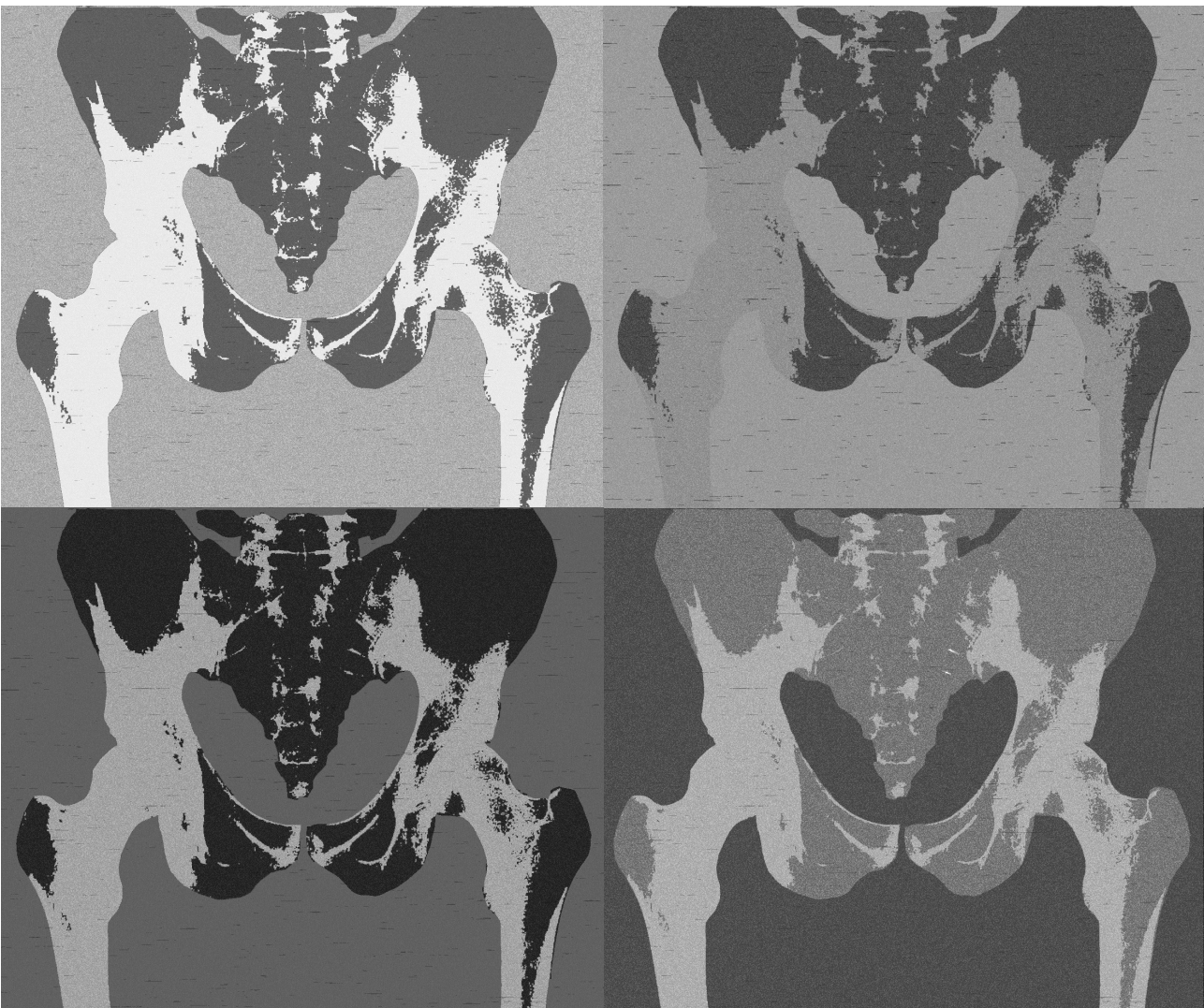
7

Metric properties of the SPARCC-score of the sacroiliac joints – Data from baseline, 3 and 12 months follow-up in the SPACE-cohort

Submitted

R. van den Berg
M. de Hooge
P. Bakker
F. van Gaalen
V. Navarro-Compán
K. Fagerli
R. Landewé
M. van Oosterhout
R. Ramonda
M. Reijnierse
D. van der Heijde



Abstract

Objectives

To evaluate metric properties of the SPARCC-score of the sacroiliac joints.

Methods

Patients ≥16 years with back pain (≥3 months, ≤2 years, onset <45 years) were included in the SPondyloArthritis Caught Early (SPACE)-cohort. Patients with (possible) axial spondyloarthritis (axSpA) had follow-up visits after 3 and 12 months. Patients were treated according to usual clinical practice. MRI-SIs were scored in two independent campaigns (1: baseline to 3 months and 2: baseline to 3 months to 12 months) by two different blinded reader pairs, applying the ASAS definition (positive versus negative MRI-SI) (discordant cases were adjudicated by a third reader) and the SPARCC-score (mean of two agreeing readers) was obtained. Agreement (kappa; positive/negative agreement) between SPARCC-score cut-off values and a consensus judgment of a positive MRI (ASAS definition) as external standard, change in SPARCC-score and smallest detectable changes (SDCs) over 3 and 12 months were calculated.

Results

SPARCC-score ≥2 showed best agreement with a positive MRI (both campaigns). In campaign 1, SPARCC-score changed (increased/decreased) in 70/151 patients; 26/70 change >SDC (3.4) of which 20 on stable treatment. In campaign 2, 20/68 patients changed in SPARCC-score; 11/20 change >SDC (2.1) of which 8 patients on stable treatment (3 months). Over 1 year, 23/74 patients changed in SPARCC-score; 14/23 change >SDC (2.4) of which 7 on stable treatment.

