estimates of time, but rather on their

95% confidence intervals (CI). The narrower the 95%CI the higher the precision. We have compared the 'integrated analysis' with two types of completers analysis: i. a completers-only analysis, including only patients with complete 5-year follow-up, using scores from individual readers from wave 3 (adjusted for reader; 2-level); and aggregated completers analysis, using a combination algorithm (thus without reader adjustment; 1-level).

This analytical experiment proved the superiority of the 'integrated analysis' in comparison with both types of completers analysis in several ways. First, the 'integrated analysis' was more inclusive: out of 413 patients, the 'integrated analysis' models could include between 399 and 411 patients (depending on the outcome), whereas both 'completers analyses' included 364 at maximum. Second, we have proven that adding all data from individual readers and from all waves, without combination algorithms, does not affect the precision of the estimates of change. In fact, the 95% CI intervals were mostly similar across the three analytical approaches for most outcomes. Of note, the subtle increase in binary X-SIJ structural lesions (e.g. worsening of ≥ 1 grade in ≥1 SIJ with a grade ≥2 in the worsened joint at 5 years) was detected with more precision by the 'integrated analysis' analysis (95% CI: 1.06; 2.46) as compared to both completers analyses (2-level model 95% CI: 0.78; 2.32; 1-level model 95% CI: 0.81; 3.28). These data confirm that the 'integrated analysis' increases external validity (less patients excluded) without compromising (or even improving) internal validity.

We list above several of the imaging outcomes which have been developed and validated to assess inflammation and structural damage over time in patients with axSpA. However, direct comparisons of their sensitivity to change are mostly absent in the literature especially in early axSpA. [57-61] This knowledge gap precludes informed decisions on which outcome to prioritize in the follow-up and monitoring of patients with axSpA. Therefore, in **Chapter 11** we applied the 'integrated analysis' to study the sensitivity to change of the same scores used in chapter 10. Different to Chapter 10, however, here we focused on the point estimates of time (interpreted as change per year of the outcome) and not on the 95% CI. Since scores differ in the units of change and some are continuous while other binary, all variables were standardized (difference between the individual's value and the population mean divided by the population standard deviation [SD]) to allow comparisons. Each standardized variable has a mean of 0 and a variance of 1 and reads as the number of SD above (positive) or below (negative) the mean (standardized rate of change). We have compared the imaging outcomes sensitivity to change by calculating their relative standardized rate of change, i.e. the standardized yearly rate of change of an outcome divided by the corresponding rate of a reference imaging outcome.

We found that MRI outcomes of inflammation are more sensitive to change at the SIJ level than in the spine (e.g. range of relative standardized rate of change of spinal outcomes compared to the ASAS definition of a positive MRI-SIJ: 0.094; 0.531; i.e. all values far below 1). In addition, pelvic radiographs yield low sensitivity to change in detecting structural damage, while fatty lesions detected on MRI-SIJ emerged as a promising alternative (relative rate of change of ≥ 3 fatty lesions vs mNY: 6.2; i.e. far above 1). In contrast, MRI-spine (range rate of change: -0.013; 0.027) is not better than X-spine spine (range rate of change: 0.037; 0.043) in detecting structural changes in early axSpA patients.

Further discussion and future perspectives

The main strength of the proposed 'integrated analysis' is its ability to use all available imaging data in an assumption-free way with no intervention by the analyst. We have shown that this approach does not compromise precision, unlike what was expected since more data in principle implies more 'noise'. On the contrary, for outcomes that occurred infrequently over time, precision was even improved. Thus, this approach may be of special interest in studies with long-term follow-up, and/or when the outcomes are expected to occur infrequently over time. Our results are aligned with a previous study in RA, and may as well apply to other diseases. In fact, long-term studies evaluating imaging outcomes are highly relevant in the field of rheumatology. Over the years, several cohorts have been started to address some of the most fundamental and across-diseases long-lasting research questions: What is the natural history of the disease? Are we able to distinguish the patients who will follow a fairly benign path from those with a worse prognosis? Can we intervene in this process and ultimately drive lasting therapeutic benefits? The 'integrated analysis' may help researchers solving these questions in future studies focusing on imaging.

