Statistical analysis plan for analysis of one-year MRI findings in the AIM trial

Statistical analysis plan for:

One-year MRI findings in patients with chronic low back pain and Modic changes treated with amoxicillin or placebo (the AIM trial)

Publication date: 22.02.2021

Document version: 1.0

EudraCT Number: 2013-004505-14

This document is a supplement to the AIM trial protocol [1] and comprises a statistical analysis plan (SAP) for analyses of one-year MRI findings in the AIM trial. The current SAP was prepared based on guidelines for Statistical analysis plans in clinical trials [2], and after results were available for clinical treatment efficacy in the main trial, in clinical subgroups, and in MRI subgroups.

Scientific board

John-Anker Zwart, professor dr.med
Kjersti Storheim, professor, PT, PhD
Jens Ivar Brox, professor dr.med
Ansgar Espeland, professor dr.med
Margreth Grotle, professor, PT, PhD
Olav Lutro, MD
Anne Froholdt, MD, PhD
Lars Grøvle /Anne Julsrud Haugen, MD, PhD
Audny Anke, professor dr.med
Øystein P Nygaard, professor dr.med
Jan Sture Skouen, professor dr.med
Christian Hellum, MD, PhD
Benedicte A Lie, PhD
Karianne W Gammelsrud, MD, PhD
Thor Einar Holmgaard (patient representative)

Statistical advisor

Jörg Aßmus, PhD

Sponsor

Oslo University Hospital

Coordinating investigator

John-Anker Zwart, professor dr.med
FORMI, OUS Ullevål, Bygg 37B, Postboks 4956 Nydalen, 0424 Oslo
Tel: 22 11 86 21, E-mail: j.a.swarz@medisin.uio.no

Contributors to current SAP

Ansgar Espeland, professor dr.med
Per Martin Kristoffersen, MD, PhD-student
Lars Christian Bråten, MD, PhD
Lars Grøvle, MD, PhD
Margreth Grotle, professor, PT, PhD
Anne Julsrud Haugen, MD, PhD
Jens Ivar Brox, professor dr.med
Nils Vetti, MD, PhD
Christian Hellum, MD, PhD
Mads Peder Rolfsen, MD, PhD-student
Anne Froholdt, MD, PhD
Audny Anke, professor dr.med
Jan Sture Skouen, professor dr.med
Kjersti Storheim, professor, PT, PhD
John-Anker Zwart, professor dr.med
Jörg Aßmus, PhD, statistician

Signature page

Title	**One-year MRI findings in patients with chronic low back pain and Modic changes treated with amoxicillin or placebo (the AIM trial)**
Trial registration:	ClinicalTrials.gov, ID: NCT02323412
EudraCT no:	2013-004505-14

I hereby declare that I have reviewed and approved the statistical analysis plan:

To be signed by person writing the SAP, senior statistician responsible, project manager, and coordinating investigator.

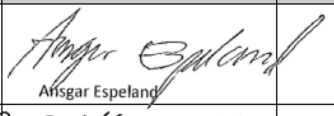
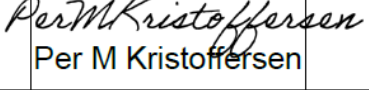
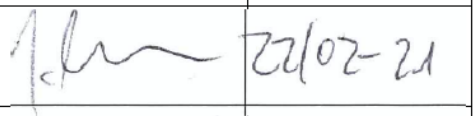
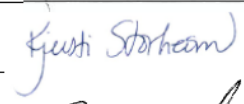
Name	Title	Role	Signature	Date
Ansgar Espeland	Professor, dr.med	Main author of current SAP		18.02.21
Per M Kristoffersen	MD, PhD-student	Contributor to current SAP		19.02.21
Jörg Aßmus	PhD	Senior statistician responsible		
Kjersti Storheim	Professor, PT, PhD	Project manager		10.02.21
John-Anker Zwart	Professor dr.med	Coordinating investigator		10.02.21

Table of contents

Abbreviations

Abbreviation	Explanation
AIM	Antibiotics in Modic changes
BMI	Body mass index
CI	Confidence interval
CSF	Cerebrospinal fluid
ITT	Intention to treat
MCs	Modic changes
MRI	Magnetic resonance imaging
NRS	Numerical rating scale
ODI	Oswestry disability Index
PP	Per protocol
RMDQ	Roland-Morris disability questionnaire
SAP	Statistical analysis plan
SD	Standard deviation
STIR	Short tau inversion recovery
T1/T2	Non-fat suppressed T1- and T2 weighted sequences

Background and rationale

Modic changes (MCs) are MRI signal changes in the vertebral bone marrow extending from the endplate and are classified as type 1 (oedema type), 2 (fatty type) and 3 (sclerotic type, less common) based on standard (non-fat suppressed) T1- and T2 weighted sequences (T1/T2) [3].

The association between MCs and low back pain is inconsistent [4]. Type 1 MCs were associated with pain in some studies [5-10]. Different explanations for MCs are hypothesized [11]. One hypothesis is that low back pain and MCs are caused by a low-grade infection with *Cutibacterium acnes* in the intervertebral disc. The AIM trial, described in more detail below, investigated the effect of amoxicillin versus placebo in patients with chronic low back pain and type 1 or 2 MCs.

