PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

This paper was submitted to a another journal from BMJ but declined for publication following peer review. The authors addressed the reviewers' comments and submitted the revised paper to BMJ Open. The paper was subsequently accepted for publication at BMJ Open.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Minimal important differences for improvement in shoulder condition patient-reported outcomes: a systematic review to inform a BMJ Rapid Recommendation
AUTHORS	Hao, Qiukui; Devji, Tahira; Zeraatkar, Dena; Wang, Yuting; Qasim, Anila; Siemieniuk, Reed; Vandvik, Per; Lähdeoja, Tuomas; Carrasco Labra, Alonso; Agoritsas, Thomas; Guyatt, Gordon

VERSION 1 – REVIEW

REVIEWER	Ferrer, Montse IMIM (Hospital del Mar Research Institute), Health Services Research Unit
REVIEW RETURNED	16-Nov-2018

GENERAL COMMENTS	This article describes a systematic review of the Minimal Important Differences (MIDs) for patient-reported outcomes (PROs) measuring pain, function and health-related quality of life (HRQoL), estimated in studies of patients with shoulder conditions. Without any doubt, this review presents an interesting and well-developed approach to selecting the most adequate MIDs, which could be very useful in the PROs field. However, some details of the process need to be clarified, and findings merit further discussion. 1. What this study adds, page 3. "…we offered median estimates for the systematic review team." In my opinion, the median estimates selected in this study are of interest for a wider audience and purposes than the systematic review on effectiveness of the subacromial decompression surgery. My suggestion is to not restrict the offer of results to this survey, but open the offer for the multiple purposes where MID can be useful. 2. What this study adds, page 3. "Our survey informed a systematic review on effectiveness and BMJ Rapid Recommendations addressing subacromial decompression surgery for shoulder pain." Since this study was promoted to inform a systematic review previously planned, this sentence is not appropriated in the section of 'What this study adds'. Perhaps in 'What is already known on this topic'?

3. Background, page 4. “Although there is no “gold standard” anchor-based methodology, our group has used the existing literature and expert input to develop an instrument that measures the credibility of anchor-based MIDs. The instrument has proved reliable (manuscript in preparation). Among desirable criteria to establish a trustworthy MID (Table 1) is a requirement for at least a moderate correlation (e.g., >0.4) between change in the target PROM instrument and the change on the anchor (20 22).” In my opinion, information about reliability of the instrument and table 1 would be more suitably located in the methods section. It would be more informative to just list here the 5 aspects considered by this instrument to assess credibility.

4. Background, page 4. “No systematic survey has thus far summarized anchor-based MIDs of PROMs used in shoulder conditions applying an assessment of credibility.” It would be more informative to mention the systematic reviews previously published about this (such as references 49, 52-55), just highlighting that they did not apply any assessment of credibility.

5. Background, page 4. “Our review informed an associated BMJ Rapid Recommendations and facilitated interpretation of critical outcomes of interest, including shoulder pain, function, and health-related quality of life (HRQoL). The BMJ Rapid Recommendations project is a collaboration between the MAGIC foundation (www.magicproject.org) and the BMJ, with the goal of providing timely, trustworthy practice guidelines in response to new, potentially practice-changing evidence (23).” This ‘potentially practice-changing evidence’, what is it about? It is also a bit confusing that the first time that the ‘BMJ Rapid Recommendations’ are mentioned in the introduction section they are commented together with the general usefulness of the study (of facilitating interpretation). The reasons to promote this study are much more clearly stated in the sentence of ‘What this study adds’: “Our survey informed a systematic review on effectiveness and BMJ Rapid Recommendations addressing subacromial decompression surgery for shoulder pain.”

6. Methods, Protocol, page 5. “We conducted this systematic survey based on a registered published PROSPERO protocol (No.CRD42018106531).” Please clarify the difference between ‘systematic survey’ and ‘systematic review’.

7. Methods, Guideline panel and patient involvement, page 5. “The BMJ Rapid Recommendations guideline panel provided critical oversight to this systematic survey.” Please include a brief description on the actions comprising the ‘critical oversight’: besides the patients’ involvement in the identification of the outcomes of interest, in which other steps did they participate?

8. Methods, Instruments under consideration, page 5. “We addressed each of the PROMs corresponding to these constructs included as outcomes in the RCTs that proved eligible for the systematic review addressing the impact of shoulder surgery (Table 2).” Please describe briefly how the constructs identified by the panel were linked with PROs used in RCTs: was this process performed in the systematic review of effectiveness of the subacromial descompression surgery?

