Dear dr. Weber, dear editors, dear reviewers Oslo, April 30th 2016

Responses to reviewers and editors

Thank you for offering us the opportunity to revise our manuscript for further consideration by the BMJ. We appreciate the thorough review and helpful comments, which have led to constructive discussions among the authors, and an improved report of the ARCTIC trial. Please find our responses (point-by-point) to the reviewers' comments below, and a revised manuscript, including a version with tracked changes, attached.

Sincerely,

Espen A. Haavardsholm, on behalf of the authors

Manuscript ID BMJ.2016.031902

"Ultrasound in the management of rheumatoid arthritis: ARCTIC randomised controlled strategy trial"

Response to:

Report from The BMJ's manuscript committee meeting

Members of the committee were: Wim Weber (Chair), Rafael Perera (Statistics advisor), Elizabeth Loder, José Merino, Rubin Minhas, George Roeggla, Tiago Villanueva.

Committee comment 1:
We thought your study addresses an interesting and important research question. We had the following queries:

The event rates (remission) are quite low and considerably lower than the planned/anticipated event rates. At those low rate of events it seems that the study was underpowered to demonstrate a clinically important difference. Might you discuss this more extensively?
Response: Thank you for your comment. We are aware that the observed event rate of about 20% in the conventional group is considerably lower than the anticipated rate that was the basis for the sample size calculations. There are probably a number of reasons for this discrepancy, but most importantly this is due to the fact that no study before this trial reported results that could be used directly to estimate the combination of sustained deep DAS remission and inhibition of joint damage in an early RA population. We thus had to choose between a less strict, but easily estimated primary endpoint, or a strict outcome that would be more difficult to estimate for power calculations. The latter was chosen as researchers and clinicians, both nationally and internationally, considered such a combined stringent primary endpoint important if the study should influence future clinical recommendations.

Our choice of two readers to evaluate radiographic progression in known chronological order may have led to lower remission rates, as this method is more sensitive to change than the blinded time order that is applied in industry studies. It has been estimated that this sensitive scoring method may capture up to 50% more progressors compared to a blind reading. We also imputed missing values of the components of the primary endpoint at the end of the study using worst outcome (not meeting the primary endpoint), which is a conservative approach that also lead to lower event rates.

The observed event rate that was considerable lower than the anticipated event rate in the power calculations has implications for determining sample size if the study was to be repeated. In our sample size calculations we aimed for 80% power to detect a 20% difference between the interventions, with estimated rates of 45% and 65% (Protocol). If the study was to be repeated, the power to detect a 20% difference in the primary endpoint from the observed 19% in the control group would have been 89%. Thus, the observed event rate of 19% increases the power, and as such the completed trial was not underpowered to demonstrate a clinically important difference. This can also be deducted from the results

presented in the current manuscript. The estimated treatment difference of the primary endpoint was 3.3% with a 95% confidence interval of -7.1 to 13.7. The confidence interval is completely within the ±20% margin, ruling out a clinically important difference between the treatments according to our estimate of an important clinical effect. Even a stricter choice of a clinically important effect using a 15% margin is outside the confidence interval of the main results.

Action: We have expanded the discussion, and inserted the following: "Although remission rates were excellent, fewer patients than expected in the power calculations reached the strict composite primary outcome. In our sample size estimations, we aimed for 80% power to detect a 20% difference between the interventions (Protocol). If the study was to be repeated, the power to detect a 20% difference in the primary endpoint from 19% in the control group would have been 89%, supporting that the completed trial was not underpowered to demonstrate a clinically important difference. This can also be deducted from the results presented in the current manuscript. The estimated treatment difference of the primary endpoint was 3.3% with a 95% confidence interval of -7.1 to 13.7. The confidence limits of the primary efficacy outcome is completely within the ±20% margin, excluding a clinically significant effect of the intervention (online supplementary, section 7)."

Committee comment 2:

We would appreciate some minor clarifications in significant effects shown for some radiological (secondary outcomes) that are not discussed, as well as your choice of imputation for the primary outcome (missing = deterioration).

Response: No statistically significant effects in radiographic outcomes were seen between the groups, using a p-value of <0.05. The change in total van der Heijde modified Sharp score at 24 months is borderline significant with a p-value of 0.05. This effect is mainly due to a small difference in the erosion score between the groups, but this change in score at 24 months is not significant, with a p-value of 0.06. However, in sensitivity analyses of the same variables in the completer dataset consisting of 204 patients (table S4), a significant difference in radiographic damage over 24 months was found, with a difference in change of van der Heijde modified Sharp score of 0.45 units (95% CI -0.86 to -0.39, p-value 0.03) favouring the ultrasound tight control group.

We are of course cautious when interpreting the results from sensitivity analyses of subgroups, especially the completer data set of an RCT, but an effect of the ultrasound tight control strategy with regard to radiographic progression cannot be ruled out. A possible explanation for this may be the higher proportion of patients receiving biologic therapy, which has been shown to inhibit radiographic progression more than synthetic DMARDs, regardless of disease activity. The ultrasound group did also receive more i.a. steroid injections. The clinical implications of this potential subtle difference are small; as also pointed out in comment 13 from reviewer 3 .

We chose a conservative method to impute components of the primary outcome, but sensitivity analyses in the completer population showed similar results. In addition, we have now performed sensitivity analyses with imputation of best outcome instead of worst outcome, which also resulted in similar results (estimated treatment difference of the primary endpoint was then 3.5% with a 95% confidence interval of -8.6 to 15.6). These analyses strengthen our conclusion.

Action: Further details regarding the radiographic results have been included in table 2, and we have expanded the discussion on the possible implications of the subtle difference in radiographic outcomes:
"Still, there is only a very subtle difference in the progression in the total modified Sharp score between the groups. In a study of early RA patients who were followed over 10 years, a longitudinal association between total modified Sharp scores and HAQ-assessed functional outcome was demonstrated, and an increase of 10 units in the radiographic score was associated with a 0.03 unit increase (worsening) in HAQ score, which has a total range of 0–3 units.[1] Thus, the observed difference of 0.45 modified Sharp score units over 24 months in the current study is not clinically meaningful. The difference is only present in the erosion score, and not the joint space narrowing score, which has been found to be more strongly associated with irreversible loss of function than erosive damage.[2] The observed trend in the erosive score may be due to more frequent initiation of biologic drugs in the ultrasound tight control group, which is known to inhibit radiographic progression independent of disease activity.[3]"
We have updated the results section (added: "Similar results were also found in analyses of the primary endpoint with imputation of best outcome instead of worst outcome (data not shown)."), and included the following in the discussion (on the topic of imputations):
"Sensitivity analyses in the completer population, and analyses in the full analyses set with imputation of best outcome instead of worst outcome resulted in similar results, supporting our conclusion."

Committee comment 3:
The primary outcome initially seemed to be biased against a null as it focuses on clinical (sustained) improvement but we guess that is the most important outcome from a patient's point of view.
Response: We agree that this is an important outcome from a patient's point of view. Please also see our response to reviewer 7, comment 1, regarding the choice of the primary outcome.

Response to:

Reviewer 1

Name: Savia de Souza
Job Title: Honorary Patient Expert in Rheumatology
Institution: King's College London

Reviewer 1 comment 1:
Are the study's aims and the issue and questions that the paper addresses relevant and important to you as a patient? Do you think it would be relevant to other patients like you? What about carers?
I find the objective of both interest and importance. As a person with a chronic illness, I am interested in receiving the best care. If outcomes are better for patients with the routine use of advanced technology (in this case ultrasound imaging), that is important to know and recognise. Any study which addresses ways to improve patient care is of relevance to improving the health of patients and reducing the burden on carers.
Response: Thank you, we appreciate your comment.

Reviewer 1 comment 2:
Are there any areas that you find relevant as a patient or carer that are missing or should be highlighted?
See comment in outcome question below on pain, stiffness and fatigue.
Response: Please see responses and actions below.

Reviewer 1 comment 3:
From your perspective as a patient, would the treatment, intervention studied, or guidance given actually work in practice? Is it feasible? What challenges might patients face that should be considered?
As the trial concluded that integrating ultrasound imaging in routine practice provided no or little added benefit, the answer to this question is of less relevance. The intervention is feasible though in all likelihood patients would not wish to be having numerous ultrasound scans a year and extra injections into their joints based on the results, unless absolutely necessary. For trial design I can see why this is so but if it were implemented in practice, it would no doubt be scaled down.
Response: We agree that some patients might wish to avoid the additional time and possible intra-articular injections associated with the ultrasound examinations. The reviewer correctly points out that the results make this point less relevant.

Reviewer 1 comment 4:
Are the outcomes that are being measured in the study or described in the paper the same as the outcomes that are important to you as a patient? Are there others that should have been considered?
The outcomes measured are important to me as a patient. I would have liked a greater emphasis on measuring pain, stiffness and fatigue as the outcomes seem to primarily be around disease activity scores (which I am aware incorporate these components to some extent) and functional disability.
I believe patients should have been involved with the design of such a large trial as there

appears to be quite a large burden on participants in terms of the number of visits, tests/investigations, assessment scales to complete and the potential for additional intra-articular steroid injections.

Response: We agree that pain, stiffness and fatigue outcomes are important outcome measures to both patients and clinicians, and therefore we have included these measures in our study. We agree that these measures should be further emphasized in the paper. We have included parameters incorporating these aspects in the revised table 1 and table 2.

The ARCTIC study was designed in 2008-2010, and at that time we did not have a patient advisory board. We carried out a number of informal discussions with patients as well as researchers, clinicians, and nurses in the process of designing the study protocol, but today we would have involved patients in the study design in a formalized manner. We do agree that such discussions are especially important when designing large-scale studies with extensive examinations. In general, patients have been positive towards the implementation of ultrasound examinations in the consultations.