The AIM trial found a small, not clinically important effect of amoxicillin on clinical outcome in the type 1 MC group and no effect for type 2 [12]. A larger effect was found in subgroups with abundant MC related bone oedema on baseline short tau inversion recovery (STIR) series [13]. STIR is sensitive to oedema, both in MCs classified as type 1 and type 2, based on non-fat supressed T1/T2 series [14].

In this sub study of AIM, we assess changes in MC oedema from baseline to one-year MRI using variables defined in section “MRI variables”. We hypothesise that reduced oedema is related to better clinical outcome. The rationale is that, if MC oedema contributes to symptoms, the symptoms should improve when the oedema diminishes. Additionally, based on the hypothesis that low-grade infection can cause MCs, we will examine whether amoxicillin reduces MC oedema.

The AIM trial

The AIM trial was a six centre, randomised, parallel-group, placebo-controlled trial on the effects of three months of treatment with Amoxicillin in chronic low back pain patients with type 1 or 2 MCs at the level of a prior lumbar disc herniation (verified on MRI during the preceding 2 years) [1].

Randomisation to either amoxicillin or placebo was stratified by MC type group (type 1/type 2) and prior disc surgery (yes/no) with a 1:1:1:1 allocation and random block sizes of 4 and 6.

At a given endplate, MC types are defined as primary or secondary based on their relative extent. In primary type 1 MCs, type 1 is the most extensive (or only) type. In secondary type 1 MCs, type 1 is present, but type 2 or 3 is more extensive. Patients were classified in the MC type 1 group if they had primary and/or secondary type 1 MCs at a level with previous disc herniation, i.e. type 1 at the endplate superior and/or inferior to the disc. Patients were classified in the MC type 2 group if they had type 2 MCs, but not any type 1 MCs (primary or secondary), at a previously herniated disc level.

MCs with height < 10% of vertebral height or diameter ≤ 5 mm did not lead to inclusion in the trial. These size criteria concern total MC size (all MC types) at the given endplate.

The primary outcome in AIM was pain-related disability measured by the RMDQ (score range 0 to 24) at one-year follow-up [15]. Secondary outcomes included pain-related disability assessed by the ODI [1] and low back pain intensity on a 0-10 NRS. The minimal clinically important between-group difference in mean RMDQ score at one year (treatment effect) was defined as a difference of 4.

In each MC type group, the AIM trial was designed to detect ($\beta = 0.1$, two-sided $\alpha = 0.05$) a difference of 4 (SD 5) in mean RMDQ score between the amoxicillin group and the placebo group at one-year follow-up. This required 66 patients in each MC type group. In total 180 patients were included, 118 in the type 1 and 62 in the type 2 MC group, and 155 completed the trial without major protocol deviations [12]. See the trial protocol for further details on sample size calculation and methods [1].

The treatment effect of amoxicillin was -0.1 RMDQ points in the type 2 MC group and -2.3 RMDQ points in the type 1 MC group [12]. The treatment effect was - 5.1 RMDQ points (95% CI - 8.2 to - 1.9) -but not evident for back pain intensity- in a subgroup ('STIR3') with the most abundant MC related high signal on STIR at baseline [13]. This subgroup is defined in the "MRI variables" section.

Objectives of the present study

The present study concerns patients with chronic low back pain and type 1 or 2 MCs at the level of a previous lumbar disc herniation who were randomized to amoxicillin treatment or placebo. In this study, we will describe changes in MCs on STIR and T1/T2 from baseline to one-year follow-up MRI.

The primary objective (ranked as key secondary objective 5 in the AIM trial protocol [1]) is to assess:

(1) whether change in MC related STIR signal (volume/intensity) is related to RMDQ score at one year

Further objectives (indicated but not ranked in the original AIM trial protocol) are to explore:

(2) the effect of amoxicillin *versus* placebo on change in MC related STIR signal

(3) whether change in MC type 1 volume is related to RMDQ score at one year

(4) the effect of amoxicillin *versus* placebo on change in MC type 1 volume

Hypotheses

The primary hypothesis for this study (ranked as hypothesis seven (G) in the AIM trial protocol [1]) is:

(1) *Reduced MC related high STIR signal (volume/intensity) from baseline to one-year follow-up is associated with lower RMDQ score at one-year follow-up adjusted for baseline RMDQ score*

Further explorative hypotheses (no 3 and 4 indicated in the original AIM trial protocol) are:

(2) Amoxicillin reduces MC related high STIR signal from baseline to one-year follow-up within the STIR3 group (i.e., within the group with the most abundant STIR signal at baseline)

(3) Reduced MC type 1 volume from baseline to one-year follow-up is associated with lower RMDQ score at one year adjusted for baseline RMDQ score

(4) Amoxicillin reduces MC type 1 volume from baseline to one-year follow-up

Statistical principles

This SAP was developed after locking the databases of the AIM clinical trial and the baseline MRI assessments. All papers from the trial published before completing the SAP are referenced here [12; 13; 16-18]. All analyses described herein are considered *a priori* defined either in this SAP or in the AIM trial protocol. Any *post hoc* analyses will be identified as such in the article. All statistical tests will be 2-sided, and the nominal p value will be reported. All CIs presented will be 95% and 2-sided. The assumption of normal distribution will be checked by visual inspection of a QQ-plot.