Conclusions

SPARCC-score ≥2 can be used as a surrogate for a consensus judgment of a positive MRI (ASAS definition) in clinical trials. The SDCs ranged from 2.1-3.4 dependent on reader pair and these are close to the proposed minimum important change of 2.5.

Introduction

A positive MRI of the sacroiliac joints according to the ASAS definition ('positive-MRI') [1] is part of the ASAS axial spondyloarthritis (axSpA) criteria [2] and is increasingly used to test eligibility of axSpA patients for clinical trials [3-5]. Within clinical trials, MRI-SI is often repeated over short periods of time (e.g. 12 weeks) to test efficacy of (especially biological) treatment in terms of changes in inflammation. For this efficacy read, the SPondyloArthritis Research Consortium of Canada (SPARCC)-score is frequently used as it measures inflammation on a continuous scale with good sensitivity to change [6, 7]. It is unknown what SPARCC-score cut-off value the equivalent is of a 'positive-MRI', which is needed to link the read for eligibility and the efficacy reading. This information would be useful for example to define groups with MRIs scored according to SPARCC-scores as having either or not a 'positive-MRI', to study differences in treatment response over time [3].

Treatment with biologicals may dramatically influence inflammatory signs on MRI [8-11] but inflammation may also spontaneously change over time in patients without treatment and in patients on stable non-biological treatment [12-14]. However, it is not clear how many SPARCC-score units these spontaneous changes represent [12-14]. Moreover, these spontaneous changes are likely to be different with variable lengths of follow-up. A minimally important change (MIC) of 2.5 SPARCC-units is proposed based on the patient global assessment as external anchor [15]. It is known that interreader reliability of SPARCC-scores at a fixed time point is acceptable to high (ICC 0.69-0.96) [16, 17], but reliability on change in SPARCC-scores over time has sparsely been reported and appeared to be moderate (ICC 0.52) in one small study with 20 patients [7]. Therefore, it would be of additional value to have knowledge about interreader reliability in terms of smallest detectable change (SDC), in order to be able to judge whether the SDC is sufficiently small to detect the proposed MIC.

The aim of this study is threefold: first, to define which SPARCC-score best approximates a 'positive-MRI' judgment; second, to establish an SDC for a 3-month period and for a 1-year period; third, to describe which variation in SPARCC-score over a 3-month and 1-year period can be expected in patients without (change in) treatment.

Methods

Study population

Data from the SPondyloArthritis Caught Early (SPACE)-cohort is used for this analysis. An extensive description of the SPACE-cohort is given elsewhere [18]. In short, the SPACE-cohort is an ongoing cohort started in January 2009, including patients aged 16 years and older with back pain (≥3 months, ≤2 years, onset <45 years) visiting the rheumatology outpatient clinics of five participating centers. Patients were not included if they had other painful conditions (not related to SpA) that could interfere with the evaluation of the disease. After signing informed consent, all patients underwent a diagnostic work-up at baseline, including MRI and plain radiographs of the SI-joints, HLA-B27 testing and examining for other SpA-features. Patients fulfilling the ASAS axSpA criteria or patients with possible axSpA were included for follow-up visits after 3 and 12 months. Possible axSpA was defined as the presence of at least one specific SpA-feature with a high positive likelihood ratio (LR+ above 6) or at least two less specific SpA-features (LR+ below 6), but not fulfilling the ASAS axSpA criteria [19].

MRI-SI

MR imaging was performed on a 1.5T scanner, acquiring T1-weighted Turbo Spin Echo (T1TSE) (TR 550/TE 10) and Short Tau Inversion Recovery (STIR) (TR 2500/TE 60) sequences, obtaining slices of 4mm thickness in coronal oblique view of the SI-joints.

All readers in this study were extensively trained in reading MRIs according to the ASAS definition and the SPARCC-score during a calibration session, supervised by a senior radiologist (MR) and a senior rheumatologist (DvdH), discussing definitions of lesions, examples and pitfalls. Next, all readers independently read 30 blinded MRIs to calculate agreement (κ=0.75 to κ=0.87 for the different pairs of readers), followed by a consensus meeting in which the supervising rheumatologist and radiologist of the calibration session participated too. The agreement was considered sufficiently high to start scoring the SPACE-cohort.

Two reading campaigns were performed, at different moments in time, by different pairs of readers (RvdB and MdH in campaign 1; PB and MdH in campaign 2) with partly overlapping patients and images. Patients in the first reading campaign were included between January 2009 and November 2012 in five different centers and patients in the second reading campaign were included between January 2009 and October 2013 in one center. In campaign 1, baseline and 3-month MRI-SIs were evaluated; in campaign 2, baseline, 3-month and 1-year MRI-SIs were evaluated. In both campaigns, MRI-SIs were independently read by the two trained readers on the fulfilllment of the ASAS definition [1] and according to the SPARCC-score [6], blinded for the time sequence of the MRI-SIs as well as for clinical and laboratory data.

An MRI-SI can be marked positive according to the ASAS definition if ≥1 bone marrow edema (BME) lesion highly suggestive of SpA is present on ≥2 consecutive slices, or if several BME lesions highly suggestive of SpA are visible on a single slice. The presence of only synovitis, enthesitis or capsulitis without BME is not sufficient for a positive MRI-SI [1]. In case the two readers disagreed on the presence of a 'positive-MRI', a third trained reader served as adjudicator (VNC in campaign 1; RvdB in campaign 2).