Action: Baseline patient's global assessment of pain on a 0 – 100 mm visual analogue scale (Pain VAS) has been included in table 1, and results for 12 and 24 months have been included in table 2. The Rheumatoid Arthritis Impact of Disease (RAID) score is a feasible and patient-derived outcome measure that measures seven domains that are perceived to be of particular importance with regard to quality of life, and the domains included pain, functional disability, fatigue, sleep problems, emotional well-being, physical well-being and coping. The results from the RAID at baseline have been included in table 1, and 12 and 24 months changes have been included in table 2.

Reviewer 1 comment 5:

Do you have any suggestions that might help the author(s) strengthen their paper to make it more useful for doctors to share and discuss with patients?

I feel the ins and outs of the paper would be beyond the comprehension of most patients, however, the abstract provides a nice summary which is easy enough to share with and be understood by patients.

Response: Thank you, we are glad that the abstract is informative to patients.

Response to:

Reviewer 2

Name: Ronald van Vollenhoven
Job Title: Professor of Rheumatology
Institution: Amsterdam Rheumatology and Immunology Center

Reviewer 2 comment 1:
Espen and all, thank you for this excellent study. Having followed it from (almost) the beginning I have been very impressed by both the concept and the execution, you have all done an outstanding job on this. The results were, as I think we all agree, surprising and disappointing. We have discussed this previously and I think part of the reason for the "failure" of the study was the fact that it was trying to do two things at the same time: it wanted to show that ultrasound assessment changes the management (which it did, but not by much) and also that it changes the outcomes (which it did not). I would say that the reason for the second part being negative was that the change in the management was not very big: your data clearly show that the management of the two groups ended up being quite similar. So it is not so strange that the clinical results did not differ much, either.
Therefore, my only recommendation for making this paper even better would be to discuss the fact that, apparently, the clinicians in this study were making the same choices most of the time with or without ultrasound. The question then becomes whether the routine use of ultrasound can be recommended so as to benefit the small minority of patients where it does change the management (and I don't think such a benefit can be ruled out based on your data).
Response: Thank you for your comment, we agree that the finding that treatment decisions were rather similar regardless of ultrasound information is both interesting and to some extent surprising. This is definitely a key message of the study. Nevertheless, we think that our study is answering the research question that was posed: if you follow optimal management of patients following a tight control strategy and adjusting treatment if you do not reach a strict outcome, does ultrasound and the related treatment actions lead to more patients achieving the strict outcome? We show that ultrasound is associated with more treatment (more intra-articular injections and more use of biologics) but not to a higher frequency of strict remission.
We do agree that there may be a benefit to a small minority of patients, but the observed trend in difference between the treatment groups in radiographic score is very subtle, and we doubt that this difference may translate into any clinically meaningful difference in terms of functional outcomes. We have also included some more results regarding the differences in ultrasound outcomes between the groups, and have expanded the discussion; please also see our responses and actions to prof. Smolen (reviewer 3) and dr. Pascal (reviewer 6).
Action:
We have included further results on the differences in radiographic and ultrasound outcomes in table 2, and included a new figure S1 in the supplement. Updated and expanded the discussion of these aspects, please see separate responses/actions to reviewer 3 and 6.

Reviewer 2 comment 2:

If you wanted to explain to the reader why ultrasound is used in the first please, you could also refer to papers showing it helps in several other settings, including diagnosis and procedures.

Response: We agree that additional references and more information regarding the use of ultrasound in other settings would be helpful to the reader, and have included more background information in the introduction. We have also expanded the discussion regarding this topic. Please also see our response to comment 3 from reviewer 3.

Action: We have inserted the following sentences, including new references, in the introduction:

"Ultrasound has been shown to be more sensitive than clinical examination in detecting joint inflammation, and to improve the certainty of a diagnosis of RA.[4-7] Ultrasound may also be helpful in procedures such as aspiration of joint fluid and intra-articular corticosteroid injections.[8 9]"

We have also expanded on the role of ultrasound in the management of RA in the discussion, and revised the last sentence in the conclusion: "There may be an important role for ultrasonography in the diagnosis of RA and in procedures such as intra-articular injections. Future studies should focus on the potential benefit of ultrasound in these areas, as well as the possible role of ultrasonography in evaluating disease activity and tailoring treatment in established RA patients."

Response to:

Reviewer 3

Name: Josef Smolen
Job Title: Professor of Medicine, Chairman
Institution: Dept of Medicine 3, Medical Univ. of Vienna

Reviewer 3 comment 1:
Haavardsholm et al present the results of a very important investigator initiated study in rheumatoid arthritis patients. Employing a tight control treatment strategy with predetermined therapeutic steps, they assessed the value of targeting clinical plus sonographic remission (defined as no power Doppler signal) in comparison to just clinical remission (defined as DAS<1-6 and no swollen joint). The authors should be commended for embarking upon such a well designed and timely study.
The primary endpoint was a combination of the following outcomes: DAS<1.6 and no swollen joint (both for at least the last 8 months of the 2nd year of the study) and no radiographic progression by the van der Heijde Sharp score. The study revealed no significant difference between the groups for the primary and across most secondary endpoints with only a subtle, marginal advantage of the sonographic remission group regarding radiographic progression. The secondary endpoints included the acute phase response, SDAI remission, EULAR good response, physical function and quality of life.
The strength of the study is in the stringency of the target, namely absence of power Doppler signals in all assessed joints as well as absence of any clinically swollen joint for the sonographic group and absence of any clinically swollen joint for the clinical group. From this perspective, one may conclude that this trial provides the ultimate information that targeting sonographic remission is not worth the effort required nor the additional therapies needed.
Response: We thank the reviewer for the positive comments and the summary of the importance of our trial. We agree with his conclusion.

Reviewer 3 comment 2:
Some additional information would be desirable.
Introduction.
1. The authors state that subclinical joint inflammation is present in a majority of patients in clinical remission (2nd paragraph). This statement is based on a selective interpretation of the literature, since several studies revealed that only few patients had subclinical sonographic inflammation when stringent remission criteria were applied (e.g. Sakellariou et al, ARD 2013) and that subclinical inflammation was not associated with radiographic progression (e.g. Gartner et al, ARD 2015). Indeed, the authors themselves prove these latter and invalidate the former studies, since in their clinical remission group more than 60% of the patients had no power Doppler signal in any joint. Thus, the reference to the literature should be addressed more comprehensively and also the results related to the various studies mentioned.
Response: We agree that different studies show conflicting results on the prevalence of true subclinical synovitis. A meta-analysis by Nguyen et al in 2014 based on 19 studies (including the above mentioned reference of Sakellariou et al ARD 2013) provides substantial evidence that ultrasound detected synovitis is frequent and present in the majority of patients in clinical

remission, also with stringent remission criteria (82% overall, ranging from 74% to 86% with different remission criteria).[10] The meta-analysis also concludes that ultrasound findings predict the risk of relapse and structural progression in RA. The original study by Sakellariou et al. assessed wrist and MCP joints in an early RA cohort. The lack of available assessments of the feet (and other joints) in this population will of course lead to lower rates of ultrasound findings. In addition, they do not report findings of grey-scale synovitis.[11]
The study by Gärtner et al. assessed the wrist and finger joints. Even without assessing the feet Gärtner reports that subclinical synovitis (either PD or GS) was present in 1255 of the 1852 clinically asymptomatic joints assessed, i.e. in 68% of joints, which is in line with our statement. We have revised and modified our statement in the introduction and expanded the discussion. The latter now includes a reference to the study by Gärtner, showing that only a small proportion of joints with subclinical activity progresses radiographically.
Action: We have revised the statement in the introduction, which now reads:
"Joint inflammation visualized by ultrasound is present in a majority of RA patients in clinical remission, and several studies have shown that power Doppler activity is associated with radiographic progression and disease flare in these patients.[4 10 12-16]"
In addition, we have added the following to the discussion: "However, a recent study demonstrated that radiographic progression was rare in joints with subclinical inflammation.[17]"

Reviewer 3 comment 3:
2. It may be worthwhile mentioning that sonography is also used for diagnostic reasons, not only for management. This point is also important for the discussion, since the results presented discourage the use of ultrasound for the management of arthritis, but not for diagnostic purposes, such as to evaluate tenosynovitis or joint effusion.
Response and action: We agree, and have updated the introduction, including additional references. We have also expanded the discussion. Please see our response and actions to comment 2 by reviewer 2 on the same matter.

Reviewer 3 comment 4:
Methods
1. Why was the personnel performing the clinical assessments including joint counts unblinded? This could have biased the results in favour of the sonographic group. While this apparently did not matter, since there was no overall difference between the groups, it is still a limitation, which may have to be mentioned in the Discussion.
Response: We agree that the unblinded clinical examination is a limitation. Due to the personnel available at the participating centres, blinded joint examination was not feasible. However, joint counts were performed before the ultrasound examination, to avoid bias in this arm. We have changed the discussion of the limitations of this study as described below.
Action: We have expanded the section on limitations in discussion: "This was an open study, and two components of the primary endpoint, the tender and swollen joint counts, were not blinded. This leaves a potential for bias in treatment decisions based on these parameters."

Reviewer 3 comment 5:
2. Data in supplements are often not easily accessible, especially in the long term. It would be good to see the treatment escalations and their timepoints in a Table, if approved by the

Editor.

Response: We agree, and if approved by the editor we suggest including a simplified version of Table S1 "Treatment regimen in the ARCTIC trial" in the manuscript, as a new figure 1A.