We will use a Bonferroni corrected significance level of 0.05/7 = 0.007 for hypothesis 1 since it was ranked as the seventh hypothesis in the AIM trial protocol. Note, that the Bonferroni adjustment is conservative since the tests are not independent. This reduces the chance of type I error and strengthens positive findings but could result in type II error. The significance level is otherwise set to 0.05, as advised for explorative 'hypothesis-setting studies' to avoid type II errors [19]. We will not interpret the results based on p values alone but will also consider the size and CIs of the estimates. Several explorative analyses are performed in AIM. It is important not to over-interpret the results.

Analysis populations

MC related high STIR signal cannot diminish if it is lacking. Thus, the association between reduced STIR signal and clinical outcome (hypothesis 1) will be analysed in a sample (**n=162**) consisting of all AIM patients except 16 lacking high STIR signals at baseline and 2 diagnosed with a specific cause of back pain during the follow-up (spondyloarthritis, symptomatic disc herniation requiring surgery).

The effect of amoxicillin in reducing STIR signal (objective 2) will be analysed in a subsample (**n=141**) consisting of all patients in the per protocol (PP) population (all patients completing the trial without major protocol deviations) except 14 PP patients lacking high STIR signals at baseline. Supportive analysis may be performed in the intention to treat (ITT) population (n=180, all randomized patients).

MC type 1 volume cannot diminish in the MC type 2 group without MC type 1. Thus, the association between reduced MC type 1 volume and clinical outcome (hypothesis 3) will be analysed in the MC type 1 group (**n=117** after excluding the disc herniation patient; the spondyloarthritis patient was in the MC type 2 group). The effect of amoxicillin in reducing MC type 1 volume (hypothesis 4) will be analyzed in the MC type 1 PP group (**n=104**). PP analyses can create prognostic differences between

treatment groups [20]. However, when studying the effect of treatment on MRI changes, we regard patients who follow their assigned treatment most relevant, i.e., we consider assessing the effect of amoxicillin more important than assessing the effect of being randomized to amoxicillin.

Major protocol deviations were defined as (a) intake of < 80% of the pills (amoxicillin or placebo), (b) pause of the study medication for ≥ 2 weeks (in the antibiotic group: without other 'relevant' antibiotic treatment in that period), (c) 'relevant' antibiotic treatment in the placebo group for ≥ 4 continuous weeks between baseline and one-year follow-up, and (d) back surgery during the one-year follow-up. 'Relevant' antibiotic treatment was defined as antibiotic treatment likely to affect a C. acne discitis. Further events registered as major protocol deviations were faulty inclusion (2 patients treated with antibiotics last month prior to inclusion), both amoxicillin and placebo given to patient by mistake (1 patient), and spondyloarthritis diagnosed during follow-up (1 patient).

Trial population

The main report of the AIM trial [12] includes a flow diagram detailing enrolment, intervention allocation and follow-up. This study's report(s) will refer to that diagram and, as a minimum, include the number of patients randomised and the number in the PP population for each treatment arm.

MRI variables

This study is based on baseline and one-year follow-up MRI performed using identical protocols and 1.5-T scanners (Siemens Magnetom Avanto B19; Avanto fit E11 was used for 16 follow-up MRIs). Sagittal STIR and sagittal T1/T2 images are used. Three radiologists blinded to treatment group and clinical outcome independently evaluate MRI findings on baseline and one-year images. They score changes in MRI findings from baseline to one-year images by comparing the two sets of images. Baseline and follow-up image slices from the same anatomical location are compared side by side, using all slice pairs (usually 11-12) covering any part of any of the 12 vertebral endplates Th12-S1.

MRI evaluations – overview

MRI findings are evaluated at Th12-S1. However, the MRI variables used in this study concern the index level(s) i.e., the level with type 1 or 2 MCs and a prior disc herniation at baseline. This level was hypothesized to contain the low-grade discitis that was target for the treatment. MCs existed at the endplate superior and/or inferior to the index level disc. One patient can have multiple index levels.

All MRI variables are evaluated separately at each endplate superior and inferior to all index levels. E.g., if two index levels exist, all MRI variables are rated at each of four endplates by each of three observers. However, in analyses for the study objectives, a single conclusive rating per patient for all index level endplates is used for each MRI variable. This conclusive rating is based on the observers' majority or median rating at each endplate. E.g., if the majority rating of MC oedema is 'smaller' at one endplate and otherwise 'unchanged', the conclusive rating across all endplates is 'smaller'. For conclusive rating if the majority rating is both 'smaller' and 'larger' at different endplates, see below.

The STIR and T1/T2 variables evaluated per endplate and per patient are detailed below. All these variables are categorical, not continuous. In brief, the main final conclusive MRI variables per patient are (variable name in **bold**):

- change in MC related STIR signal (**1ySTIR change;** volume/intensity change; objectives 1 and 2)

- change in MC type 1 volume (**1yMC1vol change**; objectives 3 and 4)

All MRI variables used in this study were defined before conducting any of the analyses described in this SAP. The change-variables were defined specifically for the current study, after performing pilot studies to optimize the rating criteria and align the observers' understanding of them. The change-variables will be examined for inter-rater agreement (kappa).