	9. Methods, Literature search and study identification, page 5. “The MID database project included the development of an instrument to assess the credibility of anchor-based MID estimates and tested its reliability (it proved reliable – manuscript in preparation, data available upon request).” This information is more appropriately located in the ‘Credibility assessment’ subsection of methods, in order to avoid repetition (see comments 3 and 11). 10. Methods, Study selection, page 6. “Eligible studies… Two reviewers independently performed title and abstract screening and, subsequently, full-text screening of studies included by either reviewer. At full-text screening, reviewers resolved the disagreement by discussion or, if needed by consultation with a third reviewer. We included studies with any intervention, including expectant management.” Please, move this last sentence before description of the title and abstract screening process in order to specify this together with the other inclusion criteria. 11. Methods, Credibility assessment, page 6. “…using an abridged version of the MID credibility tool used as part of the MID database (Table 1).” Table 1 shows the 5 aspects considered for the credibility assessment (patient’s perspective anchor instrument, its understandability, correlation with PRO, precision of MID estimate, and representation of small but important change) and response options, which for most aspects are 5: Definitely no; Not so much; To a great extent; Definitely yes; Not reported. Is there any kind of quantification for these labels? For example, about the correlation between the anchor and the PRO: any instruction for helping to qualify r=0.4 as ‘Not so much’ or ‘To a great extent’? 12. Methods, Credibility assessment, page 6. “We deemed that the MID estimate had high credibility if 3 or more of the 5 criteria were met; otherwise, we deemed that the MID had low credibility.” As response options are not dichotomic, please specify what would mean that a criterion was met: ‘Definitely yes’ only? Or either ‘Definitely yes’ or ‘To a great extent’? 13. Methods, Practical application of MID estimates in BMJ Rapid Recommendations development, page 7. “The authors of the systematic review addressing effectiveness of shoulder surgery applied the most credible MID estimates identified from our review in interpreting the magnitudes of effect in the GRADE Summary of Findings table. The systematic review informed the BMJ Rapid Recommendations panel in their development of the guideline.” In my opinion, the practical application of the study results would be better located in discussion than methods (see comments 17 and 18). On the other hand, it is important to describe in methods the reasons/process to decide selecting the median of the MIDs. 14. Results, page 7. In my opinion, it would be more informative about the different steps followed to first comment table 2 and then table 3. Please review duplicities between these two tables for score range and construct measured, and add follow-up period in table 3, if possible. 15. Results, page 7-8. “For the MID estimates with high credibility, MIDs for the Constant Score and overall pain VAS were consistent across the 2 available estimates. The MIDs for the Constant Score (3 to 16.6), DASH (4.4 to 25.4), and OSS (4.0 to 14.7) were,

	however, inconsistent among 6-10 estimates provided.” I do not understand if/when MIDs for the Constant Score were consistent (first sentence) or inconsistent (second sentence). 16. Discussion, first paragraph, page 8. “MIDs of high credibility for pain and function outcomes and of low credibility for HRQoL.” In my opinion, it is important to add a paragraph commenting which aspects where more frequently met and which ones were more frequently not. The latter information is especially relevant to understand why MIDs for HRQoL instruments presented low credibility. Information about credibility of the MIDs examined is very valuable to detect methodological gaps and provide recommendations to improve quality in the PRO field. 17. Discussion, first paragraph, page 8.”MIDs estimates often varied widely; we offered median estimates for the systematic review team and guideline panel.” It is necessary to add in the discussion section a paragraph commenting the implication of selecting the median, and the potential effects of this decision on the infra or supra- detection of benefits in the systematic review on effectiveness of the subacromial decompression surgery. 18. Discussion, first paragraph, page 8. “Authors of the linked review applied these MIDs in GRADE evidence summaries; the BMJ Rapid Recommendations guideline panel used them to inform their judgements of magnitude of effect in formulating their recommendations.” Instead of the first paragraph of discussion (where generally a summary of main results are provided), it would be suitable to add a paragraph commenting the practical application of the MIDs selected for the systematic review which promoted this study, but also beyond this. 19. Conclusion, page 9. “The MID estimates inform the interpretation for a linked systematic review and guideline on arthroscopy for shoulder pain. Authors of reviews of other interventions for shoulder conditions can in future make use of our summary MIDs.” I would recommend to open the usefulness of the study findings to all type of studies (not only reviews of intervention), for example for helping to interpret the magnitude of findings, and also to calculate statistical power/sample size in primary studies.

REVIEWER	Angst, Felix Rehabiliation Clinic ("RehaClinic), Bad Zurzach, Research
REVIEW RETURNED	17-Nov-2018

GENERAL COMMENTS	Positive criticism This study describes the reviewed literature results of M(C)ID values/levels for improvements in pain, function and quality of life (QOL) in various shoulder conditions. The authors have made big efforts to perform this work. The report is very clearly written, short and concise. Especially positive characteristics are: 1. The rating of the quality of the reviewed studies by the credibility assessment. 2. P5/50, line 27ff: The specification and choice of only anchor-based MIDs, i.e. MCIDs. Distribution-based MIDs are – per definition – not clinically relevant, as described in the text. They are, in principle, only dependent on the sample size, and, by that, a type of smallest detectable differences (SDDs) (1,2).