Action: Included a new "**Figure 1A:** Protocol for escalation of disease-modifying anti-rheumatic (DMARD) therapy"

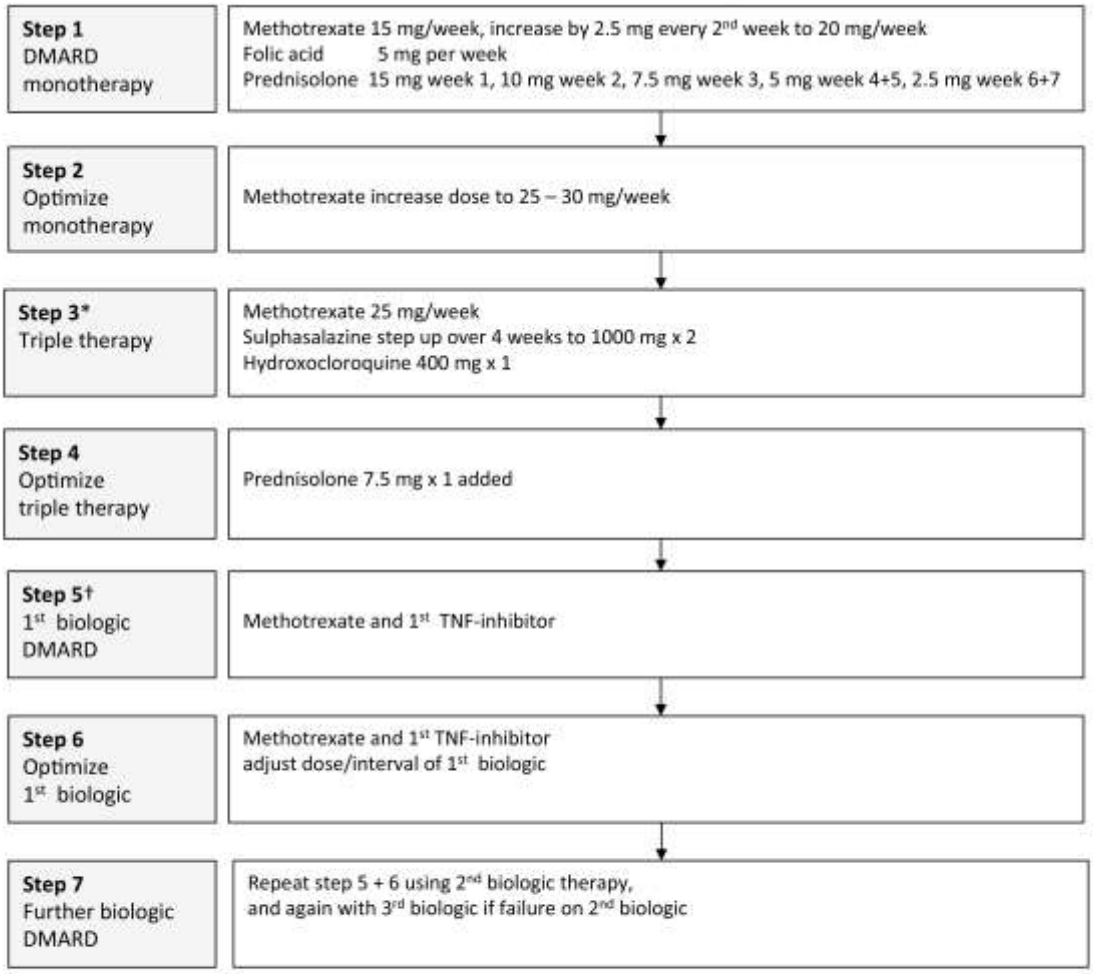


Figure 1A: Protocol for escalation of disease-modifying anti-rheumatic (DMARD) therapy

If response or reached target, current treatment is continued.

* In patients with high disease activity and risk factors for progressive joint destruction a rescue option is available which includes moving to step 5, i.e. introduce 1st biologic

† This step requires signs of ongoing inflammatory activity

Reviewer 3 comment 6:

3. The definition of no radiographic progression is stated as <0.5. With the score used and 2 readers, <0.5 actually means <0. Or is it <0.5?

Response: The reviewer is correct; we used <0.5 as a definition of no radiographic progression. This value can correspond to exactly 0, for example if both readers scored a change of 0, in addition to negative scores if one or both readers scored improvement, as radiographic change may be both positive and negative. If one reader scores a progression of 1 and the other reader scores 0, by this definition the patient will be classified as a progressor. This is a conservative approach, which may also to some extent contribute to the lower event

rate than anticipated for the primary endpoint. We have performed sensitivity analyses including a looser definition of no radiographic progression (cut-off for radiographic progression $\leq$ 0.5 units change in the van der Heijde modified Sharp score), which showed similar results as our primary analysis (estimated treatment difference of the primary endpoint 6.2% with a 95% confidence interval of -5.8 to 18.1, p-value 0.32). The proportion of patients reaching the primary endpoint increased (to 28.6% and 34.8%), and in total 11 patients in the conventional arm and 15 patients in the ultrasound arm had a progression of exactly 0.5 units of the modified Sharp score.

Reviewer 3 comment 7:

4. Why did the authors chose a 20% difference between groups for the primary endpoint to determine the sample size? Was this number a pure assumption or was it based on some published data? At the end of the study, the difference was only 3.3%, suggesting that even a much larger study would not have revealed a significance between the groups. However, this could be emphasized more strongly in the Discussion, since the study may otherwise be criticized for being underpowered.

Response: Thank you for the comment, highlighting the difficulty in predicting study outcomes, even when taking into account previous data. The assumption of a 20% difference between groups for the primary endpoint was to a certain degree based on results from previous strategy trials (as listed in table 1 p. 22 in protocol), as well as results from longitudinal observational studies with ultrasound. However, at this time there were no results from studies that would allow us to accurately estimate the effect of the planned intervention, so there was also an element of what difference would translate to a clinically important effect. Our choice of 20% can be debated, and in two recent equivalence trials of biosimilar drugs in RA, although a very different setting, the equivalence margin was set to ±15%, i.e. this was the estimate of a clinically important effect. The estimated treatment difference of the primary endpoint in the ARCTIC trial was 3.3% with a 95% confidence interval of -7.1 to 13.7. The confidence interval is completely within both the ±20% and ±15% margin, ruling out a clinical important difference between the treatments according to both our estimate of an important clinical effect, and the stricter definition used for assessment of biosimilar drugs Details regarding the rationale for the sample size are outlined in the enclosed full protocol, in section 8.2.2 p. 22 and table 1 same page, and discussed in the online supplement "section 7: Statistical considerations: the conclusion of the ARCTIC study". As outlined here, if the study was to be repeated, the power to detect a 20% difference in the primary endpoint from 19% in the control group would have been 89%, i.e. the power of the current study would be higher than originally planned.

Action: We have expanded the discussion, please see our response and action to committee, comment 1.

Reviewer 3 comment 8:

Results

1. The readers may not be familiar with the 44 joint count nor with the Ritchie index or the DAS. Therefore, the authors should add a 28 joint count for swollen and tender joints and a DAS28 to Table 1.

Response: We agree that in clinical practice and many studies DAS28 is more often used than the original DAS, however we chose DAS as it includes the joints of the feet, which are often involved in early RA. We worry that it might be confusing to readers to also include information of DAS28 scores and separate joint counts for 28 swollen and tender joints, but have included DAS28 scores in table 1 and change scores for DAS28 in table 2.
Action: We have included baseline results of DAS28 in table 1, as well as changes in DAS28 at 12 and 24 months in table 2.

Reviewer 3 comment 9:
2. Table 1 should show ranges and upper limits of normal (e.g. CRP, ESR, ultrasound score)
Response: We agree and have included the ranges for the ESR, the van der Heijde modified Sharp Score and the Ultrasound total score in table 1. We would have liked to add ranges for CRP and upper limits of normal for CRP and ESR, but this will depend on the local laboratory (and for ESR the patient's gender), and inclusion of this information in the table will significantly reduce the readability.
Action: Table 1 has been updated with ranges for the ESR, ultrasound score and the van der Heijde modified Sharp score.

Reviewer 3 comment 10:
3. On p. 11, the CDAI is mentioned – presumably a typographical error?
Response: Thank you, yes, it should be SDAI.
Action: We have corrected CDAI to SDAI.

Reviewer 3 comment 10:
4. It would be good to see the CRP values and ACR-EULAR Boolean remission rates in Table 2. It would also be helpful to see the mean DAS and SDAI values at 12 and 24 months in Table 2.
Response: Thank you for the comment; we agree that this information can be useful to the reader.
Action: We have updated table 2 with CRP values, ACR-EULAR Boolean remission rates, and results of DAS and SDAI at 12 and 24 months.

Reviewer 3 comment 11:
5. It may be worthwhile to show mean radiographic scores over time, since radiographs were obtained at several timepoints.
Response: We agree that this may be of interest; although the changes were very small as pointed out by the reviewer in comment 13. We have now included the median 12-month change scores in table 2 for the total score, the erosion score and the joint space narrowing score, to complement the 24-month change scores that were reported previously.
Action: We have updated table 2 accordingly.

Reviewer 3 comment 12:
Discussion
1. In line with the comment provided in the very first point and the authors' note in the discussion that there was growing evidence on the importance of subclinical inflammation, the contradictory views expressed in the literature should not be concealed.
Response: As pointed out by the reviewer, the literature has evolved somewhat in the years

after the ARCTIC trial was initiated, including studies with stricter definitions of clinical remission. We have updated the discussion accordingly.
Action: We added the following to the discussion: "However, a recent study demonstrated that radiographic progression was rare in joints with subclinical inflammation.[17]"

Reviewer 3 comment 13:
2. The authors mention the subtle difference in radiographic outcomes between the groups. This difference amounts to only 0.45 points on the van der Heijde-Sharp score scale of over 400 points after 2 years of treatment. Is this clinically meaningful? How does this relate to functional impairment? How many years would it take to see this difference manifested in terms of clinical or functional consequences, quality of life or ability to work? Moreover, the difference in radiographic score relates exclusively to the erosion score, while there was no difference between the groups regarding joint space narrowing. All these aspects may deserve being discussed.
Response: We fully agree with the reviewer, and have expanded this section in the discussion including these aspects.
Action: Please see our addition to the discussion, in response and action to comment 2 from the committee.