Variables describing MCs and MC related high STIR signal at a given time point (not change) were previously defined and assessed for inter-observer agreement [18]. Mean Fleiss' kappa across L4-S1 (four endplates) for the three observers indicated very good agreement for the presence of any MCs (0.88), the presence of type 1 MCs (0.81) and the presence of high STIR signal (0.86), and moderate agreement for high STIR signal volume (0.56). Agreement on STIR signal volume was better in one observer pair (Cohen's kappa: mean 0.69, range 0.56-0.80 across L4-S1). The conclusive ratings are likely more reliable than each observer's ratings [21].

STIR variables

We define MC related high STIR signal as visible high signal compared to normal vertebral body marrow, formed and located as an MC and/or located in or abutting a region with MC on T1/T2.

STIR evaluations per endplate

The following STIR variables are evaluated separately at each endplate:

(a) change in STIR signal volume from baseline to one year (**endplate 1ySTIRvol change**; smaller, unchanged, larger). Volume change is not measured as a continuous variable but categorized directly into one of the three categories based on visual assessment, as explained below.

(b) change in STIR signal i.e., overall oedema change (intensity and volume combined) from baseline to one year (**endplate 1ySTIR change**; decreased, unchanged, increased). Not measured, but categorized into one of the three categories based on visual assessment, see below

(c) baseline volume score of high STIR signal (**endplate STIR volume**; scored 0-4 based on % affected vertebral body marrow volume; 0% = 0, <10% = 1, <25% = 2, 25-50% = 3, >50% = 4). The volume of the high STIR signal is not measured; it is scored directly into one of the five volume categories by taking into account (summing up) the visually estimated affected area on all images.

(d) one-year volume score of high STIR signal (**endplate 1y STIR volume**; scored 0-4 as above). If STIR volume change in (a) is rated 'unchanged', the score in (d) is set equal to that in (c).

A change in STIR findings in (a-b) is noted only if it is clearly present on ≥ 2 slices. In addition, the change (e.g., 'smaller') must exist on ≥ 2 more slices than any opposite change (e.g., 'larger'). Changed STIR signal volume in (a) does not necessarily imply changed STIR volume score in (d) compared to (c), since clear changes in volume may occur within each volume score category.

The assessment of change in STIR signal in (b) intends to identify cases in which the volume change variable (a) alone clearly does not reflect the overall oedema change. For example, cases in which the high signal region has unchanged volume but clearly reduced intensity, indicating reduced oedema.

On each slice, change in intensity and volume (area) of high signal in (b) is visually assessed within the whole region with high STIR signal at baseline and/or at one year (**Fig 1**). Increased intensity and unchanged or larger volume imply increased STIR signal (oedema) (**Panel 1**). Unchanged intensity and larger volume also imply increased STIR signal. Unchanged intensity and smaller volume denote decreased STIR signal (oedema) (**Fig 1**). Reduced intensity and unchanged or smaller volume also denote decreased STIR signal. In cases of opposite change in intensity vs volume (e.g., reduced vs larger), it is visually judged whether the intensity change clearly 'outweighs' the volume change (or vice versa). If this is not the case, the STIR signal (oedema) is rated as 'unchanged' on the given slice.

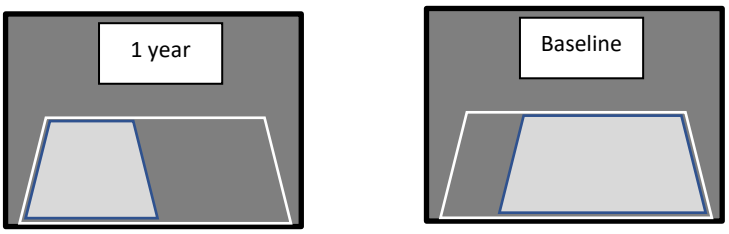


Fig 1 – The figure illustrates corresponding sagittal STIR slices of a vertebral body. Areas brighter than the remaining vertebral body represent high signal. The high signal is usually heterogeneous but illustrated here as homogeneous with average intensity judged visually. At 1 year, the area (volume) of high signal is smaller and the intensity of the high signal is unchanged. The change in STIR signal (overall oedema change) is 'decreased' across the whole region with high STIR signal seen at baseline and/or at one year (marked with a white border).

Panel 1 - Change in STIR signal i.e., overall oedema change (intensity and volume combined) at each endplate and slice described by change in intensity and area (volume) of high STIR signal			
	Volume ↑	Volume →	Volume ↓
Intensity ↑	Increased STIR signal/ oedema	Increased STIR signal/ oedema	Visual judgement of overall oedema change
Intensity →	Increased STIR signal/ oedema	Unchanged STIR signal/ oedema	Decreased STIR signal/ oedema
Intensity ↓	Visual judgement of overall oedema change	Decreased STIR signal/ oedema	Decreased STIR signal/ oedema

Baseline and one-year STIR images are compared, with one-year T1/T2 images open to decide if new STIR signal regions are MC related. STIR signal regions that were MC related at baseline remain MC related at one year. Follow-up and baseline STIR images are evaluated using similar level/window (default 200/400, can be changed when needed if the same change is applied to both image sets, and if a different change is required to make the image sets comparable, e.g., for Avanto vs Avanto fit).