According to the SPARCC-score, the presence of increased signal corresponding to BME lesions highly suggestive of SpA is marked on the six middle slices of an MRI-SI, representing the largest proportion of the synovial compartment of the SI-joints. Each SI-joint is divided into four quadrants (upper iliac, lower iliac, upper sacrum and lower sacrum). The maximum score for two SI-joints on each slice is eight. In addition, a score for 'intensity' may be assigned to each SI-joint if an 'intense signal' is seen in any quadrant on each slice resulting in a maximum score of 12. The signal from presacral blood vessels defined a lesion that is scored as intense. Furthermore, a score for 'depth' may be assigned to each SI-joint if an homogeneous and unequivocal increase in signal is extending over a depth of at least 1 cm from the articular surface on each slice resulting in a maximum score of 12. A lesion is graded as deep if there is a homogeneous and unequivocal increase in signal extending over at least 1 cm from the articular surface. The total maximum SPARCC-score is 72 [6]. The mean SPARCC-scores of the two readers were used; in case there was a third reader involved, the mean of the SPARCC-scores of the two readers in agreement of a 'positive-MRI' for that particular case were used.

Treatment

Patients in the SPACE-cohort are not treated according to a fixed protocol, but according to usual clinical practice by their rheumatologist. Treatment with NSAIDs was recorded according to the ASAS recommendations, resulting in a 0-100 score whereby 0 means no NSAID intake at all, and 100 means a daily intake at a full dose over the whole period of interest [20]. Treatment with DMARDs and anti-TNF therapy was recorded as present or absent. To investigate variation in SPARCC-scores over time, patients were categorized according to their treatment over the period of interest: no treatment, stable NSAID and/or DMARD intake, and change in NSAID and/or DMARD intake. Patients receiving anti-TNF therapy during the period of interest were excluded from the analysis on variation in SPARCC-scores.

Statistical analysis

Baseline characteristics of patients in both groups were investigated using descriptive statistics. Agreement (Cohen's kappa) between MRI-positivity based on several SPARCC-score cut-off values (≥1, ≥2, ≥3 and ≥4) and the consensus judgment of a 'positive-MRI' , as external standard, was calculated using cross-tabulation. Agreement on positive cases (positive agreement) and on negative cases (negative agreement) was also calculated [21].
Changes in SPARCC-score over the period of interest (baseline - 3 months (both campaigns); baseline - 1 year (campaign 2)), were visualized in cumulative probability plots in which patients were grouped based on treatment. Next, SDCs were calculated based on a 95% level of agreement (95%LoA) between the two readers on the change scores for both baseline to 3-month and baseline to 1-year intervals, using the following formula: SDC = (1.96 * $SD_{\Delta change\text{-}scores}$) / (√2 *√k), whereby k represents the number of readers (equals 2 in this study) [22]. The SDCs are also displayed in Bland Altman plots, that plot the mean SPARCC-score changes of the two readers (X-axis) and the inter-reader differences in SPARCC-score changes (Y-axis). In addition, the mean of the inter-reader differences in SPARCC-score changes (which is a reflection of the systematic error between the two readers) and the 95% levels of agreement (LoA) are presented in these plots. SPSS software version 20.0 was used for statistical analysis.

Results

Patients with available baseline MRI-SI were included in the analysis of the agreement between the SPARCC-score cut-off value and 'positive-MRI' (n=294 (campaign 1) and n=249 (campaign 2)). There is a partial overlap (49.1%) between patients included in campaign 2 and those included in campaign 1. In both campaigns the population is young, with short symptom duration, around 1/3 of the patients is male and around 1/3 fulfilled the ASAS axSpA criteria (table 1).
A 3-month follow-up MRI-SI was available in 154 patients in campaign 1. However, 3/154 patients received anti-TNF therapy during this period and were therefore excluded from the follow-up part of the analysis of the SPARCC-score changes over time and SDCs. In campaign 2, a 3-month follow-up MRI-SI was available in 70 patients and in 76 patients a 1-year follow-up MRI-SI was available. Two patients received anti-TNF therapy, leaving MRI-SIs of 68 (campaign 1) and 74 patients (campaign 2) for follow-up analyses.

SPARCC-score cut-off

In both campaigns, there was a high level of agreement between MRI-positivity based on all tested SPARCC-score cut-off values and the consensus judgment of a 'positive-MRI' as external standard (table 2). A cut-off value of ≥2 showed the highest kappa values (0.94 in campaign 1 and 0.98 in campaign 2) and provided the best balance in terms of misclassifications in comparison to the external standard; 5 false-positive and 1 false-negative classifications in campaign 1; zero false-positive and 1 false-negative classification in campaign 2.

Smallest detectable change of SPARCC-score

Of the patients with available follow-up MRI, the mean SPARCC-score at baseline was 4.0 (SD 8.3) and 2.3 (SD 5.7) (campaign 1 and 2, respectively). At 3 months, the mean SPARCC-score was 3.4 (SD 6.7) and 1.6 (SD 3.8) (campaign 1 and 2, respectively), and at 1 year the mean SPARCC-score was 1.4 (SD 4.0) (campaign 2).
Bland and Altman plots show the mean of the two readers in SPARCC-score changes over the 3-month (campaign 1 and 2) and 1-year period (campaign 2) against the difference between the two readers in SPARCC-score changes over those periods (figure 1).

Table 1: Baseline characteristics of patients in reading campaign 1 and patients in reading campaign 2. A proportion (49.1%) of the patients was included in both campaigns.