Reviewer 3 comment 14:
3. Another important piece of information is contained in the manuscript and could be briefly mentioned in the discussion. In this study, patients were started on MTX plus prednisone and after 2 years more than 70% of those in the conventional and more than 50% of those in the ultrasound group were still on MTX monotherapy and presumably in remission according to the respective definition (otherwise they would have escalated therapy). Moreover, despite a treatment strategy targeted at remission, only 17% of the patients in the conventional group required a biologic agent. These data reveal the power of both a treatment approach targeting remission and a start with MTX plus prednisone.
Response: We agree that this is an important lesson from the ARCTIC trial, with clinical implications. We have expanded the discussion.
Action: We inserted the following in the discussion: « Our results showcase the power of both a treatment approach targeting deep clinical remission, follow-up with tight control and initiation of treatment with methotrexate with prednisolone, combined with i.a. injections in swollen joints. After 24 months, more than 70% of patients in the conventional arm were still on methotrexate monotherapy and only 17% required a biologic agent.»

Response to:

Reviewer 4

Name: Michael Gill
Job Title: Mr
Institution: Patient Reviewer

Reviewer 5 comment 1:
This research study seems to be well conducted with a conclusion that is entirely reasonable. On closer consideration from a patient's perspective, the approach taken by the researchers does suggest a number of concerns.
Participant selection was based on age range, classification criteria for RA, no commencement of DMARD therapy and less than two years since diagnosis. Part of the usual classification of a patient's symptoms requires a swollen joint count. A proportion of patients do not have sustained swollen joints, instead they may demonstrate swelling for short periods of time, swelling that rotates across the body or swelling during flare episodes. As such, I believe, that excluding such patients could be considered a weakness.
Response: Thank you for your comment, which addresses an important issue, the selection of patients to be included in the trial. We chose to include patients based on a diagnosis of RA and fulfilment of the 2010 ACR/EULAR classification criteria, which requires at least one clinically swollen joint. In addition, the clinician should explicitly confirm that the patient was in a condition that required initiation of DMARD therapy. We believe this is a reasonable inclusion criterion, as the study protocol implies aggressive therapy with DMARDs and intra-articular steroid injections, as well as subsequently biologic therapy if the treatment target of remission is not reached. Although we agree that some patients might have periods without swollen joints during the disease course, but inclusion of patients with no current signs of arthritis into a predefined aggressive treatment protocol would cause potential harmful over-treatment. Considering the risk/benefit of the treatment we strongly believe that inclusion of patients without swollen joints is not justified.

Reviewer 4 comment 2:
The endpoint definition for patients between 16 and 24 months of disease activity was defined as clinical remission, no swollen joints and non-progression of radiographic joint damage. Other secondary outcomes were also considered. Having read numerous research papers on clinical remission and attempts to define and quantify the state, I remain suspicious that reductionism is triumphing over what patients feel. In other words irrespective of the measured definition, if the patient still feels unwell, is still unable to live as they did prior to RA then remission has not been achieved. As such I suggest that broad claims "that remission has become achievable" (p4) would be better described as "a large proportion of RA patients reach improved quality of life today than was historically the case due to tight control strategies and new therapies."
Response: We agree with the reviewer that while remission is often defined as total absence of signs and symptoms of disease activity, the disease activity cut-offs of remission may rather correspond to a state of low disease activity, "near-remission" or "partial remission", where patients still may experience pain and feel unwell. The concept of remission in RA implies absence of disease, i.e. no signs and symptoms of active disease, but is not the same

as "cure", which implies that the disease process will not return. Remission definitions based on composite disease activity indices have gained widespread use during the past 10-15 years.

With improved treatment strategies and advanced therapy, the need for consensus on how to define remission is increasingly pertinent. During the last years the European and American rheumatology organisations have collaborated to re-evaluate the concept of remission with the aim to reach consensus on a new, stringent definition of remission. For example, the ACR/EULAR Boolean remission criteria require a patient global score of maximum 1 on a scale from 0 to 10. With the new stringent remission criteria, the concept of remission is closer to the patient definition of remission. Remission is now a defined treatment target in early RA in both European and American treatment recommendations, and a number of studies during the last few years have confirmed that this is an achievable goal. We have assessed all these outcomes in our trial and also many patient oriented outcome measures, which all confirm the findings of the primary outcome. We refer to the response to reviewer 1 on RAID, which is a fully patient reported outcome developed with input by patients. The RAID score has now been added to the paper.

Reviewer 4 comment 3:
The survey design based on 11 centres in Norway does seem limited on two grounds: sample size and cultural peculiarity. Not all health care delivery jurisdictions in other parts of the world would be dominated by university hospitals. In a number of jurisdictions familiar to this writer, private practice and community health centre support for patients with RA are the dominant form of delivery.
Response: We agree that the organization of health care differs between countries, potentially influencing also the results of the current study. Private practice and community health centres are rare within Norwegian rheumatology, but we were able to recruit one private practice centre. Most patients receive care from either a local hospital or a university hospital. In the current trial 4 out of 11 centres were university hospitals, but in Norway these hospitals also to a large extent have local hospital responsibilities due to the low population density. Thus, the large majority of patients with RA, and of the patients participating in ARCTIC, are taken care of in a secondary health care situation, not a tertiary care center.

Reviewer 4 comment 4:
Radiographic images of just three joint areas do appear to be somewhat limited. Imagery of the shoulder joints should, I believe, have been included for improved analysis sensitivity.
Response: The ARCTIC trial included radiographs of wrists and hands as well as the feet. The commonly used and most sensitive scoring systems for radiographs in RA include only these joint areas. The reason is that it has been shown that damage in the hands, wrists and feet shows most progression and is highly correlated with damage in the large joints. To limit ionizing radiation, systematic imaging of the large joints in patients with RA is not recommended. However, imaging of large joints was performed when clinically indicated.

Reviewer 4 comment 5:
It was disappointing to see that no discussion was evident on the impact of ultrasound examination itself on patient motivation and levels of concern.
Response: Thank you, this is an interesting comment. The ultrasound joint examination is

time consuming, but also creates a situation where the patient can directly see what is happening within the joints. Most patients were grateful for this additional information. Ultrasound findings may be important when discussing escalating therapy, and visualizing the inflammatory process within the joint may improve patient motivation to adhere to therapy. If no sign of disease activity is seen on ultrasound, it may reassure the patient that the current therapy is working. We have included aspects of this in the discussion.
Action: The following was added to the discussion: "The result of each ultrasound joint assessment in the ultrasound tight control arm was communicated to the patient, and despite the extra time and effort required, most patients appreciated the opportunity to directly observe the level of inflammation inside the joints. This may contribute to increased patient adherence to therapy, in that ongoing inflammation would improve patient motivation to escalate treatment, and resolution of inflammation might reassure the patient that the current therapy was effective."

Reviewer 4 comment 6:
Assessing pain and fatigue are not simple matters. The interplay between adequate sleep, mental health and shear trauma of the RA experience have not been considered except by the ACR core set and VAS. It would have been useful to understand the relationship between levels of pain and positive ultrasound images. The reliance on global measure like DAS and ACR and the lack of patient involvement in the development and design of this study is, perhaps, the main shortcoming.
Response: In line with this and comments from the other reviewers, we have included additional results in table 2, including RAID and measures of pain in addition to the data reported previously. The comment addresses an important question, and we agree that these aspects of RA are of great interest. Although we find it difficult to address this fully within the current manuscript, we will go further into these aspects in future reports from the ARCTIC trial. Please see the response to reviewer 1 for details on patient involvement in the development and design of this study,
Action: We have included RAID scores in table 2 and VAS pain in table 1 and table 2.

Reviewer 4 comment 7:
This is an interesting and worthwhile study.
Response: Thank you.

Response to:

Reviewer 5

Response to: Reviewer 5

Name: Johannes Bijlsma
Job Title: Professor of rheumatology
Institution: University Medical Center Utrecht, NL

Reviewer 5 comment 1:
This is a timely study, evaluating the possible additional value of adding imaging in the decision process on tight control in treatment of patients with early RA. The results are quite clear, not only in showing that there is no additional value, but also resulting in a caveat re increasing perhaps unnecessary additional treatment. The border between under-treatment and over-treatment becomes smaller in sophisticate tight control strategies; this is an important to realise.
The study is well performed, and the authors were clever to design an acceptable concealing strategy to prevent bias in the evaluation of the primary outcome. Power calculations and analysis plan are up to date and adequate.
Response: Thank you for your comments.

Reviewer 5 comment 2:
I have one major question with regard to the actual performance of the study. In many strategy studies treating physicians have clear rules how they should come to their treatment decisions, but they may always deviate when they are convinced that patients interest prevails. E.g. in the well-known BEST study this has been evaluated as having happened in 25% of the decision-moments. In other studies this might be lower. The authors of the present study do not mention this problem at all. I wondered whether this indeed didn't happen? Please comment and evaluate.
Response: Thank you for your comment. We agree that this is an important aspect in assessing the treatment strategy. There are limitations to algorithm-driven treatment decisions, for example in patients with co-morbidities or adverse events/reactions where the clinician will take into account factors that are not part of the treatment strategy. Our results indicate that 18.8% of treatment changes in the ARCTIC trial deviated from the decisions rule, which is in line with what is reported from the BEST study.
Action: The discussion has been updated with the following paragraph: "Patient and physician adherence to a pre-specified treatment protocol targeting remission can be challenging, and in a number of clinical situations rheumatologists may be reluctant to base treatment decisions solely on an algorithm rather than integrating all available information. A recent report from the BeST study evaluated rheumatologists' adherence to a DAS-steered treat-to-target strategy, and in average the protocol adherence was 79%.[18] In the current trial, we found that 19% of treatment changes deviated from the decision rules of the treatment algorithm."