Conclusive STIR findings per patient

For each patient, the three observers' majority (or median) rating of endplate 1ySTIRvol change is determined at each index level endplate. If this majority rating of endplate 1ySTIRvol change is smaller at ≥ 1 endplate and larger at zero endplates, the change in volume of high STIR signal across all index level endplates at one year (**1ySTIRvol change**) is 'smaller'. The 1ySTIRvol change is 'unchanged' if the majority rating of endplate 1ySTIRvol change is unchanged at all endplates, and 'larger' if that majority rating is larger at ≥ 1 endplate and smaller at zero endplates.

If both 'smaller' and 'larger' majority ratings exist at individual endplates, the conclusive 1ySTIRvol change is rated as follows. It is rated as 'smaller' if the total volume of the high STIR signal across all index level endplates is clearly smaller, based on visual assessment and findings on ≥ 2 slices. The 1ySTIRvol change is rated as 'larger' if it is clearly larger. It is otherwise rated as 'unchanged'.

The change in STIR signal (intensity and volume) across all index level endplates at one year (**1ySTIR change**) is determined in the same way, based on majority ratings of endplate 1ySTIR change.

For each patient, we also record the conclusive highest score at any index level endplate for baseline **STIR volume** (**Table 1**) and **1y STIR volume** (**Table 2**).

The conclusive highest baseline **STIR volume** score was reported in a prior study [13], together with a baseline **STIR composite** index level variable with three categories (STIR 1/2/3). STIR3 was defined by MC related high STIR signal fulfilling all of the following criteria: volume ≥25% and height >50% of vertebral body, maximum intensity increase ≥25% (0%, normal vertebral body marrow; 100%, CSF), and presence on both sides of the disc. STIR1 implied MC related STIR signal volume <25% and maximum intensity increase <25%. STIR2 was defined as not STIR3 or STIR1.

T1/T2 variables

We define type 1 MCs as primary or secondary MC types that are clearly hypointense on T1 and hyperintense on T2. Borderline type 1 vs type 2 MCs (near isointense on T1) are rated as type 2.

T1/T2 evaluations per endplate

The following T1/T2 variables are evaluated separately at each endplate:

(e) change in MC type 1 volume from baseline to one year (**endplate 1yMC1vol change**; smaller, unchanged, larger; we record separately endplates lacking type 1 both at baseline and at one year). Change in the volume of the type 1 part of any MC is visually estimated and directly categorized. The evaluation concern type 1 MCs that change type/diminish/disappear and new/increasing type 1 MCs.

(f) change in MC volume (any MC type) from baseline to one year (**endplate 1yMCvol change**; smaller, unchanged, larger). Change in MC volume is visually estimated and categorized.

(g) baseline MC volume score (**endplate MC volume**; scored 0-4 based on % affected vertebral body marrow volume; 0% = 0, <10% = 1, <25% = 2, 25-50% = 3, >50% = 4). Scored into one of the five categories by taking into account (summing up) the visually estimated MC area on all images.

(h) one-year MC volume score (**endplate 1y MC volume**, scored 0-4 as above). If MC volume in (f) is rated as 'unchanged', the score in (h) is set equal to that in (g).

All T1/T2 ratings are completed with only baseline and one-year T1/T2 images open. A change in MC type 1 volume or in MC volume is noted only if it is clearly present on ≥ 2 slices and on ≥ 2 more slices than any opposite change. Changed volume in (f) does not necessarily imply changed volume score in (h) compared to (g), since clear changes in volume may occur within each volume score category.

Conclusive T1/T2 findings per patient

For each patient, the three observers' majority (or median) rating of endplate 1yMC1vol change is determined at each index level endplate. If this majority rating of endplate 1yMC1vol change is smaller at ≥ 1 endplate and larger at zero endplates, the change in MC type 1 volume across all index level endplates at one year (**1yMC1vol change**) is 'smaller'. The 1yMC1vol change is rated as 'unchanged' if the majority rating of endplate 1yMCvol change is unchanged at all endplates, and 'larger' if that majority rating is larger at ≥ 1 endplate and smaller at zero endplates.

If both 'smaller' and 'larger' majority ratings exist at individual endplates, the conclusive 1yMC1vol change is rated as follows. It is rated as 'smaller' if the total MC type 1 volume across all index level endplates is clearly smaller, based on visual assessment and findings on ≥ 2 slices. The 1yMC1vol change across all endplates is rated 'larger' if it is clearly larger. It is otherwise rated as 'unchanged'.

The change in MC volume (any MC type) across all index level endplates at one year (**1yMCvol change**) is determined in the same way, based on majority ratings of endplate 1yMCvol change.

For each patient, we also record the conclusive highest score at any index level endplate for baseline **MC volume** (which was reported in a prior study [13]; **Table 1**) and for **1y MC volume** (**Table 2**).