Negative criticism

1. The first major issue that has room for improvement and needs revision refers to the credibility rating (Table 1).
1.1. Although all listed criteria of the items are very plausible and – by not-cited literature and experience – important for a robust MID evaluation (face and content validity), describe information about the validity of this assessment in greater detail.
1.2. Did you perform a validity study of the credibility rating? If yes, describe those results and the characteristics of the form in a short paragraph.
1.3. How do you quantify the credibility, by the sum of the single items score points?
1.4. "High" credibility was rated if the score is ≥3, "low" if the score is 0,1 or 2 – in order to stratify the MIDs, e.g. in Table 4. With other words, higher scores mean higher credibility.
1.4.1. How was this threshold chosen? Again: its validity?
1.4.2. Item 1 "impossible to tell" gives 2 points and item 2 "impossible to tell" gives 4 points. The summary would be =6 meaning high credibility – this makes no sense. Please comment.
1.5. Credibility rating, item 3: Correlation (r) anchor (Likert rating: categorical) and PRO (score changes: continuous). You defined a minimal level of r≥0.40 (p 5/50, line 42). Please outline that in Table 1.
1.5.1. Please give arguments for this level of 0.40. Previous literature rather defined r≥0.30 (2). Are there empiric data for the validity of this level or is this rather a rule of thumb? Please specify and comment.
1.5.2. Please outline that only the Spearman rank, not Pearson product-moment correlations are appropriate and used for the study because the anchor is not continuous.

2. The second and more important issue refers to the different methods of MID determination. This is important for generalisation of the presented MID levels and their use for future studies. This may also be a reason for the wide variation of the MID levels of the single studies reviewed.
2.1. Of course, for a meta-analytical (MA) pooling, you need as many studies as possible to obtain a pooled parameter that is substantially more valid than the single estimate of one study. By that, many studies with different (baseline, methodological etc.) characteristics are summarised doing so as if there were no differences – the well-known methodological shortcoming of every review/MA. Please comment this weakness as limitation.
2.2. In the case of the present study, this has done (beside other disease-relevant characteristics of the sample, diagnosis, type of intervention etc.) for the type of method how to determine the MID. Please comment that.
2.3. Correctly, the three main methods are very shortly described (on p 8/50, line 41ff): 1) the simple method of Jaeschke et al. 1989 (3) using the score change of the "somewhat better" transition group, 2) the "mean change" method of Redelmeier & Lorig 1993 (4) using the difference of the score changes of "somewhat better" and "almost equal", and 3) the ROC curve method according to Youden (see in 2) using a single point estimate of the receiver operating characteristic curve between those two groups.
2.3.1 These three methods have to be described in the methods section in greater extent and the corresponding original literature has to be referenced.

2.3.2. Discuss shortly advantages/disadvantages of the three methods. For example, that 2) is the most often used and accepted method in literature and possible the most valid one and that 3) has shortcomings because it is a single point estimate; consider also (1,2).

2.4. On p. 9/50, line 37ff, it is stated:

“The results of our systematic survey have limitations. The range of reported MIDs was wide for some of the PROMs (e.g., 0.3 to 30 for Constant score; 4.4 to 25.41 for Disabilities of the Arm, Shoulder and Hand (DASH)). Baseline characteristics (participants’ disease/conditions, sample size, PROMs or instruments), anchors and analytic methods varied among included studies, but these differences did not explain variability in the estimates.”

In contrast to that, empirical comparison of the three methods at the same setting found considerable differences between the MID levels: 13.51 points (scale 0-100) by method 1), 8.74 by 2), 15.00 by 3) (1). Thus, the statement “but these differences did not explain variability in the estimates” must be wrong. In contrary, pooling data across those three methods is likely to be one of major the reasons for the wide variation of the MID levels of the single studies reviewed.

3. Relative MIDs – the issue of bias of MIDs.

3.1. In Table 4, p. 25/50, line 41ff, relative MIDs are shown (in %). The level of a relative MID is biased by the baseline score, especially at both ends of a closed scale – see example and discussion in (1). Please comment.

3.2. To me, it goes too far to discuss all possibilities and problems of bias of parameters that express MIDs in this paper (1). However, some short discussion has to be conducted and added about possibilities to reduce bias of MIDs in order to present a valuable and informative BMJ Rapid Recommendation in this theme, based on the information of (1,5) and possibly of other references.

Minor points:

4. The simple shoulder test (SST) has considerably low MIDs (Table 4, p. 25/50, line 17,45) when compared to all other reviewed scores: For example: SST: 1.8 score points, compared to the DASH: 10.2 and the Constant score: 8.3 etc. – all instruments are scaled by 0-100. Please comment and explain why. State hypotheses.