Arguably, by including all scoring data without 'hidden' assumptions, we may better approximate the 'true' point-estimates (the 'signal'). In fact, despite similar levels of precision, differences in the point-estimates were found between methods. That is, the estimated rates of change for the various imaging outcomes differed across analytical approaches. This might be, at least, in part explained by the fact that different patients are included depending on the method, with the 'integrated analysis' being the most inclusive. An alternative explanation might find ground on the old 'wisdom of the crowd' theory. An article published more than 30 years ago explains how this theory works by using a simple 'bean jar experiment'.[62] Briefly, a classroom of students is asked about the number of beans contained in a transparent jar. First, with no specific instructions each student yields an estimate and the average is calculated. Then the students are instructed about how to best estimate the number of beans and the exercise is repeated. Almost all students failed the exact number, but the average estimate came very close to the real number. Against expectations, however, the second average estimate (after instructions) was far less close to the exact number. The explanation for this counterintuitive finding is that the errors in each guess in the first exercise was independent from each other but became dependent in the second exercise when all students learned about the same instructions. This simple experiment tells us that combining multiple observations approximates the 'truth' provided the independence assumption holds. In theory, the larger the number of observations, the closer to the truth we will get.

The 'integrated analysis' uses a far greater number of observations than any of the 2 'completers analysis'. Let's use as starting point the 413 patients from DESIR who were included in the analysis of Chapter 10. For simplicity, let's assume that all 413 patients (p) complete the 5-year follow-up. These patients had at least one available score from at least one of the visits (t=4) read by at least one reader/adjudicator (r=3) from all available 'reading-waves' (w=3). In the 'completers analysis' using wave 3 only and aggregated outcomes (e.g. 2 out of 3) the statistical model could include at maximum 1,239 observations (413p * 3t *1r * 1w). This figure increases to 3,717 observations if we use individual-reader data instead of a combined outcome (413p * 3t *3r * 1w). Finally, with the 'integrated analysis' an impressive 14,868 observations can be

used at maximum (413p * 4t *3r * 3w). Obviously, the actual number of observations is variable depending on missing data, but in theory this approach can increase by almost 15 times the number of observations used to estimate our coefficient of interest. By applying a statistical technique that handles correlated data, we can then apply the principle of the 'wisdom of the crowd' and use all this information to approximate the 'true' estimate of change of each imaging outcome better than any of the 'completers analysis'.

The application of the 'integrated analysis' to compare the sensitivity to change of imaging outcomes yielded important insights, which may help in prioritizing imaging scoring methods in subsequent observational or interventional studies in early axSpA. Imaging abnormalities were found to be scarce and to hardly change over the period of 5 years at the spinal level regardless of the outcome and imaging modality. The opposite was observed at the SIJ level, which is aligned with the literature supporting that structural damage usually starts at the SIJ level and that the spine gets involved later on, and only in some of the patients.[63] We add to the literature by showing, for the first time, that MRI-SIJ outcomes of structural damage are more sensitive to change than the 'conventional' pelvic radiographic outcomes. This finding can be used to plan future studies aiming at studying progression of structural damage at the SIJ level, including those testing interventions aiming at disease modification.

Final comments

In this thesis we have pursued innovative analytical solutions for some of the most challenging questions in the field of SpA. We have gained better insights into the concept of axSpA by studying it independently of the rheumatologist's opinion. Our findings likely add knowledge to what axSpA really is. Future studies will learn us how much of these insights will translate into a better recognition of the disease in clinical practice and in better classifying them for research purposes. Since SpA is a slowly progressing disease, several years are needed to see meaningful changes in imaging abnormalities of the axial skeleton, which poses methodological challenges. We have shown that thoughtful analytical approaches, that make best use of imaging data, are helpful in better estimating progression, in unravelling its determinants and in clarify which outcomes are best to monitor disease. Efforts are made to further improve outcome measurement in axSpA, including the development of new imaging techniques, which can benefit from our proposed solutions to long-term imaging scoring. No question is too difficult when methodological rigor and creativity are put to work together:

Aut viam inveniam aut faciam

REFERENCES

1. Rudwaleit M, van der Heijde D, Landewé R, et al. The development of Assessment of SpondyloArthritis international Society classification criteria for axial spondyloarthritis (part II): validation and final selection. Ann Rheum Dis. 2009 Jun; 68(6):777-783.
2. Rudwaleit M, van der Heijde D, Landewé R, et al. The Assessment of SpondyloArthritis International Society classification criteria for peripheral spondyloarthritis and for spondyloarthritis in general. Ann Rheum Dis. 2011 Jan; 70(1):25-31.
3. van den Berg R, de Hooge M, van Gaalen F, et al. Percentage of patients with spondyloarthritis in patients referred because of chronic back pain and performance of classification criteria: experience from the Spondyloarthritis Caught Early (SPACE) cohort. Rheumatology (Oxford). 2015 Jul; 54(7):1336.
4. Dougados M, d'Agostino MA, Benessiano J, et al. The DESIR cohort: a 10-year follow-up of early inflammatory back pain in France: study design and baseline characteristics of the 708 recruited patients. Joint Bone Spine. 2011 Dec; 78(6):598-603.
5. Deodhar A, Mease PJ, Reveille JD, et al. Frequency of Axial Spondyloarthritis Diagnosis Among Patients Seen by US Rheumatologists for Evaluation of Chronic Back Pain. Arthritis Rheumatol. 2016 Jul; 68(7):1669-1676.
6. Lin Z, Liao Z, Huang J, et al. Evaluation of Assessment of Spondyloarthritis International Society classification criteria for axial spondyloarthritis in Chinese patients with chronic back pain: results of a 2-year follow-up study. Int J Rheum Dis. 2014 Sep; 17(7):782-789.
7. Molto A, Paternotte S, Comet D, et al. Performances of the Assessment of SpondyloArthritis International Society axial spondyloarthritis criteria for diagnostic and classification purposes in patients visiting a rheumatologist because of chronic back pain: results from a multicenter, cross-sectional study. Arthritis Care Res (Hoboken). 2013 Sep; 65(9):1472-1481.
8. Strand V, Rao SA, Shillington AC, et al. Prevalence of axial spondyloarthritis in United States rheumatology practices: Assessment of SpondyloArthritis International Society criteria versus rheumatology expert clinical diagnosis. Arthritis Care Res (Hoboken). 2013 Aug; 65(8):1299-1306.
9. van den Berg R, van Gaalen F, van der Helm-van Mil A, et al. Performance of classification criteria for peripheral spondyloarthritis and psoriatic arthritis in the Leiden Early Arthritis cohort. Ann Rheum Dis. 2012 Aug; 71(8):1366-1369.
10. Tomero E, Mulero J, de Miguel E, et al. Performance of the Assessment of Spondyloarthritis International Society criteria for the classification of spondyloarthritis in early spondyloarthritis clinics participating in the ESPERANZA programme. Rheumatology (Oxford). 2014 Feb; 53(2):353-360.
11. Landewé RB. Magnetic resonance imaging in the diagnosis of ankylosing spondylitis: be aware of gold standards and circularity. J Rheumatol. 2010 Mar; 37(3):477-478.
12. Aletaha D, Neogi T, Silman AJ, et al. 2010 rheumatoid arthritis classification criteria: an American College of Rheumatology/European League Against Rheumatism collaborative initiative. Ann Rheum Dis. 2010 Sep; 69(9):1580-1588.
13. Radner H, Neogi T, Smolen JS, et al. Performance of the 2010 ACR/EULAR classification criteria for rheumatoid arthritis: a systematic literature review. Ann Rheum Dis. 2014 Jan; 73(1):114-123.
14. Baraliakos X, Braun J. Non-radiographic axial spondyloarthritis and ankylosing spondylitis: what are the similarities and differences? RMD Open. 2015; 1(Suppl 1):e000053.
15. Sieper J, Hu X, Black CM, et al. Systematic review of clinical, humanistic, and economic outcome comparisons between radiographic and non-radiographic axial spondyloarthritis. Semin Arthritis Rheum. 2017 Jun; 46(6):746-753.
16. Sieper J, van der Heijde D. Review: Nonradiographic axial spondyloarthritis: new definition of an old disease? Arthritis Rheum. 2013 Mar; 65(3):543-551.
17. Lopez-Medina C, Ramiro S, van der Heijde D, et al. Characteristics and burden of disease in patients with radiographic and non-radiographic axial Spondyloarthritis: a comparison by systematic literature review and meta-analysis. RMD Open. 2019; 5(2):e001108.
18. Dougados M, Demattei C, van den Berg R, et al. Rate and Predisposing Factors for Sacroiliac Joint Radiographic Progression After a Two-Year Follow-up Period in Recent-Onset Spondyloarthritis. Arthritis Rheumatol. 2016 Aug; 68(8):1904-1913.
19. Poddubnyy D, Rudwaleit M, Haibel H, et al. Rates and predictors of radiographic sacroiliitis progression over 2 years in patients with axial spondyloarthritis. Ann Rheum Dis. 2011 Aug; 70(8):1369-1374.
20. Ramiro S, van der Heijde D, van Tubergen A, et al. Higher disease activity leads to more structural damage in the spine in ankylosing spondylitis: 12-year longitudinal data from the OASIS cohort. Ann Rheum Dis. 2014 Aug; 73(8):1455-1461.
21. Ez-Zaitouni Z, Landewé R, van Lunteren M, et al. Imaging of the sacroiliac joints is important for diagnosing early axial spondyloarthritis but not