Description of baseline characteristics

In the main report of the AIM trial [12], patients in each treatment group are described with respect to age, sex, body mass index (BMI), smoking, education level, comorbidity, emotional distress, fear-avoidance beliefs, leg pain, duration of back pain, former disc herniation surgery, physical workload,

compensated work injury or sick leave, concomitant medication use, and index level(s) with both MC and previously herniated disc. Baseline MRI findings are published for each treatment group [13].

In the report of the present study, we will refer to the prior descriptions of baseline data and will, as a minimum, present age, sex, baseline RMDQ score, prior disc herniation surgery, physical workload, BMI, smoking, **MC type group**, index levels, **STIR composite group**, and highest **STIR volume** score and highest **MC volume** score at an index level endplate in the total AIM cohort (**Table 1**). We may also present baseline data for analysed subsamples as a basis for discussing the results.

Categorical variables will be summarised by numbers and percentages, and continuous variables by mean and standard deviation (if normal distribution) or median and interquartile range (if skewed distribution). We will evaluate the clinical importance of any imbalance in baseline data between the treatment groups without statistical testing.

Analyses

MRI findings at one-year follow-up will be reported descriptively for each treatment group (**Table 2**). Interobserver agreement for change in MRI findings will be analysed using kappa statistics.

Association of 1ySTIR change with RMDQ score at one-year (objective 1, **n=162**)

One-year and baseline mean RMDQ scores will be calculated for each of the three categories of 1ySTIR change. Linear regression will be done with one-year **RMDQ** score as dependent variable adjusted for baseline RMDQ score, and with:

(a) **1ySTIR change** as independent variable, dichotomized into 'decreased' or not, as hypothesis 1 indicates that reduced MC oedema is related to lower RMDQ score

(b) **STIR3** (yes or no) added as a second independent variable and potential confounder, since baseline STIR signal abundance might affect both 1ySTIR change and one-year RMDQ score

(c) adjustments also made for age, physical workload, smoking, and BMI (factors known to affect clinical outcome that might perhaps also affect MC oedema)

We will report the unstandardized regression coefficient B with 95% CI and the p value.

- *To evaluate hypothesis 1, we will use the result (B with 95% CI and the p value) for **1ySTIR change** in the adjusted model and will regard p < 0.007 as statistically significant.*

We will also assess whether adding the interaction **STIR3** (yes or no)***1ySTIR change** (decreased or not) improves the model, based on the likelihood ratio test and/or change in R squared.

QQ-plots will be visually checked for normal distribution of residuals.

We may repeat the analysis of **1ySTIR change** with **ODI** and **back pain** scores as dependent variable.

Effect of amoxicillin *versus* placebo on 1ySTIR change (objective 2, **n=141**)

Logistic regression adjusted for prior disc surgery will be done with **treatment group** as independent variable and **1ySTIR change** as dependent variable, dichotomized into 'decreased' or not (i.e., 'unchanged' or 'increased'). Hypothesis 2 is that amoxicillin reduces MC oedema, so both unchanged and increased oedema represent 'treatment failure'. Ordinal regression using all three categories of 1ySTIR change is regarded less relevant in this situation and provides less power. Yet, the frequency of 'increased' in each treatment group will be considered when interpreting the regression results.

To explore the effect of amoxicillin on **1ySTIR change** within the STIR3 subgroup (hypothesis 2), we will include the following independent variables: prior disc surgery, age[a], **STIR3** (yes or no), **treatment group** and the interaction **STIR3** (yes or no)***treatment group**.

[a]Since age (<40 vs ≥40 years) modified the effect of amoxicillin on clinical outcome in explorative analysis [12].

We will perform a pairwise comparison of marginal means (i.e., probabilities of 'decreased' 1ySTIR change) for amoxicillin vs placebo within each STIR3 (yes or no) group. We will report odds ratios, absolute risk reduction, and/or number needed to treat with 95%CIs and p values.

- *To evaluate hypothesis 2, we will use the result for pairwise comparison of marginal means within the **STIR3 (yes) group (n=41)** and the p value from the Interaction test*
- *We will also report results for all analysed patients (**n=141**) and/or for the STIR3 (no) group*

Association of 1yMC1vol change with RMDQ score at one-year (objective 3, **n=117**)

We will calculate one-year and baseline mean RMDQ score for each of the three 1yMC1vol change categories. Linear regression will be done with **RMDQ** score at one year as dependent variable adjusted for baseline RMDQ score, and with:

(a) **1yMC1vol change** as independent variable, dichotomized into 'smaller' or not, as hypothesis 3 indicates that reduced MC oedema is related to lower RMDQ score

(b) adjustments made for age, physical workload, smoking, and BMI.

We will report the unstandardized regression coefficient B with 95% CI and the p value.

- *To evaluate hypothesis 3, we will use the result (B with 95% CI and the p value) for **1yMC1vol change** (smaller or not) in the adjusted model.*

QQ-plots will be visually checked for normal distribution of residuals.