Reviewer 5 comment 3:

In addition, were the sonographers trained, preferably repeatedly during the trial, and aligned in order to get as good as possible the same cut off points in their judgement?
Response: This is an important aspect of an ultrasonography trial. The sonographers were extensively trained and calibrated before the study, and ultrasound training was repeated once yearly during the conduct of the study.
Action: The methods section has been updated with the following sentences: «All the sonographers participating in the study were experienced and underwent extensive training with both static and dynamic hands-on exercises to calibrate readers before the inclusion of the first patient, and an ultrasound workshop to ensure calibration was repeated annually during the conduct of the data collection. A published atlas with ultrasound images showing the range of scores of both power Doppler and grey-scale synovitis in the assessed joints was available at all study centres for reference.[19]»

Reviewer 5 comment 4:
Minor remark: p 6, line 42 DAS score: is this DAS 28 or 44, ESR or CRP ?
Response: The original DAS score, based on 44 joints and the Ritchie Articular Index, was used throughout the study. We agree that this was not sufficiently clear.
Action: We have changed this section, which now reads: «The DAS score (range 0 to 10, higher score indicating more disease activity) is a composite index of four variables: the number of swollen joints among 44 examined joints, the Ritchie Articular Index with a graded assessment 0 to 3 of the tenderness in 26 joint regions, the erythrocyte sedimentation rate (ESR) and the patient's global assessment on a visual-analogue scale (VAS) ranging from 0 to 100.[20]»

Response to:

Reviewer 6

Name: Zufferey Pascal
Job Title: PD
Institution: Rheumatology unit /DAL / CHUV / Lausanne, Switzerland

Reviewer 6 comment 1:
General considerations:
This article is interesting and original since it is one of the rare published studies performed in multiple centers using a strict computer based tight control design for the monitoring of RA treatments. It is also as the authors mentioned one of the first study using, in one arm ultrasound as a tool to monitor therapy.
I have no special remarks regarding the design of the study, the selection of the patients, the number of patients included, the schedule of the visits and the treatment algorithm, the clinical and radiological targets used to assess the response to treatment. They have all been validated in previous studies. The high amount of steroid infiltrations can however be questioned.
Response: Thank you for your comment. We agree that the aggressive steroid injection probably is more common in Scandinavian countries than in the rest of the world. Data from the CIMESTRA trial (and later the OPERA trial) has shown this approach to be very effective,[21] but also in the TICORA trial patients in the tight control arm received significantly more steroid injections than in the comparison arm.[22]

Reviewer 6 comment 2:
I just regret that the authors have not included a third arm for which treatment adaptations would have been made according to physician decision in collaboration with the patient. Even if previous studies have already shown that a stringent tight control algorithm performed better in term of remission than the classical physician based approach, in clinical practice, decisions made by a computer without any discussion and previous acceptance from the patients does not seem really applicable and even desirable. Moreover clinical and US results of this third arm would have been very interesting. This point could be discussed by the author.
Response: We agree that the suggested addition to the trial design would add information about the outcome of less stringent treatment. However, the additional value of this arm would be limited by the convincing results from the TICORA trial, and it might even be argued that such an arm would be unethical due to the large differences in patient outcomes across many domains that were reported[22].

Reviewer 6 comment 3:
As a specialist of the use of US in RA, I have some reservations concerning the ultrasound component of the study. The author stated that they used a validated score but in fact they (reference 31) only showed that the score was reliable in terms of inter-reader performances. To my knowledge, this score has never been evaluated for its sensitivity to change and its performances compared to clinical disease activity measurements such as the DAS.

Response: The US score used in this study was developed in close cooperation with dr. Hilde Berner Hammer, an internationally recognized specialist in musculoskeletal ultrasound. This ultrasound score was based on previous work starting with a comprehensive 78-joint score, which in a series of exercises was reduced to yield the current US score. The sensitivity to change of the full 78-joint score and several reduced scores have been published, as well as the performance of this approach compared to clinical assessments[23]. We agree that the description of the score could be more precise, and have changed the manuscript accordingly.
Action: The sentence "…, according to a validated scoring system of 32 joints" has been changed to "…, according to a scoring system of 32 joints with high intra- and inter-rater reliability."

Reviewer 6 comment 4:
Moreover there is no explanation for the US changes chosen (10%, 20% of the total score) to decide treatment rules in the US arm. Are those criteria clinically relevant? Why choosing total B and Doppler score for treatment adaptations and just no Doppler as the final target? DAS relevant changes (>0.6, >1.2) have been previously validated so the US rules should also be validated.
Response: We agree with the reviewer that this information should be provided. We used a data-driven approach to derive cut-offs for the US change score, by assessing the magnitude of changes in US-score corresponding to relevant DAS changes.[24] These analyses were performed in a previously collected data set using the same US score. The chosen cut-offs of 10% and 20% US change scores corresponded to DAS changes of >0.6 and >1.2, respectively.
Action: The following sentence has been added to the methods section: "A data-driven approach was applied to derive cut-offs for the US change score in a previously collected data set using the same US score, by assessing the magnitude of changes in US-score corresponding to relevant DAS changes. The chosen cut-offs of 10% and 20% US change scores corresponded to DAS changes of >0.6 and >1.2, respectively."

Reviewer 6 comment 5:
Why the authors did chose only clinical and radiological outcomes and no ultrasound outcomes either primary or secondary, although one of the final targets in the US group was no Doppler activity? The design of study has nevertheless included US evaluation in both groups. Some of these US outcomes were expressed and compared between the two groups at baseline, during and at the end of follow-up. In contrary to clinical and radiographic outcomes, they are significantly different between the two groups after treatment adaptation (table1 and 2).For me, this point is not adequately discussed.
Response: A number of ultrasound parameters were included as secondary outcomes. We agree with the reviewer that these results are of importance, and have added the ultrasound total score at 12 and 24 months in table 2. After 12, and to a lesser extent after 24 months there were significant differences in the grey-scale and power Doppler scores in favour of the ultrasound tight control group. As with the radiographic differences, these are very small and it is questionable if this is clinically relevant. Moreover, this is an expected finding, as in contrast to the conventional tight control group, in the ultrasound tight control group treatment should be intensified if joints showed a power Doppler signal.

Please also see our responses to comments below.

Action: Table 2 has been updated to include ultrasound scores, both total and individual scores for power Doppler and grey-scale at 12 and 24 months.

Reviewer 6 comment 6:
According to previous studies mentioned in the introduction by the authors, the better Doppler results in the US arm could suggest that the remission obtained based on US evaluation is more robust and could led to less relapses. The conclusion of the article should therefore be a little bit modulated, not rejecting so straight fully ultrasound as an additional tool for treatment tailoring in RA.
Response: We agree that this is an interesting hypothesis, and we are currently conducting a follow-up study of the ARCTIC trial that would allow us to assess this at a later stage. However, the results for the primary outcome and the secondary outcomes are convincing with regard to the effect during the first two years of such a strategy. Based on the support to our conclusion from other reviewers we have not made any changes to the first part of the conclusion, but we have revised the last part of the conclusion.
Action: Revised the last sentence of the manuscript: "There may be an important role for ultrasonography in the diagnosis of RA and in procedures such as intra-articular injections. Future studies should focus on the potential benefit of ultrasound in these areas, as well as the possible role of ultrasonography in evaluating disease activity and tailoring treatment in established RA patients"

Reviewer 6 comment 7:
Some comment on the US operators and machines used could also be useful if one really wants to appreciate how the results can be applied to clinical practice and to other countries: how many different operators, how many different machines, which ones, repartition and balance of operators in the different centers.etc).
Response: We agree, and have included information regarding the machines and settings. All the sonographers are listed in the supplement, section 1.
Action: The following sentence has been added to the methods section: «The ultrasound assessments were performed with Siemens Antares or GE Logiq E9 ultrasound machines with linear probes (11.4/13.0 MHz). Power Doppler parameters were adjusted according to the device used (pulse repetition frequency 391/600Hz; Doppler frequency 7.3/10.0MHz).[25] There were no changes in ultrasound settings during the study, and no upgrading of software.»

Reviewer 6 comment 8:
P4: Line 35 : I don't think that a majority of patient in clinical remission have US signs of activity it is more around 30 to 50%.
Response: We agree that background on subclinical inflammation is of interest in this paper, and that we could provide further details. A meta-analysis by Nguyen et al in 2014 based on 19 studies concludes that ultrasound-detected synovitis is frequent and present in the majority of patients in clinical remission (82% overall, ranging from 74% to 86% with different remission criteria). Please also see our response to reviewer 3 regarding the same topic.
Action: We have revised the statement in the introduction, which now reads:
"Joint inflammation visualized by ultrasound is present in a majority of RA patients in clinical remission, and several studies have shown that power Doppler activity is associated with radiographic progression and disease flare in these patients.[4 10 12-16]"

Reviewer 6 comment 9:
P4: Line 40 : imaging remission discussed here and not taken into account in the paper although you have some US data collected for that. (no Doppler activity in the Us group and even data to compare in the non US group).
Response: In the introduction we explain the rationale for conducting the current study, and at the time when the protocol was developed such data were not available.[26]

Reviewer 6 comment 10:
Participants:
P5 line 46: consent : was it clear that that the patient had to accept and follow strictly the computer defined rules.
Response: The inclusion criteria for the study have been provided in the online supplement, and item 7 states that patients should be "able and willing to give written informed consent and comply with the requirements of the study protocol". It was explained in the patient information that it was the intention to follow the treatment strategy driven by achieving/not achieving the target. However, there were also a number of exceptions when the treatment strategy did not need to be followed, for example due to adverse events, co-morbidity or other factors.
Action: Please see response/actions to reviewer 5 concerning the number of treatment decisions not adhering to the "decision rules".