	Reading campaign 1, n=294	**Reading campaign 2, n=249**
Age (years) at inclusion, mean ± SD	31.2 ± 10.4	31.1 ± 11.5
Male, n (%)	102 (34.7)	81 (32.5)
Duration of back pain (months), mean ± SD	13.1 ± 7.1	13.3 ± 7.4
HLA-B27 positive, n (%)	113 (38.4)	79 (31.7)
Pos. Fam. History SpA, n (%)	113 (38.4)	89 (35.7)
IBP, n (%)	195 (66.3)	142 (57.0)
Psoriasis, n (%)	28 (9.5)	26 (10.4)
Dactylitis, n (%)	16 (5.4)	8 (3.2)
Enthesitis, n (%)	49 (16.7)	24 (9.6)
Uveitis, n (%)	24 (8.2)	18 (7.2)
IBD, n (%)	20 (6.8)	19 (7.6)
Good response to NSAIDs, n (%)	112 (38.1)	69 (27.7)
Elevated CRP/ESR, n (%)	58 (19.7)	42(16.9)
Asymmetric lower limb arthritis, n (%)	48 (16.3)	26 (10.4)
Radiographic sacroiliitis*, n (%)	23 (7.8)	24 (9.6)
Sacroiliitis MRI**, n (%)	67 (22.8)	31 (12.4)
SPARCC-score, mean ± SD	2.9 ± 7.7	1.3 ± 4.4
CRP, mean ± SD	6.9 ± 13.0	7.3 ± 11.6
ASDAS, mean ± SD	2.6 ± 1.1	2.7 ± 0.8
BASDAI, mean ± SD	4.6 ± 2.5	4.6 ± 2.1
BASFI, mean ± SD	3.0 ± 2.3	3.2 ± 2.4
ASAS axSpA criteria positive, n (%)	119 (40.5)	83 (33.3)

*Radiographic sacroiliitis according to the modified New York criteria [25]. **Sacroiliitis on MRI according to the ASAS definition (consensus judgment) [1]. HLA-B27, Human Leukocyte Antigen; IBP, Inflammatory Back Pain; IBD, Inflammatory Bowel Disease; CRP, C-reactive protein; ESR, erythrocyte sedimentation rate.

The plots show that a large number of observations is clustered around the mean difference of 0, and that differences between readers occur with similar amplitude across the entire range of the SPARCC-score (a homoscedastic pattern). To visualize the high number of overlapping observations, series of ranges were defined in which all observations were grouped into their corresponding range, exponentially displayed on the X-axis. The SDC in campaign 1 over the 3-month period is 3.4 SPARCC-units, depicted in figure 1a as the dark grey area reflecting the SDC of both increased and decreased SPARCC-scores over time. The SDC in campaign 2 over the 3-month period is 2.1 SPARCC-units (figure 1b) and over the 1-year period 2.4 SPARCC-units (figure 1c).

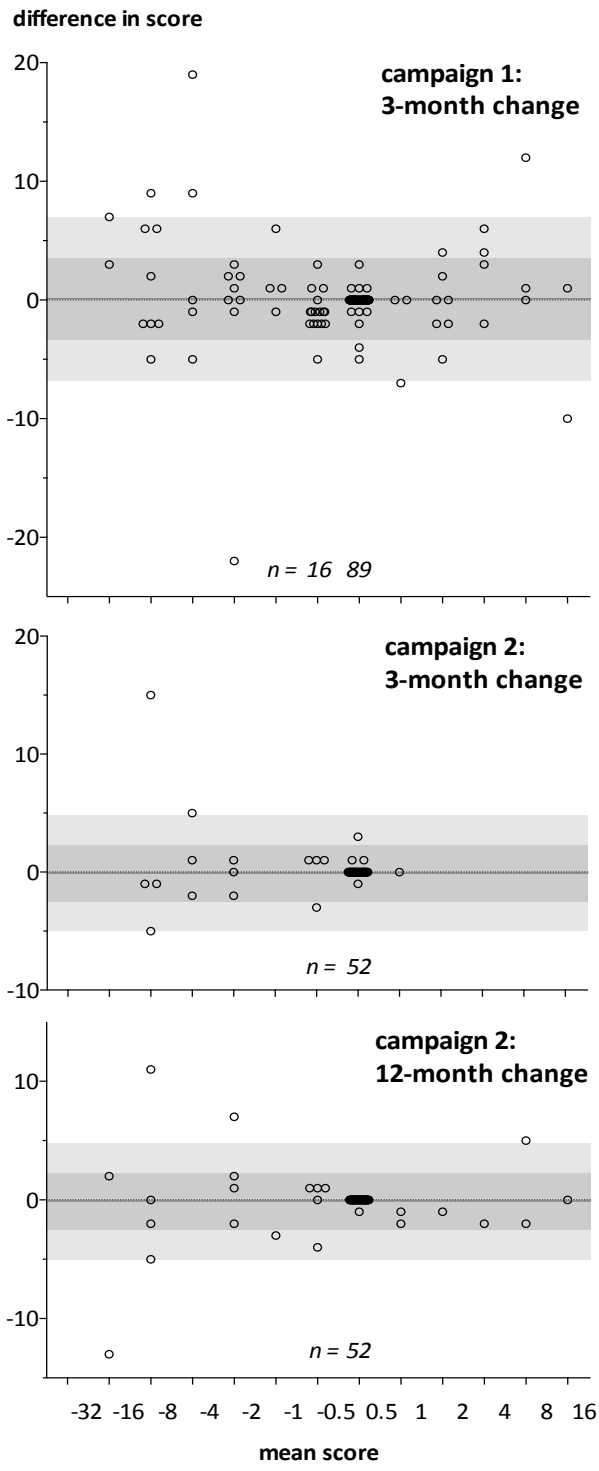


Figure 1: Bland Altman plots showing the mean SPARCC-score change of the two readers (X-axis) versus the delta SPARCC-score changes of the two readers (Y-axis). The large number of overlapping observations clustered around the mean difference of zero are displayed in series of ranges increasing exponentially on the positive side of zero and decreasing exponentially on the negative side (X-axis). The 'n' above the X-axis show the number of observations per group. The solid grey line represent the overall mean of the delta SPARCC-score changes (equivalent to systematic error between the two readers). The light grey area represents the 95% levels of agreement (LoA), and the dark grey area represents the smallest detectable change (SDC) in both directions (increase in SPARCC-score and decrease in SPARCC-score over time). The reader is referred to the text for further clarification.