Effect of amoxicillin *versus* placebo on 1yMC1vol change (objective 4, **n=104**)

Logistic regression adjusted for prior disc surgery will be done with **treatment group** as independent and **1yMC1vol change** as dependent variable, dichotomized into 'smaller' or not (i.e., 'unchanged' or 'larger'). Hypothesis 4 indicates that amoxicillin reduces MC oedema, so both unchanged and larger oedema represent 'treatment failure'. Ordinal regression using all three categories of 1yMC1vol change is regarded less relevant in this situation and provides less power. However, the frequency of 'larger' in each treatment group will be considered when interpreting the regression results.

We will report odds ratios and/or number needed to treat with 95%CIs and p values.

- *To assess hypothesis 4, we will use the regression result for **treatment group**.*

Missing data

Of 360 baseline and one-year clinical outcome values for the 180 AIM patients, 13 are missing for RMDQ, 14 for ODI, and 13 for back pain intensity. All relevant MRI data are complete at baseline but missing at one year for 8 (4%) of the 180 patients and for 2 (1%) of 155 PP patients.

Missing values of **1ySTIR change** (decreased or not), **1yMC1vol change** (smaller or not), or of baseline or one-year RMDQ, ODI or back pain intensity (**8 variables**) will be substituted with imputed values from a multiple imputation model (n=180). This model will use 50 imputations and the following **further variables**: RMDQ and back pain intensity at 3/6/9 months, ODI at 3 months, EuroQol's health related quality of life (EQ5D-5L) score at baseline/3/12 months, age, leg pain (0-10 NRS), comorbidity, fear avoidance, emotional distress, physical workload, prior surgery for disc herniation, smoking, BMI, treatment group, **MC type group**, **STIR3** (yes/no), highest index level **STIR volume** score (0-4) and **STIR height** (continuous, STIR signal height relative to vertebral body height), **STIR intensity** (maximum intensity of the STIR signal; <25%, 25-40%, >40%), and largest index level **MC volume** score (1-4) and **MC height** (continuous, MC height relative to vertebral body height). We will also add **STIR3** (yes/no)***1ySTIR change** (decreased or not) and **STIR3** (yes/no)***treatment group** to the imputation model. We will reduce the model if the imputation does not converge.

Table 1 – Baseline characteristics by treatment group

	Amoxicillin group N = 89		Placebo group N = 91	
	n	%	n	%
Age, mean (SD)	44.7 (9.0)		45.2 (9.0)	
Sex, women	53	60	52	57
RMDQ score, mean (SD)	12.7 (4.7), n=88		12.8 (3.7), n=90	
Prior disc herniation surgery, yes	18	20	20	22
MC type group, type 1	58	65	60	66
Body mass index, mean (SD)	26.1 (4.1), n=89		25.9 (4.0), n=90	
Smoking, yes	25	28	21 (of 89)	24
Physical workload				
Mostly sitting	37 (of 77)	48	26 (of 74)	35
Job requires a lot of walking	20 (of 77)	26	20 (of 74)	27
Job requires a lot of walking and lifting	17 (of 77)	22	24 (of 74)	32
Job requires physically heavy work	3 (of 77)	4	4 (of 74)	5
Index level(s)				
L2/L3	2	2.2	2	2.2
L3/L4	7	7.9	5	5.5
L4/L5	48	53.9	29	31.9
L5/S1	58	65.2	74	81.3
STIR composite group				
STIR 1	23	25.8	25	27.5
STIR 2	42	47.2	45	49.5
STIR 3	24	27.0	21	23.1
STIR volume*				
0 (0%)	11	12.4	5	5.5
1 (<10%)	21	23.6	27	29.7
2 (<25%)	31	34.8	36	39.6
3 (25-50%)	21	23.6	18	19.8
4 (>50%)	5	5.6	5	5.5
MC volume*				
1 (<10%)	17	19.1	15	16.5
2 (<25%)	33	37.1	35	38.5
3 (25-50%)	30	33.7	29	31.9
4 (>50%)	9	10.1	12	13.2

SD, standard deviation. RMDQ, Roland-Morris disability questionnaire. IQR, interquartile range. MC, Modic change. STIR, short tau inversion recovery.
* Highest score at an index level endplate (% of vertebral body marrow volume; visually estimated)

Table 2 – One-year index-level MRI findings by treatment group

Variable	Amoxicillin group* (N = xx)		Placebo group* (N = yy)	
	n	%	n	%
1ySTIR change – change in MC related STIR signal (intensity and volume combined) from baseline to one year				
decreased				
unchanged, STIR signal at one year and baseline				
unchanged, no STIR signal at one year or baseline				
increased				
1ySTIRvol change – change in volume of MC related STIR signal from baseline to one year				
smaller				
unchanged, STIR signal at one year and baseline				
unchanged, no STIR signal at one year or baseline				
larger				
Largest 1y STIR volume – highest score at an index level endplate (% of vertebral body marrow volume)				
0 (0%)				
1 (<10%)				
2 (<25%)				
3 (25-50%)				
4 (>50%)				
1yMC1vol change – change in MC type 1 volume from baseline to one year				
smaller				
unchanged, MC type 1 at one year and baseline				
unchanged, no MC type 1 at one year or baseline				
larger				
1yMCvol change – change in MC volume (all types) from baseline to one year				
smaller				
unchanged				
larger				
Largest 1y MC volume – highest score at an index level endplate (% of vertebral body marrow volume)				
0 (0%)				
1 (<10%)				
2 (<25%)				
3 (25-50%)				
4 (>50%)				

MRI, magnetic resonance imaging. STIR, short tau inversion recovery. MC, Modic change.
*One-year MRI is lacking for x of 89 patients in the amoxicillin group and y of 91 patients in the placebo group.