Reviewer 6 comment 11:
Randomization:
P6 line 6 : personel: Who did the US evaluation in the US group, was it the same physician or a different one?
Response and action: The US evaluation in the US group was performed by the treating physician, and we have clarified this in the manuscript: "In the ultrasound strategy arm, the sonographer was also the treating physician, and patients were informed of the US results."

Reviewer 6 comment 12:
How many physicians how many ultrasonographers took part to the study?
Response: The investigators of the ARCTIC trial are listed in section 1 in the online supplement, specified for each arm. All investigators in the ultrasound tight control group were sonographers. A total of 27 sonographers (all physicians) were involved in the study, and in the conventional tight control arm there were 25 physicians.

Reviewer 6 comment 13:
Were patients randomized adequately in each center or did some centers evaluate most patient of US group and some others most of patients of the clinical arms since these factors can have a great impact on results especially if no prior standardization was performed?
Response and action: The randomization was stratified for study centre (p. 5 line 60 and p. 6 line 1).
There was a good balance in the allocation of the two strategies in all centres. There was extensive training and calibration of ultrasonographers prior to the study, as well as during the

conduct of the study. The methods section has been updated with information on the prior standardization, see also response and action to reviewer 5, comment 3.

Reviewer 6 comment 14:
Assessments
P6: Line 17: validated score, see above remarks
Response: Please see above response with description of the modifications to the manuscript.

Reviewer 6 comment 15:
P6 line 30: blinded to the results, what about the US group blinded or not to US?
Response: The US group was not blinded to the US score. See also response to comment 11 above.
Action: Added the following under the "assessments" subheading: "In the ultrasound strategy arm, the sonographer was also the treating physician, and patients were informed of the US results."

Reviewer 6 comment 16:
Treatments strategy
P6 line 52 : methotrexate , oral or subcutaneous?
Response: The initial therapy was oral methotrexate, but patients could be switched to subcutaneous methotrexate. The treatment regimen in the ARCTIC trial is provided in online supplement table S1, but we acknowledge that many readers will be interested in further details than what was previously provided in the main manuscript.
Action: Please also see response and action to comment from reviewer 3. We have added a new figure 1 A.

Reviewer 6 comment 17:
Outcomes
P7, Line 46 , why so many early X-ray evaluations in patients with only moderate disease activity ?
Response: There are several studies indicating that radiographic progression may occur early in the disease, and that X-rays evaluated using the van der Heijde modified Sharp score is sensitive to change over periods as short as 3 months.

Reviewer 6 comment 18:
Results:
P10, table1: it would be interesting to know the % of patients with Doppler activity at baseline in both groups in order to compare them with follow-up data (table 2)
Response: We agree, and we have included information on the proportion of patients in each group that has power Doppler signal in any joint.

Action: We have added this information to table 1, and details are now also available in the new Figure S1. Please also see our response and actions to reviewer 7, comment 3 regarding this subject.

Reviewer 6 comment 19:
P10 , line 53, why 8 and not 6 months ?

Response: This was to accommodate the tight control strategy, in which treatment response was assessed every 2 months for the first year, then every 4 months. Thus, for feasibility and practical reasons 8 months was chosen instead of 6 months. In the definition of a "complete clinical response" defined by the FDA (Food and Drug Administration. Guidance for industry. Clinical development programs for drugs, devices and biological products for the treatment of rheumatoid arthritis. US Department of Health and Human Services, FDA, Feb 1999), the requirement of radiographic arrest is at least 6 months, thus 8 months was found to be an acceptable time frame.

Reviewer 6 comment 20:
P12, table 2: Doppler results are significantly different between the two groups in contrary of baseline data (see above)
Gray scale and total score results during follow-up like in the baseline data would have been interesting. to know. Are they also significantly different between the two groups?
Response and action: We agree that this is of interest and have provided the data requested in table 2. There are significant differences in most of the parameters, as expected in favour of the ultrasound group.

Response to:

Reviewer 7

Name: Kei Ikeda
Job Title: Assistant Professor
Institution: Chiba University Hospital

Reviewer 7 comment 1:
The Authors investigated the additional benefit of performing ultrasound over conventional treat-to-target approach in the treatment of DMARDs-naïve patients with rheumatoid arthritis (RA). As the Authors argue, this is a randomized trial that is necessary to determine whether ultrasound imaging remission should be used as a treatment target of RA in routine practice. However, the primary endpoint is not optimal and the Reviewer does not agree with the Authors' interpretation of the results.

1. Two out of three primary endpoint components are identical to the treatment target of the conventional arm (i.e. no swollen joints, DAS remission). Because the improvement in ultrasound-detected synovitis is often not reflected by swollen joint count or DAS, these two components obviously contribute to the negative result when treatment was adjusted to meet these components in both arms. The primary endpoint should have been no radiographic progression, which is independent of treatment targets in both arms, and the study seems to be underpowered to detect the difference in this component. In fact, the statistically significant differences in ultrasound outcomes and the trends towards radiographic benefit in ultrasound group are consistent with what have been implied in the previous studies supporting the use of ultrasound. The Authors' arguments should be revised according to these points.
Response: We would like to emphasize that the clinical targets were identical in both treatment arms, with the addition of the ultrasonography target in the ultrasound arm. We understand your concern regarding the choice of endpoint, which was one of the main discussion points when designing this study. We performed a series of meetings and discussed these matters with clinicians and researchers before initiating the study, both nationally and internationally, to be able to understand what evidence would be needed to convince rheumatologists that a treatment target of imaging remission would be better than a target based on conventional assessments. A large majority stated that an endpoint of clinical remission, preferably sustained remission, combined with no radiographic progression would provide the necessary evidence to change clinical practice. We still think that this is true. The primary endpoint in the ARCTIC trial was based on the original US Food and Drug Administration definition of a "Complete clinical response", which was defined as sustained remission combined with radiographic arrest for at least 6 months (Food and Drug Administration. Guidance for industry. Clinical development programs for drugs, devices and biological products for the treatment of rheumatoid arthritis. US Department of Health and Human Services, FDA, Feb 1999).

Reviewer 7 comment 2:

2. The treatment target of the conventional arm (i.e. no swollen joints, DAS remission) is too strict for the study to reflect real-world practice and to demonstrate the benefit of performing ultrasound in daily practice. Rheumatologists do not always escalate treatment for a single swollen joint or a minor flare just above DAS 1.6 as they are aware that these clinical manifestations can be a false positive. Ultrasound plays an important role in decision making at such an equivocal disease activity. The very strict clinical treatment target that was applied to both arms is not likely to have allowed the study to demonstrate this role of ultrasound. This point should be discussed in the manuscript.

Response: We do agree that the choice of treatment target is important in a treat-to-target trial. According to the international Treat-to-target recommendations and recent EULAR and ACR treatment recommendations the treatment goal in early RA should be remission, and abrogation of inflammation is necessary to obtain this measure. In the current study our aim was to establish if ultrasound guided treatment would contribute to a better outcome over and above what could be achieved by conventional treatment adhering to good clinical practice based on available knowledge. We agree that it is possible that ultrasound guided treatment might be more important if tight control is not applied, but such a design would not answer our research question.

The treatment goal of no swollen joints has been an important part of regular clinical practice in the Nordic countries for many years, and even the CIMESTRA protocol, developed in 1998, states that this is an explicit goal in early RA[21]. We acknowledge that subtle changes in the DAS may be due to many factors, and that this not always will translate into a clinical decision of changing treatment. This was part of the treatment strategy, in that clinicians were allowed to take into account other factors, e.g. adverse events and co-morbidity. To escalate treatment to biologic DMARDs it was explicitly stated in the protocol that it should be evidence of ongoing inflammatory activity. We have added a new figure 1A to the main manuscript, making this part of the treatment algorithm more accessible to the general reader.

Action: Added a new figure 1A to the main manuscript.

Reviewer 7 comment 3:

3. Page 16, 2nd line, "inflammation assessed by ultrasound was suppressed to a minimum in both arms". 38.4 % of patients with positive power Doppler signal is not "minimum". Please revise the sentence. The Reviewer considers that the statistically significant difference in the positivity of power Doppler signal between two arms is one of the novel and significant findings of this study. Therefore, the time course of power Doppler positivity should be added to Figure 2.