Figure 1a: mean of the delta SPARCC-scores 0.1 (95% LoA -6.8 to 7.0); SDC 3.4. Observations are clustered in the range -0.5 to 0.5 (n=89) and the range -1 to -0.5 (n=16).

Figure 1b: mean of the delta SPARCC-scores 0.2 (95% LoA -4.0 to 4.4); SDC 2.1. Observations are clustered in the range -0.5 to 0.5 (n=52).

Figure 1c: mean of the delta SPARCC-scores -0.1 (95% LoA -5.0 to 4.8); SDC 2.4. Observations are clustered in the range -0.5 to 0.5 (n=52).

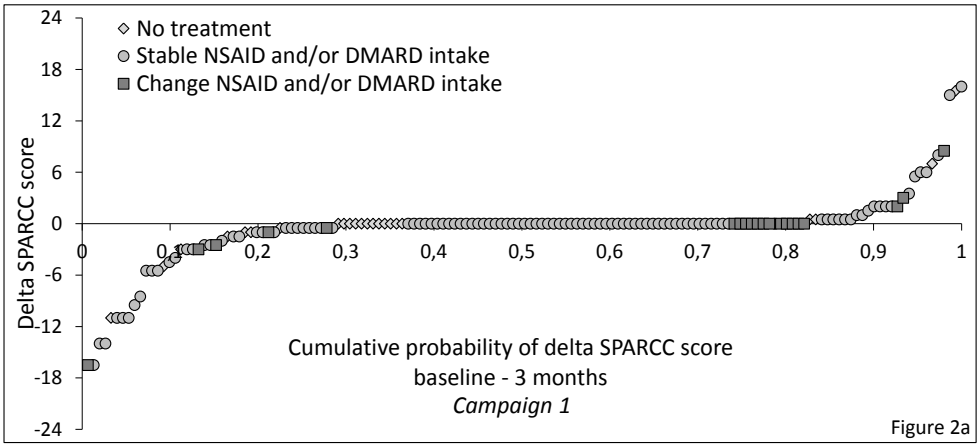
Table 2: Various SPARCC cut-off values tested against the ASAS definition of a positive MRI, in reading campaign 1 and reading campaign 2.

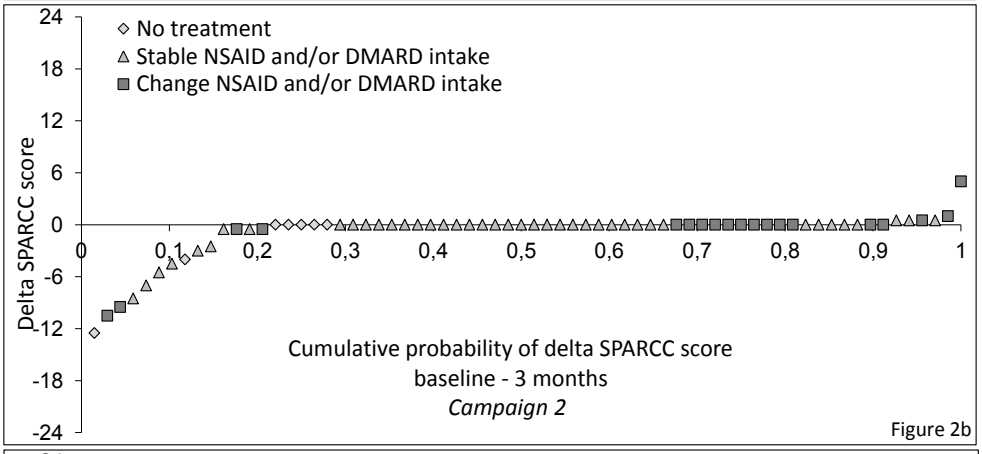
Reading campaign 1 (n=294)		
	positive MRI (ASAS)	**negative MRI (ASAS)**
SPARCC ≥1	67	21
SPARCC <1	0	206
Kappa: 0.82	PA: 95.2%	NA: 86.5%
SPARCC ≥2	66	5
SPARCC <2	1	222
Kappa: 0.94	PA: 98.7%	NA: 95.7%
SPARCC ≥3	57	1
SPARCC <3	10	226
Kappa: 0.89	PA: 97.6%	NA: 91.2%
SPARCC ≥4	47	1
SPARCC <4	20	226
Kappa: 0.77	PA: 95.6%	NA: 81.7%
Reading campaign 2 (n=249)		
	positive MRI (ASAS)	**negative MRI (ASAS)**
SPARCC ≥1	31	5
SPARCC <1	0	213
Kappa: 0.91	PA: 98.8%	NA: 92.5%
SPARCC ≥2	31	1
SPARCC <2	0	217
Kappa: 0.98	PA: 99.8%	NA: 98.4%
SPARCC ≥3	25	0
SPARCC <3	6	218
Kappa: 0.88	PA: 98.6%	NA: 89.3%
SPARCC ≥4	21	0
SPARCC <4	10	218
Kappa: 0.79	PA: 97.8%	NA: 80.8%

PA, positive agreement is the agreement on positive cases. NA, negative agreement is the agreement on negative cases.