5. The title needs the specification that only MIDs for improvement were reviewed.

6. There are two page numberings: n of 50 or m of 25. Please choose only one and number consistently.

7. At the end: Try to formulate a distinct take home message for the reader.

References

1 Angst F, Aeschlimann A, Angst J. The minimal clinically important difference (MCID) raised the significance of outcome

	effects above the statistical level, with methodological implications for future studies. J Clin Epidemiol 2017;82:128-36. 2 Revicki D, Hays RD, Cella DE, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. J Clin Epidemiol 2008;61:102-9. 3 Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. Controlled Clin Trials 1989;10:407-15. 4 Redelmeier DA, Lorig K. Assessing the clinical importance of symptomatic improvements. An illustration in rheumatology. Arch Intern Med 1993;153:1337-42. 5 Angst F, Benz T, Lehmann S, Aeschlimann A, Angst J. Multidimensional minimal clinically important differences (MCID) in knee osteoarthritis after comprehensive rehabilitation: a prospective evaluation from the Bad Zurzach Osteoarthritis Study. RMD open 2018. Oct 8;4(2):e000685. doi: 10.1136/rmdopen-2018-000685. eCollection 2018.

REVIEWER	Jeppe Vejlgaard Rasmussen MD, PhD, associate professor
REVIEW RETURNED	18-Nov-2018

GENERAL COMMENTS	Thank you for the opportunity to review this interesting manuscript. The topic is of great interest and within the scope of the BMJ Rapid Recommendation concept. The manuscript is well-written. What I miss is a recommendation regarding the MID of the included PROMs or at least a detailed discussion of the heterogenicity of the results and the fact that there is no clearly defined MID for the PROMs in the field of shoulder surgery. My specific concerns: Abstract: It is stated that the objective is "to identify credible anchor-based minimal important differences (MIDs) for patient reported outcome measures (PROMs) relevant to a BMJ Rapid Recommendations addressing subacromial decompression surgery for shoulder pain" and in the conclusion that "The MID estimates inform the interpretation for a linked systematic review and guideline addressing subacromial decompression surgery for shoulder pain. I get the impression that the authors only include studies of subacromial decompression and that the conclusion of the study is valid for this specific pathology only and not for cuff tear, glenohumeral osteoarthritis or humeral fractures. However, several eligible studies (table 3) is about shoulder arthroplasty and humeral fractures? Please clarify. Background The authors describe that "In this systematic survey, we (1) summarize MID estimates for the PROMs used in RCTs that investigate the effect of surgery on shoulder pain, and (2) assessed the credibility of these MID estimates." However, some studies are not RCTs and not all studies include surgery? It is stated that "eligible studies used any design including retrospective and prospective observational studies…. (page 7 line 11). Please clarify.

	Method: The Constant score is not exactly a patient-reported outcome, so why is this measurement tool included (It is recognized by the authors, page 6 line 47)? If the Constant score is included (the measurement tool with the highest number of papers) the paper is more about outcome measurement of shoulder pathologies and not patient-reported outcomes. Results: No comments Discussion: The authors acknowledge that the range of reported MIDs was wide and use the Constant Score as an example. For the absolute Constant score the Median MCI was 8.3 range 3 to 16.6, n=10 (high credibility). Furthermore, there is a wide range of pathologies from nonoperative treatment of impingement to shoulder arthroplasty for complex humeral fractures. The MIDs for these pathologies are, from a clinical perspective, not the same. Patients with impingement can often regain full range of motion and no pain whereas the aim of surgical treatment of complex humeral fractures is pain-relief (which only accounts for 15% of the maximum Constant Score). Thus, I am not convinced that an unweighted median value should be used to define the MID. Wouldn't it be better to use the MID that has been reported for specific treatment of specific pathologies such as shoulder arthroplasty for humeral fractures? I believe this should be discussed in more detail. Conclusion: The authors state that "The MID estimates inform the interpretation for a linked systemtatic review and guideline on arthroscopy for shoulder pain." Again, the results are not specific for patients treated with arthroscopy. Please clarify. best regards

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1

Comments:
This article describes a systematic review of the Minimal Important Differences (MIDs) for patient-reported outcomes (PROs) measuring pain, function and health-related quality of life (HRQoL), estimated in studies of patients with shoulder conditions. Without any doubt, this review presents an interesting and well-developed approach to selecting the most adequate MIDs, which could be very useful in the PROs field. However, some details of the process need to be clarified, and findings merit further discussion.

RESPONSE: We thank the reviewer for the positive evaluation of our paper.

1. What this study adds, page 3. "…we offered median estimates for the systematic review team." In my opinion, the median estimates selected in this study are of interest for a wider audience and purposes than the systematic review on effectiveness of the subacromial decompression surgery. My suggestion is to not restrict the offer of results to this survey, but open the offer for the multiple purposes where MID can be useful.

RESPONSE: We thank the reviewer's suggestion. We have addressed the issue in the abstract of the revised manuscript as follows:

> "The MID estimates inform the interpretation for a linked systematic review and guideline addressing subacromial decompression surgery for shoulder pain, and could also prove useful for authors addressing other interventions for shoulder problems."

and again, in the final sentence of the revised manuscript:

> "Researchers addressing a wide variety of shoulder conditions can in future make use of our summary MIDs to inform sample size and aid in interpretation of results."