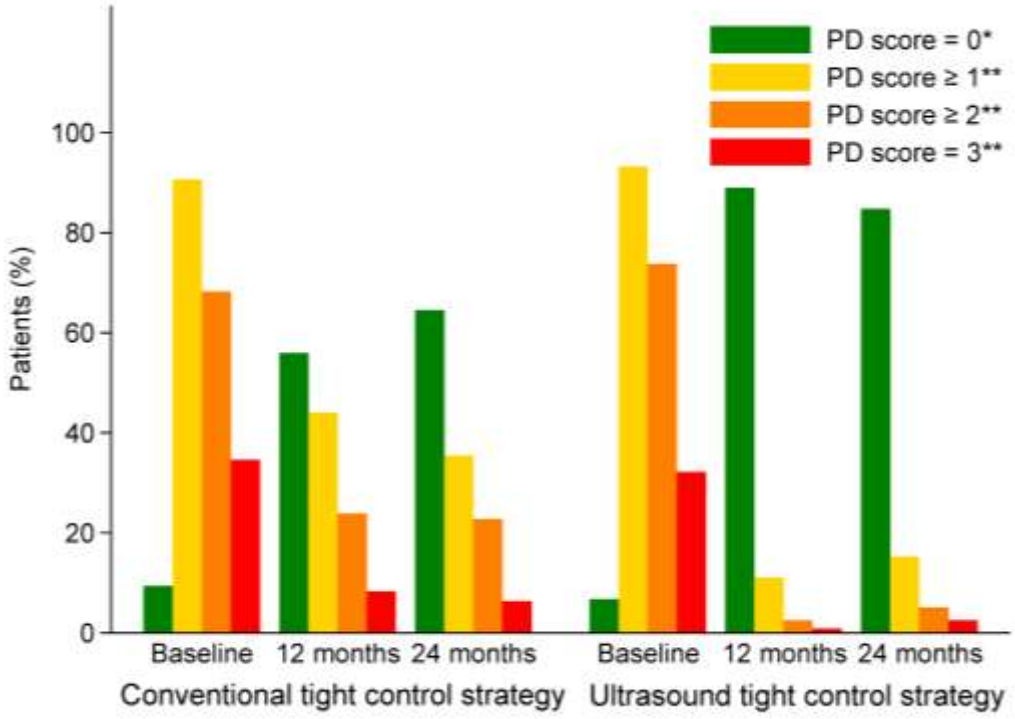
Response: We agree with the reviewer that the term "minimum" is imprecise, and have revised the sentence. At baseline, about 90% of patients displayed at least one joint with power Doppler positivity, at 12 and 24 months this was significantly reduced. As suggested by the reviewer, we have added a figure with the power Doppler data for both groups at the time points that were available. See also our response and action to reviewer 8 regarding the same issue.

Action: We revised the sentence and removed the term "minimum", the statement now reads: "Despite the more aggressive treatment in the ultrasound tight control group, inflammation assessed by ultrasound was significantly suppressed in both study arms, with a majority of patients having no power Doppler activity in any joint after two years."

We inserted a novel figure in the supplement:

Figure S1. Power Doppler activity at baseline, 12 months and 24 months

The histograms shows the proportion of patients with no power Doppler activity in any joint (PD score =0, colour green) and proportion of patients with at least one joint with power Doppler activity grade 1 (PD score ≥ 1, colour yellow), at least one joint with power Doppler activity grade 2 (PD score ≥ 2, colour orange) and at least one joint with power Doppler activity grade 3 (PD score = 3, colour red).



* in all joints ** in any joint

.

Response to:

Reviewer 8

Name: Peter Cheung
Job Title: Rheumatologist
Institution: National University Hospital, Singapore

Reviewer 8 comment 1:
This is a very well conducted randomized controlled study evaluating the utility of ultrasound as part of a tight treat to target strategy in patients with early RA. The authors wanted to see whether the additional goal of having no Power Doppler signal detected in joints would be clinically relevant, as compared to clinical remission alone in therapeutic decisions. It was concluded that US offered no additional benefits, with both primary and secondary outcomes being no different in both the US tight control arm and the clinical tight control arm. However, there was a trend for less radiographic progression at 2 years in the US tight control arm.

Major comment
-There is good reliability data for ultrasonographers. What about the clinicians who were determining the primary outcome (i.e. DAS and also clinical synovitis). Who were they? Were they the treating rheumatologists or separate metrologists? Either way, was there a multi-centre standardization exercise to document inter and intra-observer reliability?
Response: Thank you for your comment. The joint counts were performed either by the treating rheumatologist or a separate metrologist, depending on the organization at each participating study centre. We do not have formal documentation of inter and intra-observer reliability in the current study. However, a number of steps were taken to ensure reliability. There were multi-centre training and calibration exercises before the onset of the study, and at annual investigator meetings during the conduct of the study. In addition, all study personnel who performed joint counts where instructed to use the methodology recommended in the "Eular handbook of clinical assessments in rheumatoid arthritis" which was provided to the study personnel performing the joint counts (P.L.C.M. van Riel, J. Fransen, D.L. Scott. Van Zuiden Communications 2000, Leiden, the Netherlands. ISBN: 9075141904).
Action: Added the following to the methods section: "There were multi-centre training and calibration exercises of clinical joint examinations before the onset of the study, and at annual investigator meetings during the conduct of the study. Examinations were performed according to the EULAR handbook of clinical assessments in RA.[27]"

Reviewer 8 comment 2:
-The mean US power Doppler score was not that high at baseline, as with the DAS score. This may have contributed to the lack of difference in the primary and secondary outcome. A discussion of this could be relevant.
Response: We agree that the mean DAS was in the high moderate range (3.4-3.5, the cut-off for high disease being 3.7) at baseline. Disease activity is difficult to define based on the US power Doppler score alone as there are no accepted cut-off of moderate and high disease activity. The inclusion criteria stated that the patients should have clinical indication for DMARD therapy, as well as a diagnosis of RA and fulfil the 2010 ACR/EULAR classification criteria. There were no formal entry criteria defining the level of disease activity, except the requirement of at least one clinically swollen joint. In the current study we aimed to include a study population capturing a broad range of RA disease activity, not a subset with high disease activity, as is often the case in studies of new pharmaceutical products. We believe this heterogeneity is a strength of the current study, as it improves generalizability to real life practice. We are not sure how or if this decision influenced the observed rates of reaching the primary endpoint for the two treatment strategies.
Action: We added the following to the discussion: "The inclusion criteria stated that the patients should have clinical indication for DMARD therapy, as well as a diagnosis of RA and fulfilment of the 2010 ACR/EULAR classification criteria. There were no formal entry criteria regarding the level of disease activity. At baseline the mean disease activity score was in the upper range of moderate. In the current study we aimed to include a study population capturing a broad range of RA disease activity, not a subset with high disease activity, as is often the case in studies of new pharmaceutical products. We believe this heterogeneity is a strength of the current study, as it reflects real life practice."

Reviewer 8 comment 3:
In addition, information on proportion of joints with power Doppler scores 0,1,2,3 would be useful to understand the burden of highly inflamed joints, as opposed to joints which are primarily Grade 1, which can often be questionable in terms of clinical relevance.
Response and action: We agree that this information would be of interest, and have inserted a new figure S1 in the supplement, see also our response/action to reviewer 7, comment 3.

Minor comments

Reviewer 8 comment 4:
-The allocation of subjects appeared valid. Patients in the clinical arm also received US assessment at baseline and every year. Were the patients totally blinded to the US images? Did the protocol specifically indicate there was no communication between the ultrasonographer and the patient?
Response: The patients were blinded to the US images and results of the US examination. The investigators in the clinical arm did not have access to the scores in the electronic case record forms. This aspect was not stated in the protocol, but the ultrasonographers were

trained and instructed specifically not to give any information about the US assessment to the patients (or treating physicians) in the clinical arm.
Action: This section in methods has been revised: "Patients in the conventional tight control arm were assessed by ultrasound yearly, but both the patient and the treating physician were blinded to the results, and the treating physicians did not have access privileges to ultrasound data in the electronic case report form."

Reviewer 8 comment 5:
-There is a belief by rheumatologists that it is not necessary to achieve absolute zero Power Doppler, as this is present in normal joints sometimes. It would be relevant to indicate this in the discussion and to explain that one important aspect is to evaluate what is an acceptable cut-off for Power Doppler score, and also ultimately, it is a clinical evaluation of the patients as a whole when making therapeutic decisions.
Response: This is an interesting topic, and we agree that the cut-off of zero for the power Doppler score was not an easy decision to make, and we had extensive discussion with clinicians and researchers regarding this topic in the process of developing the protocol. The main reason for this cut-off was the results of the study by Scire et al, in which ROC analyses revealed a cut-off for the PD score >0 as the best cut-off with regard to prediction of flares in patients in clinical remission.[16] In a recent study by Padovano et al (ARD 2015), PD activity in healthy joints was rare.[28] This supports the notion that zero power Doppler activity may be the preferred target with regard to assessing US inflammation. Please also see comment 1 from reviewer 3 regarding the importance of the stringency of the target.
Action: We included the rationale for the cut-off of zero for power Doppler activity in the methods section (subheading treatment strategies): "The choice of no ultrasound power Doppler signal in any assessed joint as the preferred treatment target was based on available literature and extensive discussions with clinicians and researchers.[16]"

Reviewer 8 comment 6:
-What are the mean cumulative steroid dose (both oral and IA) for the groups?
Response and action: The mean triamcinolone hexacetonide dose for the i.a. injections is presented in table 2, and we have also updated table 2 to include the median cumulative oral steroid dose.

Reviewer 8 comment 7:
-Proportion of NSAID use should be included. In addition, it would be useful to indicate the mean methotrexate dose for each group.
Response and action: We agree, and this information has been included in table 2, under the subheading medication.

Reviewer 8 comment 8:
-There were obviously an increased number of injections in the US group. Is there any data available on the proportion of which joint regions were injected in the 2 arms? This may be relevant as injections to some joints may not be as effective as some other joint region.

Response: We have data on the specific joint level for all injections, but believe it is beyond the scope of the current manuscript to report detailed data on the joint level. However, if requested and warranted by the editor we can provide this information. In general, injections in large joints were similar between the groups, whereas the inter-carpal joints, the radio-ulnar joint and the MTP5 were more frequently injected in the US group, and the small joints of the feet distal to the MTPs and the PIP joints of the hands were more frequently injected in the conventional group.