Change in SPARCC-scores over 3 months and 1 year

Eighty-one out of 151 patients in campaign 1 (53.6%) showed no change in SPARCC-score over the 3-month period of which 75/81 (92.6%) had a SPARCC-score of 0 at both time points. In the 70 out of 151 patients (46.4%) showing a change in SPARCC-score, 27 increased and 43 decreased (mean change -1.1 (SD 6.3); median change -0.5 (range -16.5 to 16.0)) (figure 2a & table 3). In 26 out of 70 patients (37.1%) with SPARCC-score changes, the change was more than the SDC (3.4); 16 patients decreased (2 patients without treatment, 11 with stable NSAIDs intake, 2 with stable NSAIDs and DMARD intake, 1 patient started





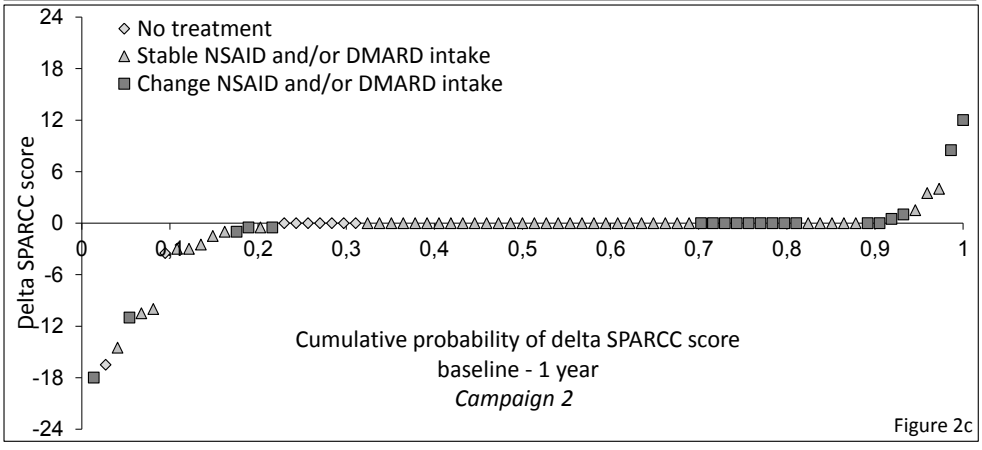


Figure 2: Cumulative probability plots of all delta SPARCC-scores over a 3-month period (2a and 2b) and a 1-year period (2c) with different symbols indicating the treatment over the investigated period.

Table 3: All changes in SPARCC-score in patients grouped according to treatment.

	No treatment	Stable NSAIDs/ DMARDs	Start NSAIDs/ DMARDs	Stop NSAIDs/ DMARDs
Campaign 1 – baseline to 3 months				
No SPARCC-score change	N=13	N=56	N=7	N=5
Increase in SPARCC-score (mean change (SD); range)	N=4 5.9 (SD 7.1) 0.5 to 15.5	N=20 3.7 (4.6) 0.5 to 16.0	N=3 4.5 (3.5) 2 to 8.5	-
Decrease in SPARCC-score (mean change (SD); range)	N=9 -2.8 (3.4) -11.0 to -0.5	N=29 -4.9 (4.9) -16.5 to -0.5	N=5 -4.7 (6.7) -16.5 to -0.5	-
Campaign 2 – baseline to 3 months				
No SPARCC-score change	N=4	N=31	N=8	N=5
Increase in SPARCC-score (mean change (SD); range)	-	N=5 0.6 (0.2) 0.5 to 1.0	N=1 5 (-) -	-
Decrease in SPARCC-score (mean change (SD); range)	N=2 -5.0 (6.4) -12.5 to -0.5	N=10 -4.6 (3.3) -10.5 to -0.5	N=2 -6.5 (8.5) -12.5 to -0.5	-
Campaign 2 – baseline to 1 year				
No SPARCC-score change	N=10	N=28	N=7	N=6
Increase in SPARCC-score (mean change (SD); range)	-	N=3 3.0 (1.3) 1.5 to 4.0	N=1 12 (-) -	N=3 3.3 (4.5) 0.5 to 8.5
Decrease in SPARCC-score (mean change (SD); range)	N=3 -6.8 (8.5) -16.5 to -0.5	N=8 -5.8 (5.1) -14.5 to -1.0	N=1 -0.5 (-)	N=4 -7.6 (8.4) -18.0 to -0.5

NSAIDs intake) and 10 patients increased (2 without treatment, 7 with stable NSAIDs intake, 1 started NSAIDs intake). In the remaining 44 patients (62.9%) the SPARCC-score changes were within the area still compatible with measurement error. intake but continued NSAID intake). In the remaining 9 patients (39.1%) SPARCC-score changes were not beyond measurement error.

In campaign 2, two follow-up intervals for the same patients are available. Over the 3-month period, SPARCC-score did not change in 48 out of 68 patients (70.6%); 46/48 patients (95.8%) had a SPARCC-score of 0 at both time points. In the remaining 20 patients (29.4%) the SPARCC-score changed; 14 patients showed a decrease and 6 patients an increase (mean change -3.1 (SD 4.6); median change -1.5 (range -12.5 to 5) (figure 2b & table 3). Eleven out of 20 patients (55.0%) showed a SPARCC-score change >SDC (2.1); 10 patients decreased (1 without treatment, 6 with stable NSAIDs intake, 2 with stable NSAIDs and DMARD intake, 1 started NSAIDs intake) and 1 patient increased (started NSAIDs intake). The remaining 9 patients (45.0%) had SPARCC-score changes still compatible with measurement error.

The results over the 1-year period in campaign 2 are similar to the results over the 3-month period in campaign 2, although more variation between patients is seen; 51/74 patients (68.9%) did not show a change in SPARCC-score, of which 50 patients (98.0%) had a SPARCC-score of 0 at both time points. The remaining 23 patients (31.1%) showed a change in SPARCC-score; 16 patients decreased and 7 increased (mean change -2.9 (SD 7.5); median

change -1.0 (range -18.0 to 12.0)) (figure 2c & table 3). Fourteen out of the 23 patients (60.9%) showed a SPARCC-score change of more than the SDC (2.4); 10 patients decreased (2 without treatment, 4 with stable NSAID intake, 2 with stable DMARD intake, 1 stopped NSAID intake, 1 started but stopped again NSAID intake) and 4 patients increased (1 with stable NSAID intake, 1 stopped NSAID intake, 1 started NSAID intake, 1 stopped DMARD
The majority of the patients showing changes in SPARCC-score of more than the SDC in both campaigns (20/26 (76.9%; campaign 1), 8/11 (72.7%; 3-month period campaign 2) and 7/14 (50.0%; 1-year period campaign 2)) were on stable NSAID and/or DMARD intake.