References

1 Storheim K, Espeland A, Grovle L et al (2017) Antibiotic treatment In patients with chronic low back pain and Modic changes (the AIM study): study protocol for a randomised controlled trial. Trials 18:596

2 Gamble C, Krishan A, Stocken D et al (2017) Guidelines for the Content of Statistical Analysis Plans in Clinical Trials. JAMA 318:2337-2343

3 Modic MT, Steinberg PM, Ross JS, Masaryk TJ, Carter JR (1988) Degenerative disk disease: assessment of changes in vertebral body marrow with MR imaging. Radiology 166:193-199

4 Herlin C, Kjaer P, Espeland A et al (2018) Modic changes-Their associations with low back pain and activity limitation: A systematic literature review and meta-analysis. PLoS One 13:e0200677

5 Hanimoglu H, Cevik S, Yilmaz H et al (2019) Effects of Modic Type 1 Changes in the Vertebrae on Low Back Pain. World Neurosurg 121:e426-e432

6 Maatta JH, Karppinen J, Paananen M et al (2016) Refined Phenotyping of Modic Changes: Imaging Biomarkers of Prolonged Severe Low Back Pain and Disability. Medicine (Baltimore) 95:e3495

7 Splendiani A, Bruno F, Marsecano C et al (2019) Modic I changes size increase from supine to standing MRI correlates with increase in pain intensity in standing position: uncovering the "biomechanical stress" and "active discopathy" theories in low back pain. Eur Spine J 28:983-992

8 Mera Y, Teraguchi M, Hashizume H et al (2020) Association between types of Modic changes in the lumbar region and low back pain in a large cohort: the Wakayama spine study. Eur Spine J. 10.1007/s00586-020-06618-x

9 Saukkonen J, Maatta J, Oura P et al (2020) Association Between Modic Changes and Low Back Pain in Middle Age: A Northern Finland Birth Cohort Study. Spine (Phila Pa 1976) 45:1360-1367

10 Brinjikji W, Diehn FE, Jarvik JG et al (2015) MRI Findings of Disc Degeneration are More Prevalent in Adults with Low Back Pain than in Asymptomatic Controls: A Systematic Review and Meta-Analysis. AJNR Am J Neuroradiol 36:2394-2399

11 Albert HB, Kjaer P, Jensen TS, Sorensen JS, Bendix T, Manniche C (2008) Modic changes, possible causes and relation to low back pain. Med Hypotheses 70:361-368

12 Braten LCH, Rolfsen MP, Espeland A et al (2019) Efficacy of antibiotic treatment in patients with chronic low back pain and Modic changes (the AIM study): double blind, randomised, placebo controlled, multicentre trial. BMJ 367:l5654

13 Kristoffersen PM, Braten LCH, Vetti N et al (2020) Oedema on STIR modified the effect of amoxicillin as treatment for chronic low back pain with Modic changes-subgroup analysis of a randomized trial. Eur Radiol. 10.1007/s00330-020-07542-w

14 Finkenstaedt T, Del Grande F, Bolog N et al (2017) Modic Type 1 Changes: Detection Performance of Fat-Suppressed Fluid-Sensitive MRI Sequences. Rofo 190:152 - 160

15 Grotle M, Brox JI, Vollestad NK (2003) Cross-cultural adaptation of the Norwegian versions of the Roland-Morris Disability Questionnaire and the Oswestry Disability Index. J Rehabil Med 35:241-247

16 Braten LCH, Schistad EI, Espeland A et al (2020) Association of Modic change types and their short tau inversion recovery signals with clinical characteristics- a cross sectional study of chronic low back pain patients in the AIM-study. BMC Musculoskelet Disord 21:368

17 Grotle M, Braten LC, Brox JI et al (2020) Cost-utility analysis of antibiotic treatment in patients with chronic low back pain and Modic changes: results from a randomised, placebo-controlled trial in Norway (the AIM study). BMJ Open 10:e035461

18 Kristoffersen PM, Vetti N, Storheim K et al (2020) Short tau inversion recovery MRI of Modic changes: a reliability study. Acta Radiol Open 9:2058460120902402

19 Kent P, Keating JL, Leboeuf-Yde C (2010) Research methods for subgrouping low back pain. BMC Med Res Methodol 10:62

20 Gupta SK (2011) Intention-to-treat concept: A review. Perspect Clin Res 2:109-112

21 Espeland A, Vetti N, Krakenes J (2013) Are two readers more reliable than one? A study of upper neck ligament scoring on magnetic resonance images. BMC Med Imaging 13:4

22 Braten LC (2018) Statistical analysis plan (SAP) for clinical outcomes in the AIM-study: A randomized trial of antibiotic treatment in patients with chronic low back pain and Modic Changes (the AIM-study). ClinicalTrialsgov NCT02323412