> 2. What this study adds, page 3. "Our survey informed a systematic review on effectiveness and BMJ Rapid Recommendations addressing subacromial decompression surgery for shoulder pain." Since this study was promoted to inform a systematic review previously planned, this sentence is not appropriated in the section of 'What this study adds'. Perhaps in 'What is already known on this topic'?

RESPONSE: The BMJ Open does not now request a section on "What this study adds" but rather a "strength and Limitations" section, we have placed the relevant sentence in the body of the paper but not in that section.

3. Background, page 4. "Although there is no "gold standard" anchor-based methodology, our group has used the existing literature and expert input to develop an instrument that measures the credibility of anchor-based MIDs. The instrument has proved reliable (manuscript in preparation). Among desirable criteria to establish a trustworthy MID (Table 1) is a requirement for at least a moderate correlation (e.g., >0.4) between change in the target PROM instrument and the change on the anchor (20 22)." In my opinion, information about reliability of the instrument and table 1 would be more suitably located in the methods section. It would be more informative to just list here the 5 aspects considered by this instrument to assess credibility.

RESPONSE: We moved table 1 to the method section (Now table 2). However, the details of the instrument, e.g. reliability, are best placed in a paper now in preparation – the BMJ is our target journal for this submission, which the BMJ should receive in the next weeks.

4. Background, page 4. “No systematic survey has thus far summarized anchor-based MIDs of PROMs used in shoulder conditions applying an assessment of credibility.” It would be more informative to mention the systematic reviews previously published about this (such as references 49, 52-55), just highlighting that they did not apply any assessment of credibility.

RESPONSE: We have amended the statement as follows:

> “Although systematic reviews addressing MIDs in shoulder PROMs are available [1-5], they are dated and have not applied an assessment of credibility.”

5. Background, page 4. “Our review informed an associated BMJ Rapid Recommendations and facilitated interpretation of critical outcomes of interest, including shoulder pain, function, and health-related quality of life (HRQoL). The BMJ Rapid Recommendations project is a collaboration between the MAGIC foundation (www.magicproject.org) and the BMJ, with the goal of providing timely, trustworthy practice guidelines in response to new, potentially practice-changing evidence (23).” This ‘potentially practice-changing evidence’, what is it about? It is also a bit confusing that the first time that the ‘BMJ Rapid Recommendations’ are mentioned in the introduction section they are commented together with the general usefulness of the study (of facilitating interpretation). The reasons to promote this study are much more clearly stated in the sentence of ‘What this study adds’: “Our survey informed a systematic review on effectiveness and BMJ Rapid Recommendations addressing subacromial decompression surgery for shoulder pain.”

RESPONSE: Here, we try to introduce the BMJ Rapid Recommendations’. We have used the reviewer’s suggested wording in the relevant section of the Background.

6. Methods, Protocol, page 5. “We conducted this systematic survey based on a registered published PROSPERO protocol (No.CRD42018106531).” Please clarify the difference between ‘systematic survey’ and ‘systematic review’.

RESPONSE: we used the term “systematic review” instead of “systematic survey” in the new submission.

7. Methods, Guideline panel and patient involvement, page 5. “The BMJ Rapid Recommendations guideline panel provided critical oversight to this systematic survey.” Please include a brief description on the actions comprising the ‘critical oversight’: besides the patients’ involvement in the identification of the outcomes of interest, in which other steps did they participate?

RESPONSE: We added this in the method section.

> “The panel members also provided input into the methodology of our review.

8. Methods, Instruments under consideration, page 5. “We addressed each of the PROMs corresponding to these constructs included as outcomes in the RCTs that proved eligible for the systematic review addressing the impact of shoulder surgery (Table 2).” Please describe briefly how the constructs identified by the panel were linked with PROs used in RCTs: was this process performed in the systematic review of effectiveness of the subacromial decompression surgery?

RESPONSE: We have expanded on the process as follows:

> The *BMJ* Rapid Recommendations panel, informed by the Outcome Measures in Rheumatology (OMERACT) shoulder core outcomes set [6], nominated shoulder pain, function, and health-related quality of life (HRQoL) as critical patient-important outcomes of interest in the management of shoulder conditions. Following guidance from the panel, the systematic review team addressing the effectiveness of surgery for subacromial pain syndrome sought evidence for each of these outcomes in the eligible RCTs. We worked closely with that review team and addressed each of the PROMs corresponding to these constructs included as outcomes in the RCTs that proved eligible for the systematic review addressing the impact of shoulder surgery (the subacromial decompression surgery) (Table 1).

9. Methods, Literature search and study identification, page 5. “The MID database project included the development of an instrument to assess the credibility of anchor-based MID estimates and tested its reliability (it proved reliable – manuscript in preparation, data available upon request).” This information is more appropriately located in the ‘Credibility assessment’ subsection of methods, in order to avoid repetition (see comments 3 and 11).

RESPONSE: We moved this sentence to the “Credibility assessment” subsection of methods as suggested.