Discussion

This study performed in the SPACE-cohort has shown in two campaigns that a cut-off value of 2 SPARCC-units is best compatible with a consensus judgment of a positive versus negative MRI according to the ASAS definition. These results were not unexpected as the ASAS definition of a positive MRI-SI includes - apart from a qualitative part (BME lesions highly suggestive of spondyloarthritis) - a quantitative part that requires at least one BME lesion visible on at least 2 consecutive slices or several lesions on a single slice [1]. However, in theory, a SPARCC-score can be high because of the presence of several small lesions (highly suggestive of SpA), scattered over several slices (e.g. one lesion on slice 1, another lesion on slice 4 and another lesion on slice 6) but still not fulfilling the ASAS definition. A SPARCC-score can also be high if one lesion is assigned as 'intense' or 'deep', while it is only visible on 1 slice. Moreover, the SPARCC-score prescribes that lesions are scored in the six middle slices, while the ASAS definition takes all slices into account [1, 6]. Occasionally, part of a lesion may be visible on only one of the six middle slices, while the remaining part of the lesion is visible outside those six middle slices, or a slice outside those middle six shows several lesions. However, these considerations are mainly theoretical and do not appear very frequently. Therefore, a SPARCC cut-off level of 2 units may serve as a surrogate for the ASAS definition of a positive MRI and could be used in clinical trials with central efficacy reading in order to derive a dichotomy (positive versus negative) for prognostic reasons.
The SDCs in campaign 2 (2.1 SPARCC-units over 3 months and 2.4 over 1 year) are close to the proposed MIC of 2.5 SPARCC-units, which was calculated using pooled changes over 12 and 52 weeks [15], but the SDC of campaign 1 (3.4) is slightly higher. This suggests that the previously proposed MIC is close to measurement error in our study based on two different reader pairs and different periods of follow-up.
A large proportion of the SPARCC-score changes seen in the patients in both reading campaigns could be considered as noise as these changes are smaller than the SDCs (62.9% and 45% (3-months, campaign 1 and 2) and 39.1% (1-year in campaign 2). To investigate the influence of non-biological treatment on inflammation on MRI-SI, only patients with SPARCC-score changes greater than the SDC were taken into account. Somewhat surprisingly, the majority of patients with a change in SPARCC-score were on stable NSAID and/or DMARD treatment. Some patients taking stable doses of NSAIDs increased in SPARCC-score while others who were also on stable NSAIDs intake decreased in SPARCC-score. These results are in line with the results found in trials where patients using NSAIDs – either in an open label trial or in a placebo group – showed also both increased and decreased inflammation scores on MRI-SI over 6 and 16 weeks, respectively [14, 23]. Moreover, also patients with stable background treatment in the placebo group of the ABILITY-1 trial slightly decreased in SPARCC-score at group level, like we found in this study [3].
Although too few patients in the SPACE-cohort used DMARDs to draw conclusions on the effect of DMARDs, comparable effects can be expected. The comparator group in the ESTHER trial using sulfasalazine showed a mean decrease of 1.7 and 1.9 SPARCC-units over 24 and 48 weeks, respectively [13]. In the comparator group of another trial where patients used methotrexate, a mean of 1.4 (95%CI -0.8 to 3.5) inflammatory lesions resolved over

30 weeks [24]. Although an overall decrease in inflammation score was seen in these trials, some patients increased in inflammation score on MRI-SI when looking at the individual level [13, 24]. These results indicate that in patients on stable treatment changes in BME on MRI-SI that are beyond measurement error may occur, which may point to true fluctuation in inflammatory activity over time.

The direct comparisons of our results with the results of drug efficacy trials is difficult as the SPACE-cohort is an observational cohort including unselected patients with back pain of short duration resulting in a heterogeneous patient population, with low numbers of a 'positive-MRI' and low baseline mean SPARCC-scores, while drug efficacy trials select patients with high levels of disease activity. In patients selected because of a high level of disease activity a decrease in scores is more likely (regression to the mean) in comparison to an unselected group of patients. Thus, the patients in the SPACE-cohort will likely not be representative of patients in trials. Nevertheless, we have also observed an overall decrease in the SPACE-cohort, just as in the trials. This might be due to the fact that patients preferably seek help in case of maximum complaints, which is by default the time point of inclusion in the SPACE-cohort. It is possible that the results would have been different if this study had been performed in a long-standing or severely diseased group of patients. Furthermore, the SPACE-cohort is not designed to investigate the effects of treatment on inflammation on MRI. For example, and in contrast to drug efficacy trials, there is not a good relation between the start date of therapy and the date of the MRI.

Another possible limitation is that the readers have given their judgement based on the ASAS definition immediately after the evaluation according to the SPARCC-score. Since the quantitative part of the ASAS definition resembles a SPARCC-score of 2, the choice of the value of 2 as the best SPARCC-score to serve as cut-off level for negative and positive MRI may not be entirely independent. It would have been better if different scores were acquired independently, or even by different readers, as is frequently the case in clinical trials.

In conclusion, a SPARCC-score of 2 as cut-off value best reflects the caesura between a positive and negative MRI according to the ASAS definition. This cut-off can be used (in clinical trials) in order to create a dichotomous MRI variable of potential prognostic interest. The SDCs we have obtained in our two experiments are close enough to the proposed MIC of 2.5 SPARCC-units, which adds credibility to a cut-off level of 2.5 units in that it represents a true difference rather than only measurement error. Surprisingly, while patients are on stable treatment, true (>SDC) changes in SPARCC-score over time (both increases and decreases) were frequently observed. This observation strongly suggests that MRI-activity fluctuates over time.