all-decisive. Rheumatology (Oxford). 2018 Jul 1; 57(7):1173-1179.

22. Rudwaleit M, Landewé R, van der Heijde D, et al. The development of Assessment of SpondyloArthritis international Society classification criteria for axial spondyloarthritis (part I): classification of paper patients by expert opinion including uncertainty appraisal. Ann Rheum Dis. 2009 Jun; 68(6):770-776.
23. de Hooge M, van Gaalen FA, Renson T, et al. Low specificity but high sensitivity of inflammatory back pain criteria in rheumatology settings in Europe: confirmation of findings from a German cohort study Ann Rheum Dis. 2019 Nov;78(11):1605-1606.
24. Poddubnyy D, Callhoff J, Spiller I, et al. Diagnostic accuracy of inflammatory back pain for axial spondyloarthritis in rheumatological care. RMD Open. 2018; 4(2):e000825.
25. Landewé RBM. Overdiagnosis and overtreatment in rheumatology: a little caution is in order. Ann Rheum Dis. 2018 Oct; 77(10):1394-1396.
26. Ez-Zaitouni Z, Bakker PAC, van Lunteren M, et al. Presence of multiple spondyloarthritis (SpA) features is important but not sufficient for a diagnosis of axial spondyloarthritis: data from the SPondyloArthritis Caught Early (SPACE) cohort. Ann Rheum Dis. 2017 Jun; 76(6):1086-1092.
27. van Lunteren M, van der Heijde D, Sepriano A, et al. Is a positive family history of spondyloarthritis relevant for diagnosing axial spondyloarthritis once HLA-B27 status is known? Rheumatology (Oxford). 2019 Sep 1;58(9):1649-1654.
28. Rudwaleit M. Fibromyalgia is not axial spondyloarthritis. Rheumatology (Oxford). 2018 Sep 1; 57(9):1510-1512.
29. van der Linden S, Valkenburg HA, Cats A. Evaluation of diagnostic criteria for ankylosing spondylitis. A proposal for modification of the New York criteria. Arthritis Rheum. 1984 Apr; 27(4):361-368.
30. van den Berg R, Lenczner G, Feydy A, et al. Agreement between clinical practice and trained central reading in reading of sacroiliac joints on plain pelvic radiographs. Results from the DESIR cohort. Arthritis Rheumatol. 2014 Sep; 66(9):2403-2411.
31. van Tubergen A, Heuft-Dorenbosch L, Schulpen G, et al. Radiographic assessment of sacroiliitis by radiologists and rheumatologists: does training improve quality? Ann Rheum Dis. 2003 Jun; 62(6):519-525.
32. Mau W, Zeidler H, Mau R, et al. Outcome of possible ankylosing spondylitis in a 10 years' follow-up study. Clin Rheumatol. 1987 Sep; 6 Suppl 2:60-66.
33. Sampaio-Barros PD, Bortoluzzo AB, Conde RA, et al. Undifferentiated spondyloarthritis: a longterm followup. J Rheumatol. 2010 Jun; 37(6):1195-1199.
34. Sampaio-Barros PD, Conde RA, Donadi EA, et al. Undifferentiated spondyloarthropathies in Brazilians: importance of HLA-B27 and the B7-CREG alleles in characterization and disease progression. J Rheumatol. 2003 Dec; 30(12):2632-2637.