10. Methods, Study selection, page 6. “Eligible studies… Two reviewers independently performed title and abstract screening and, subsequently, full-text screening of studies included by either reviewer. At full-text screening, reviewers resolved the disagreement by discussion or, if needed by consultation with a third reviewer. We included studies with any intervention, including expectant management.” Please, move this last sentence before description of the title and abstract screening process in order to specify this together with the other inclusion criteria.

RESPONSE: Revised as suggested.

11. Methods, Credibility assessment, page 6. “…using an abridged version of the MID credibility tool used as part of the MID database (Table 1).” Table 1 shows the 5 aspects considered for the credibility assessment (patient’s perspective anchor instrument, its understandability, correlation with PRO, precision of MID estimate, and representation of small but important change) and response options, which for most aspects are 5: Definitely no; Not so much; To a great extent; Definitely yes; Not reported. Is there any kind of quantification for these labels? For example, about the correlation

between the anchor and the PRO: any instruction for helping to qualify r=0.4 as ‘Not so much’ or ‘To a great extent’?

RESPONSE: Yes, we have. If the r≧0.3 and < 0.5, we judge the item as “not so much”. The details of the instrument of credibility assessment will be presented in a paper now in preparation and nearing readiness for submission – the BMJ will be our target journal.

12. Methods, Credibility assessment, page 6. “We deemed that the MID estimate had high credibility if 3 or more of the 5 criteria were met; otherwise, we deemed that the MID had low credibility.” As response options are not dichotomic, please specify what would mean that a criterion was met: ‘Definitely yes’ only? Or either ‘Definitely yes’ or ‘To a great extent’?

RESPONSE: Yes, we deemed that the MID estimate had high credibility if 3 or more of the 5 criteria were met (either ‘Definitely yes’ or ‘To a great extent’ for each item).

> “We deemed that the MID estimate had high credibility if 3 or more of the 5 criteria were met (either ‘Definitely yes’ or ‘To a great extent’ for each item); otherwise, we deemed that the MID had low credibility”

13. Methods, Practical application of MID estimates in BMJ Rapid Recommendations development, page 7. “The authors of the systematic review addressing effectiveness of shoulder surgery applied the most credible MID estimates identified from our review in interpreting the magnitudes of effect in the GRADE Summary of Findings table. The systematic review informed the BMJ Rapid Recommendations panel in their development of the guideline.” In my opinion, the practical application of the study results would be better located in discussion than methods (see comments 17 and 18). On the other hand, it is important to describe in methods the reasons/process to decide selecting the median of the MIDs.

RESPONSE: Thanks, we agree. We made changes as suggested.

> “We also provided the systematic review team with the median, minimum and maximum values across the range of high credibility trustworthy MID estimates generated from the eligible studies for the PROMs of interest.”

14. Results, page 7. In my opinion, it would be more informative about the different steps followed to first comment table 2 and then table 3. Please review duplicities between these two tables for score range and construct measured, and add follow-up period in table 3, if possible.

RESPONSE: We have deleted the score range and construct measured from Table 3 and added the follow-up period as suggested. (Table 3)

15. Results, page 7-8. “For the MID estimates with high credibility, MIDs for the Constant Score and overall pain VAS were consistent across the 2 available estimates. The MIDs for the Constant Score (3 to 16.6), DASH (4.4 to 25.4), and OSS (4.0 to 14.7) were, however, inconsistent among 6-10 estimates provided.” I do not understand if/when MIDs for the Constant Score were consistent (first sentence) or inconsistent (second sentence).

RESPONSE: Sorry, this is a mistake. In the first sentence, the Constant score should be the Simple Shoulder Test (SST). We corrected it accordingly.

> “For the MID estimates with high credibility, MIDs for the SST (1.5 to 2.1) and overall pain VAS (1.4 to 1.6) were consistent across the 2 available estimates. The MIDs for the Constant Score (3 to 16.6), DASH (4.4 to 25.4), and OSS (4.0 to 14.7) were, however, inconsistent among 6-10 estimates provided.”

16. Discussion, first paragraph, page 8. “MIDs of high credibility for pain and function outcomes and of low credibility for HRQoL.” In my opinion, it is important to add a paragraph commenting which aspects where more frequently met and which ones were more frequently not. The latter information is especially relevant to understand why MIDs for HRQoL instruments presented low credibility. Information about credibility of the MIDs examined is very valuable to detect methodological gaps and provide recommendations to improve quality in the PRO field.

RESPONSE: Revised as suggested. We added the following paragraph.

> “For the credibility assessment, we found that the anchor instrument directly addressed the patient’s perspective, and judged the understanding the anchor instrument for patients as ‘Definitely yes’ or ‘To a great extent’, for all the MID estimates. Approximately half of the estimates did not report the correlation between the anchor and the PROM. We judged the precision of the MID estimation and the threshold or difference between groups on the anchor used to estimate the MID as “Definitely no” or “Not so much” for most MID estimates.”