Reviewer 8 comment 9:
Were the rheumatologists in the US arm allowed to inject the joint or tendon, only if they had power Doppler or can clinical involvement be enough?
Response: Clinical involvement could be enough, as described in the methods section: "In both arms, swollen joints were treated by intra-articular steroids, additionally any joint with power Doppler signal in the ultrasound tight control arm should be injected. All injections in the ultrasound tight control arm were guided by ultrasound." Further description is provided in section 3 in the online supplement.

Reviewer 8 comment 10:
-It is concerning that there were much more treatment escalations in the US arm, with no apparent gains in clinical outcome as well as functional/patient reported outcomes. In the US arm, how many patients who fulfilled the clinical improvement criteria (i.e. DAS) but not the PD criteria, and thus had to have an escalation in therapy, and how many required escalation of therapy with the clinical improvement criteria alone?
Response: In the US arm there were 1412 visits with treatment decisions, and the number of visits with treatment decisions with patients not having reached the treatment target, but still displaying a clinical response, was 129. In 93 of these cases there was also an ultrasound response, so in 33 out of these 129 cases (25.6%) the US improvement criteria alone indicated a treatment change that would not have been indicated in the clinical arm. In 168 treatment escalations in the US arm, 114 did not have a clinical response, these would also have required an escalation of treatment with the clinical improvement criteria alone.

Reviewer 8 comment 11:
-At one year, did the authors have any information on the number of patients who would need to have an escalation of therapy in the tight control clinical arm, if they had included the US information, which was blinded to the rheumatologists and patients?
Response: We have looked into this, and in patients who were in DAS remission at 12 months with no swollen joints, a total of 27 patients displayed at least one joint with PD activity. However, some of these might have had an ultrasound response, but as we did not assess ultrasound at every visit in this arm it is not possible to know which of these 27 patients would not require escalation due to having a treatment response if they were in the US arm.

Reviewer 8 comment 12:

-More biologics were started for the US arm. It would be useful to provide information on which biologic for both arms, and also whether there were any switchers.

Response: We agree that it would be useful to have more information regarding biologic therapy, and we have now provided information on the proportion of patients on 1st biologic, 2nd biologic and 3rd biologic in table 2, indicating the number of switchers. The data on specific biologics are provided below, we would be happy to include this as a table in the manuscript or supplement if the editor prefers a more detailed version.

Medication	Conventional arm	Ultrasound arm	Total
Adalimumab	0	2	2
Etanercept	8	11	19
Infliximab	1	3	4
Certolizumab pegol	4	13	17
Golimumab	4	2	6
Rituximab	1	1	2
Tocilizumab	1	1	2
Abatacept	0	1	1
Total	19	34	53

Action: We have updated table 2 with the proportion of patients on 1st biologic, 2nd biologic and 3rd biologic in both arms, indicating the number of switchers.

Reviewer 8 comment 13:
-Were measures placed in the protocol to ensure patients were adherent to their oral medications?
Response: Study personnel were instructed to discuss the importance of drug compliance to their patients, and to repeat this at every study visit. Current medication was reported in the electronic case report form at every visit. No formal measures of compliance (e.g. collection of empty containers) were undertaken.

References

1. Odegard S, Landewe R, van der Heijde D, et al. Association of early radiographic damage with impaired physical function in rheumatoid arthritis: a ten-year, longitudinal observational study in 238 patients. Arthritis Rheum 2006;**54**(1):68-75.
2. Aletaha D, Funovits J, Smolen JS. Physical disability in rheumatoid arthritis is associated with cartilage damage rather than bone destruction. Ann Rheum Dis 2011;**70**(5):733-9.
3. Klareskog L, Catrina AI, Paget S. Rheumatoid arthritis. Lancet 2009;**373**(9664):659-72.
4. Colebatch AN, Edwards CJ, Ostergaard M, et al. EULAR recommendations for the use of imaging of the joints in the clinical management of rheumatoid arthritis. Ann Rheum Dis 2013;**72**(6):804-14.
5. Matsos M, Harish S, Zia P, et al. Ultrasound of the hands and feet for rheumatological disorders: influence on clinical diagnostic confidence and patient management. Skeletal Radiol 2009;**38**(11):1049-54.
6. Filer A, de Pablo P, Allen G, et al. Utility of ultrasound joint counts in the prediction of rheumatoid arthritis in patients with very early synovitis. Ann Rheum Dis 2011;**70**(3):500-7.
7. Agrawal S, Bhagat SS, Dasgupta B. Improvement in diagnosis and management of musculoskeletal conditions with one-stop clinic-based ultrasonography. Mod Rheumatol 2009;**19**(1):53-6.
8. Sibbitt WL, Jr., Peisajovich A, Michael AA, et al. Does sonographic needle guidance affect the clinical outcome of intraarticular injections? The Journal of rheumatology 2009;**36**(9):1892-902.
9. Hammer HB, Terslev L. Role of ultrasound in managing rheumatoid arthritis. Curr Rheumatol Rep 2012;**14**(5):438-44.
10. Nguyen H, Ruyssen-Witrand A, Gandjbakhch F, et al. Prevalence of ultrasound-detected residual synovitis and risk of relapse and structural progression in rheumatoid arthritis patients in clinical remission: a systematic review and meta-analysis. Rheumatology (Oxford, England) 2014;**53**(11):2110-8.
11. Sakellariou G, Scire CA, Verstappen SM, et al. In patients with early rheumatoid arthritis, the new ACR/EULAR definition of remission identifies patients with persistent absence of functional disability and suppression of ultrasonographic synovitis. Ann Rheum Dis 2013;**72**(2):245-9.
12. Saleem B, Brown AK, Quinn M, et al. Can flare be predicted in DMARD treated RA patients in remission, and is it important? A cohort study. Ann Rheum Dis 2012;**71**(8):1316-21.
13. Brown AK, Conaghan PG, Karim Z, et al. An explanation for the apparent dissociation between clinical remission and continued structural deterioration in rheumatoid arthritis. Arthritis Rheum 2008;**58**(10):2958-67.
14. Brown AK, Quinn MA, Karim Z, et al. Presence of significant synovitis in rheumatoid arthritis patients with disease-modifying antirheumatic drug-induced clinical remission: evidence from an imaging study may explain structural progression. Arthritis Rheum 2006;**54**(12):3761-73.

15. Peluso G, Michelutti A, Bosello S, et al. Clinical and ultrasonographic remission determines different chances of relapse in early and long standing rheumatoid arthritis. Ann Rheum Dis 2011;**70**(1):172-5.

16. Scire CA, Montecucco C, Codullo V, et al. Ultrasonographic evaluation of joint involvement in early rheumatoid arthritis in clinical remission: power Doppler signal predicts short-term relapse. Rheumatology (Oxford, England) 2009;**48**(9):1092-7.

17. Gartner M, Alasti F, Supp G, et al. Persistence of subclinical sonographic joint activity in rheumatoid arthritis in sustained clinical remission. Ann Rheum Dis 2015;**74**(11):2050-3.

18. Markusse IM, Dirven L, Han KH, et al. Evaluating Adherence to a Treat-to-Target Protocol in Recent-Onset Rheumatoid Arthritis: Reasons for Compliance and Hesitation. Arthritis Care Res (Hoboken) 2016;**68**(4):446-53.

19. Hammer HB, Bolton-King P, Bakkeheim V, et al. Examination of intra and interrater reliability with a new ultrasonographic reference atlas for scoring of synovitis in patients with rheumatoid arthritis. Ann Rheum Dis 2011;**70**(11):1995-8.

20. van der Heijde DM, van 't Hof MA, van Riel PL, et al. Judging disease activity in clinical practice in rheumatoid arthritis: first step in the development of a disease activity score. Ann Rheum Dis 1990;**49**(11):916-20.

21. Hetland ML, Stengaard-Pedersen K, Junker P, et al. Aggressive combination therapy with intra-articular glucocorticoid injections and conventional disease-modifying anti-rheumatic drugs in early rheumatoid arthritis: second-year clinical and radiographic results from the CIMESTRA study. Ann Rheum Dis 2008;**67**(6):815-22.

22. Grigor C, Capell H, Stirling A, et al. Effect of a treatment strategy of tight control for rheumatoid arthritis (the TICORA study): a single-blind randomised controlled trial. Lancet 2004;**364**(9430):263-9.

23. Hammer HB, Kvien TK. Comparisons of 7- to 78-joint ultrasonography scores: all different joint combinations show equal response to adalimumab treatment in patients with rheumatoid arthritis. Arthritis Res Ther 2011;**13**(3):R78.

24. van Gestel AM, Prevoo ML, van 't Hof MA, et al. Development and validation of the European League Against Rheumatism response criteria for rheumatoid arthritis. Comparison with the preliminary American College of Rheumatology and the World Health Organization/International League Against Rheumatism Criteria. Arthritis Rheum 1996;**39**(1):34-40.

25. Torp-Pedersen ST, Terslev L. Settings and artefacts relevant in colour/power Doppler ultrasound in rheumatology. Ann Rheum Dis 2008;**67**(2):143-9.

26. Haavardsholm EA, Lie E, Lillegraven S. Should modern imaging be part of remission criteria in rheumatoid arthritis? Best Pract Res Clin Rheumatol 2012;**26**(6):767-85.

27. van Riel PvG, A.M.; Scott, D.L. *EULAR handbook of clinical assessments in rheumatoid arthritis : on behalf of the EULAR Standing Committee for International Clinical Studies Including Therapeutic Trials -ESCISIT*. The Netherlands: Van Zuiden, 2000.

28. Padovano I, Costantino F, Breban M, et al. Prevalence of ultrasound synovial inflammatory findings in healthy subjects. Ann Rheum Dis 2015.