35. Schattenkirchner M, Kruger K. Natural course and prognosis of HLA-B27-positive oligoarthritis. Clin Rheumatol. 1987 Sep; 6 Suppl 2:83-86.
36. Landewé RBM, van der Heijde D. "Big Data" in Rheumatology: Intelligent Data Modeling Improves the Quality of Imaging Data. Rheum Dis Clin North Am. 2018 May; 44(2):307-315.
37. Landewé R, van der Heijde D. Radiographic progression depicted by probability plots: presenting data with optimal use of individual values. Arthritis Rheum. 2004 Mar; 50(3):699-706.
38. Poddubnyy D, Protopopov M, Haibel H, et al. High disease activity according to the Ankylosing Spondylitis Disease Activity Score is associated with accelerated radiographic spinal progression in patients with early axial spondyloarthritis: results from the GErman SPondyloarthritis Inception Cohort. Ann Rheum Dis. 2016 Dec; 75(12):2114-2118.
39. Machado PM, Baraliakos X, van der Heijde D, et al. MRI vertebral corner inflammation followed by fat deposition is the strongest contributor to the development of new bone at the same vertebral corner: a multilevel longitudinal analysis in patients with ankylosing spondylitis. Ann Rheum Dis. 2016 Aug; 75(8):1486-1493.
40. Maksymowych WP, Chiowchanwisawakit P, Clare T, et al. Inflammatory lesions of the spine on magnetic resonance imaging predict the development of new syndesmophytes in ankylosing spondylitis: evidence of a relationship between inflammation and new bone formation. Arthritis Rheum. 2009 Jan; 60(1):93-102.
41. Baraliakos X, Heldmann F, Callhoff J, et al. Which spinal lesions are associated with new bone formation in patients with ankylosing spondylitis treated with anti-TNF agents? A long-term observational study using MRI and conventional radiography. Ann Rheum Dis. 2014 Oct; 73(10):1819-1825.
42. van der Heijde D, Machado P, Braun J, et al. MRI inflammation at the vertebral unit only marginally predicts new syndesmophyte formation: a multilevel analysis in patients with ankylosing spondylitis. Ann Rheum Dis. 2012 Mar; 71(3):369-373.
43. Ostergaard M, Maksymowych WP, Pedersen SJ, et al. Structural lesions detected by magnetic resonance imaging in the spine of patients with spondyloarthritis - Definitions, assessment system, and reference image set. Journal of Rheumatology. 2009; 36(SUPPL. 84):18-34.
44. Krabbe S, Sorensen IJ, Jensen B, et al. Inflammatory and structural changes in vertebral

bodies and posterior elements of the spine in axial spondyloarthritis: construct validity, responsiveness and discriminatory ability of the anatomy-based CANDEN scoring system in a randomised placebo-controlled trial. RMD Open. 2018; 4(1):e000624.