17. Discussion, first paragraph, page 8.”MIDs estimates often varied widely; we offered median estimates for the systematic review team and guideline panel.” It is necessary to add in the discussion section a paragraph commenting the implication of selecting the median, and the potential effects of this decision on the infra or supra- detection of benefits in the systematic review on effectiveness of the subacromial decompression surgery.

RESPONSE: Revised as follows.

> “We also provided the systematic review team with the median, minimum and maximum values across the range of high credibility trustworthy MID estimates generated from the eligible studies for the PROMs of interest. The only instance in which the variability in scores was sufficiently

> great that choice of one of the extremes rather than the median could substantially influence conclusions was for the Constant score."

18. Discussion, first paragraph, page 8. "Authors of the linked review applied these MIDs in GRADE evidence summaries; the BMJ Rapid Recommendations guideline panel used them to inform their judgements of magnitude of effect in formulating their recommendations." Instead of the first paragraph of discussion (where generally a summary of main results are provided), it would be suitable to add a paragraph commenting the practical application of the MIDs selected for the systematic review which promoted this study, but also beyond this.

RESPONSE: Revised as suggested.

> "Authors of the linked review used these MIDs (Pain VAS 0-10 1.5 units, the Constant score 0-100 scale 8.3 units, and EQ 5-D, 0.07 units) to gauge the importance of possible difference patients in GRADE evidence summaries and to dichotomize the improvements (proportions of patients achieving MID or more); the BMJ Rapid Recommendations guideline panel used them to inform their judgements of magnitude of effect in formulating their recommendations. The systematic review informed the BMJ Rapid Recommendations panel in their development of the guideline."

19. Conclusion, page 9. "The MID estimates inform the interpretation for a linked systematic review and guideline on arthroscopy for shoulder pain. Authors of reviews of other interventions for shoulder conditions can in future make use of our summary MIDs." I would recommend to open the usefulness of the study findings to all type of studies (not only reviews of intervention), for example for helping to interpret the magnitude of findings, and also to calculate statistical power/sample size in primary studies.

RESPONSE: Revised as suggested – please see response to reviewer's point 1.

Reviewer: 2

Comments:
Positive criticism
This study describes the reviewed literature results of M(C)ID values/levels for improvements in pain, function and quality of life (QOL) in various shoulder conditions. The authors have made big efforts to perform this work. The report is very clearly written, short and concise. Especially positive characteristics are:
1. The rating of the quality of the reviewed studies by the credibility assessment.
2. P5/50, line 27ff: The specification and choice of only anchor-based MIDs, i.e. MCIDs. Distribution-based MIDs are – per definition – not clinically relevant, as described in the text. They are, in principle, only dependent on the sample size, and, by that, a type of smallest detectable differences (SDDs) (1,2).

RESPONSE: We thank the reviewer for the positive evaluation of our paper.

Negative criticism

1. The first major issue that has room for improvement and needs revision refers to the credibility rating (Table 1).
1.1. Although all listed criteria of the items are very plausible and – by not-cited literature and experience – important for a robust MID evaluation (face and content validity), describe information about the validity of this assessment in greater detail.

1.2. Did you perform a validity study of the credibility rating? If yes, describe those results and the characteristics of the form in a short paragraph.

RESPONSE: We have addressed the reliability of the credibility instrument but not its validity. We have added this as a limitation in the strength and limitations section as follows:

> "With respect to the assessment of credibility, a formal assessment of the validity of the instrument has not been undertaken."

1.3. How do you quantify the credibility, by the sum of the single items score points?

RESPONSE: NO. We did not sum the score points and quantify the credibility. We only count how many criteria the estimation met (either 'Definitely yes' or 'To a great extent' for each item)? We describe the process in the revised manuscript as follows:

> "We deemed that the MID estimate had high credibility if 3 or more of the 5 criteria were met (either 'Definitely yes' or 'To a great extent' for each item); otherwise, we deemed that the MID had low credibility. We regard the credibility as a dichotomous variable (high and low) and do not quantify the credibility."

1.4. "High" credibility was rated if the score is ≥3, "low" if the score is 0,1 or 2 – in order to stratify the MIDs, e.g. in Table 4. With other words, higher scores mean higher credibility.

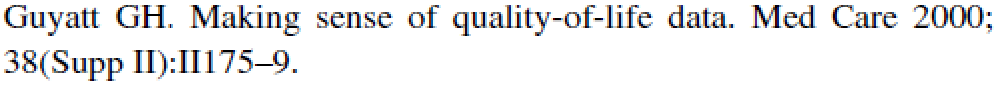
1.4.1. How was this threshold chosen? Again: its validity?
1.4.2. Item 1 "impossible to tell" gives 2 points and item 2 "impossible to tell" gives 4 points. The summary would be =6 meaning high credibility – this makes no sense. Please comment.

RESPONSE: As the same as the last response, we did not quantify the credibility. One could challenge our decision of a threshold for high and low credibility and we have added this as another limitation as follows:

> "Moreover, one might challenge our judgment in inferring high credibility if 3 or more criteria were met."