References

1. Rudwaleit M, Jurik A, Hermann K, *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;68:1520-7.
2. 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;68:777-83.
3. Sieper J, van der Heijde D, Dougados M, *et al.* Efficacy and safety of adalimumab in patients with non-radiographic axial spondyloarthritis: results of a randomised placebo-controlled trial (ABILITY-1). Ann Rheum Dis 2013;72:815-22.
4. Sieper J, Lenaerts J, Wollenhaupt J, *et al.* Efficacy and safety of infliximab plus naproxen versus naproxen alone in patients with early, active axial spondyloarthritis: results from the double-blind, placebo-controlled INFAST study, Part 1. Ann Rheum Dis 2014;73:101-7.
5. Landewé R, Braun J, Deodhar A, *et al.* Efficacy of certolizumab pegol on signs and symptoms of axial spondyloarthritis including ankylosing spondylitis: 24-week results of a double-blind randomised placebo-controlled Phase 3 study. Ann Rheum Dis 2014;73:39-47.
6. Maksymowych W, Inman R, 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;53:703-9.
7. Landewé R, Hermann K, van der Heijde D, *et al.* Scoring sacroiliac joints by magnetic resonance imaging. A multiple-reader reliability experiment. J Rheumatol 2005;32:2050-5.
8. Braun J, Baraliakos X, Hermann K, *et al.* Golimumab reduces spinal inflammation in ankylosing spondylitis: MRI results of the randomised, placebo- controlled GO-RAISE study. Ann Rheum Dis 2012;71:878-84.
9. Baraliakos X, Brandt J, Listing J, *et al.* Outcome of patients with active ankylosing spondylitis after two years of therapy with etanercept: clinical and magnetic resonance imaging data. Arthritis Rheum 2005;53:856-63.
10. Lambert R, Salonen D, Rahman P, *et al.* Adalimumab significantly reduces both spinal and sacroiliac joint inflammation in patients with ankylosing spondylitis: a multicenter, randomized, double-blind, placebo-controlled study. Arthritis Rheum 2007;56:4005-14.
11. Braun J, Landewé R, Hermann K, *et al.* Major reduction in spinal inflammation in patients with ankylosing spondylitis after treatment with infliximab: results of a multicenter, randomized, double-blind, placebo-controlled magnetic resonance imaging study. Arthritis Rheum 2006;54:1646-52.
12. van Onna M, Jurik A, van der Heijde D, *et al.* HLA-B27 and gender independently determine the likelihood of a positive MRI of the sacroiliac joints in patients with early inflammatory back pain: a 2-year MRI follow-up study. Ann Rheum Dis 2011;70:1981-5.
13. Song I, Hermann K, Haibel H, *et al.* Effects of etanercept versus sulfasalazine in early axial spondyloarthritis on active inflammatory lesions as detected by whole-body MRI (ESTHER): a 48-week randomised controlled trial. Ann Rheum Dis 2011;70:590-6.
14. Jarrett S, Sivera F, Cawkwell L, *et al.* MRI and clinical findings in patients with ankylosing spondylitis eligible for anti-tumour necrosis factor therapy after a short course of etoricoxib. Ann Rheum Dis 2009;68:1466-9.
15. Maksymowych W, Lambert R, Brown L, *et al.* Defining the Minimally Important Change for the SpondyloArthritis Research Consortium of Canada Spine and Sacroiliac Joint Magnetic Resonance Imaging Indices for Ankylosing Spondylitis. J Rheumatol 2012;39:1666-74.
16. van der Heijde D, Sieper J, Maksymowych W, *et al.* Spinal inflammation in the absence of sacroiliac joint inflammation on magnetic resonance imaging in patients with active nonradiographic axial spondyloarthritis. Arthritis Rheumatol 2014;66:667-73.
17. Weber U, Maksymowych W, Jurik A, *et al.* Validation of whole-body against conventional magnetic resonance imaging for scoring acute inflammatory lesions in the sacroiliac joints of patients with spondylarthritis. Arthritis Rheum 2009;61:893-9.

18. 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) 2013;52:1492-9.
19. Rudwaleit M, van der Heijde D, Khan M, *et al.* How to diagnose axial spondyloarthritis early. Ann Rheum Dis 2004;63:535-43.
20. Dougados M, Simon P, Braun J, *et al.* ASAS recommendations for collecting, analysing and reporting NSAID intake in clinical trials/epidemiological studies in axial spondyloarthritis. Ann Rheum Dis 2011;70:249-51.
21. Cicchetti D, Feinstein A. High agreement but low kappa: II. Resolving the paradoxes. J Clin Epidemiol 1990;43:551-8.
22. Bruynesteyn K, Boers M, Kostense P, *et al.* Deciding on progression of joint damage in paired films of individual patients: smallest detectable difference or change. Ann Rheum Dis 2005;64:179-82.
23. Barkham N, Keen H, Coates L, *et al.* Clinical and imaging efficacy of infliximab in HLA-B27-Positive patients with magnetic resonance imaging-determined early sacroiliitis. Arthritis Rheum 2009;60:946-54.
24. Marzo-Ortega H, McGonagle D, Jarrett S, *et al.* Infliximab in combination with methotrexate in active ankylosing spondylitis: a clinical and imaging study. Ann Rheum Dis 2005;64:1568-75.
25. van der Linden S, Valkenburg H, Cats A. Evaluation of diagnostic criteria for ankylosing spondylitis. A proposal for modification of the New York criteria. Arthritis Rheum 1984;27:361-8.