45. Weber U, Lambert RG, Ostergaard M, et al. The diagnostic utility of magnetic resonance imaging in spondylarthritis: an international multicenter evaluation of one hundred eighty-seven subjects. Arthritis Rheum. 2010 Oct; 62(10):3048-3058.
46. Maksymowych WP, Lambert RG, Ostergaard M, et al. MRI lesions in the sacroiliac joints of patients with spondyloarthritis: an update of definitions and validation by the ASAS MRI working group. Ann Rheum Dis. 2019 Nov;78(11):1550-1558.
47. Schett G, Stolina M, Dwyer D, et al. Tumor necrosis factor alpha and RANKL blockade cannot halt bony spur formation in experimental inflammatory arthritis. Arthritis Rheum. 2009 Sep; 60(9):2644-2654.
48. White A, Abbott H, Masi AT, et al. Biomechanical properties of low back myofascial tissue in younger adult ankylosing spondylitis patients and matched healthy control subjects. Clin Biomech (Bristol, Avon). 2018 Aug; 57:67-73.
49. Landewé R, Ostergaard M, Keystone EC, et al. Analysis of integrated radiographic data from two long-term, open-label extension studies of adalimumab for the treatment of rheumatoid arthritis. Arthritis Care Res (Hoboken). 2015 Feb; 67(2):180-186.
50. Hermann KG, Baraliakos X, van der Heijde DM, et al. Descriptions of spinal MRI lesions and definition of a positive MRI of the spine in axial spondyloarthritis: a consensual approach by the ASAS/OMERACT MRI study group. Ann Rheum Dis. 2012 Aug; 71(8):1278-1288.
51. Lambert RG, Bakker PA, van der Heijde D, et al. Defining active sacroiliitis on MRI for classification of axial spondyloarthritis: update by the ASAS MRI working group. Ann Rheum Dis. 2016 Nov; 75(11):1958-1963.
52. Maksymowych WP, Inman RD, Salonen D, et al. Spondyloarthritis Research Consortium of Canada magnetic resonance imaging index for assessment of spinal inflammation in ankylosing spondylitis. Arthritis Rheum. 2005 Aug 15; 53(4):502-509.
53. Maksymowych WP, Inman RD, Salonen D, et al. Spondyloarthritis research Consortium of Canada magnetic resonance imaging index for assessment of sacroiliac joint inflammation in ankylosing spondylitis. Arthritis Rheum. 2005 Oct 15; 53(5):703-709.
54. Rudwaleit M, Jurik AG, Hermann KG, et al. Defining active sacroiliitis on magnetic resonance imaging (MRI) for classification of axial spondyloarthritis: a consensual approach by the ASAS/OMERACT MRI group. Ann Rheum Dis. 2009 Oct; 68(10):1520-1527.
55. Creemers MC, Franssen MJ, van't Hof MA, et al. Assessment of outcome in ankylosing spondylitis: an extended radiographic scoring system. Ann Rheum Dis. 2005 Jan; 64(1):127-129.
56. de Hooge M, van den Berg R, Navarro-Compan V, et al. Patients with chronic back pain of short duration from the SPACE cohort: which MRI structural lesions in the sacroiliac joints and inflammatory and structural lesions in the spine are most specific for axial spondyloarthritis? Ann Rheum Dis. 2016 Jul; 75(7):1308-1314.
57. de Hooge M, Pialat JB, Reijnierse M, et al. Assessment of typical SpA lesions on MRI of the spine: do local readers and central readers agree in the DESIR-cohort at baseline? Clin Rheumatol. 2017 Jul; 36(7):1551-1559.
58. Lukas C, Braun J, van der Heijde D, et al. Scoring inflammatory activity of the spine by magnetic resonance imaging in ankylosing spondylitis: a multireader experiment. J Rheumatol. 2007 Apr; 34(4):862-870.
59. Ramiro S, van Tubergen A, Stolwijk C, et al. Scoring radiographic progression in ankylosing spondylitis: should we use the modified Stoke Ankylosing Spondylitis Spine Score (mSASSS) or the Radiographic Ankylosing Spondylitis Spinal Score (RASSS)? Arthritis Res Ther. 2013 Jan 17; 15(1):R14.
60. Wanders AJ, Landewé RB, Spoorenberg A, et al. What is the most appropriate radiologic scoring method for ankylosing spondylitis? A comparison of the available methods based on the Outcome Measures in Rheumatology Clinical Trials filter. Arthritis Rheum. 2004 Aug; 50(8):2622-2632.
61. Ramiro S, Claudepierre P, Sepriano A, et al. Which scoring method depicts spinal radiographic damage in early axial spondyloarthritis best? Five-year results from the DESIR cohort. Rheumatology (Oxford). 2018 Nov 1;57(11):1991-2000.
62. Treynor JL. Market Efficiency and the Bean Jar Experiment. Financial Analysts Journal. 1987; 43(3):50-53.
63. Rudwaleit M, Haibel H, Baraliakos X, et al. The early disease stage in axial spondylarthritis: results from the German Spondyloarthritis Inception Cohort. Arthritis Rheum. 2009 Mar; 60(3):717-727.