1.5. Credibility rating, item 3: Correlation (r) anchor (Likert rating: categorical) and PRO (score changes: continuous). You defined a minimal level of r≥0.40 (p 5/50, line 42). Please outline that in Table 1.
1.5.1. Please give arguments for this level of 0.40. Previous literature rather defined r≥0.30 (2). Are there empiric data for the validity of this level or is this rather a rule of thumb? Please specify and comment.

RESPONSE: We did not define a minimal level of r≥0.40 in this study. The original presentation about this is a misleading. We set a number of thresholds for the in the correlation between anchor and the PROMs (0.3: Definitely no; ≧0.3 to 0.51: Not so much; ≧0.5 to <0.7: To a great extent; ≧0.7: Definitely). The key threshold is actually 0.5. There is empirical data to support this threshold (see the following reference [7])



Guyatt GH. Making sense of quality-of-life data. Med Care 2000; 38(Supp II):II175–9.

1.5.2. Please outline that only the Spearman rank, not Pearson product-moment correlations are appropriate and used for the study because the anchor is not continuous.

RESPONSE: we added a relevant footnote to table 2 as follows.

> “@ For anchors with categorical scales the Spearman rather the Pearson’s correlation, is appropriate.”

2. The second and more important issue refers to the different methods of MID determination. This is important for generalisation of the presented MID levels and their use for future studies. This may also be a reason for the wide variation of the MID levels of the single studies reviewed.
2.1. Of course, for a meta-analytical (MA) pooling, you need as many studies as possible to obtain a pooled parameter that is substantially more valid than the single estimate of one study. By that, many studies with different (baseline, methodological etc.) characteristics are summarised doing so as if there were no differences – the well-known methodological shortcoming of every review/MA. Please comment this weakness as limitation.

RESPONSE: Because of the very point the reviewer makes we did not pool across MIDs provided the median, minimum and maximum values. We did not do a meta-analysis. We have identified the variability in MIDs as a limitation of the review as follows:

> “The range of reported MIDs was wide for some of the PROMs (e.g., 0.3 to 30 for Constant score; 4.4 to 25.41 for Disabilities of the Arm, Shoulder and Hand (DASH)).”

2.2. In the case of the present study, this has done (beside other disease-relevant characteristics of the sample, diagnosis, type of intervention etc.) for the type of method how to determine the MID. Please comment that.

RESPONSE: We are not certain of the reviewer's point, but suspect it has to do with the issue raised in section 2.4. Please see our comment there.

2.3. Correctly, the three main methods are very shortly described (on p 8/50, line 41ff): 1) the simple method of Jaeschke et al. 1989 (3) using the score change of the "somewhat better" transition group, 2) the "mean change" method of Redelmeier & Lorig 1993 (4) using the difference of the score changes of "somewhat better" and "almost equal", and 3) the ROC curve method according to Youden (see in 2) using a single point estimate of the receiver operating characteristic curve between those two groups.
2.3.1 These three methods have to be described in the methods section in greater extent and the corresponding original literature has to be referenced.
2.3.2. Discuss shortly advantages/disadvantages of the three methods. For example, that 2) is the most often used and accepted method in literature and possible the most valid one and that 3) has shortcomings because it is a single point estimate; consider also (1,2).

RESPONSE: We consider the discussion that the reviewer suggests somewhat tangential and beyond the scope of the manuscript. We do, however, think that the reviewer's essential point is worth making, and that it is worthwhile to add the citations the reviewer suggests. We have therefore addressed the issue in the limitations section of the manuscript.

> "Finally, investigators used different methods to relate the anchor to a transition rating; the optimal approach remains uncertain [8 9]."

2.4. On p. 9/50, line 37ff, it is stated:
"The results of our systematic survey have limitations. The range of reported MIDs was wide for some of the PROMs (e.g., 0.3 to 30 for Constant score; 4.4 to 25.41 for Disabilities of the Arm, Shoulder and Hand (DASH)). Baseline characteristics (participants' disease/conditions, sample size, PROMs or instruments), anchors and analytic methods varied among included studies, but these differences did not explain variability in the estimates."
In contrast to that, empirical comparison of the three methods at the same setting found considerable differences between the MID levels: 13.51 points (scale 0-100) by method 1), 8.74 by 2), 15.00 by 3) (1). Thus, the statement "but these differences did not explain variability in the estimates" must be wrong. In contrary, pooling data across those three methods is likely to be one of major the reasons for the wide variation of the MID levels of the single studies reviewed.

RESPONSE: The reviewer's prior work (we suspect) was able to demonstrate an association between the MID and the method of relating the anchor to the target instrument score. We were unable to establish such a relationship in our data and acknowledge this as a limitation of our review, at the same time acknowledging the prior work that did establish an association as below:

> "…though others have detected associations between methodological approaches and MIDs, our attempts to establish a clear relation between these variables and the MID were not successful [8]."

3. Relative MIDs – the issue of bias of MIDs.
3.1. In Table 4, p. 25/50, line 41ff, relative MIDs are shown (in %). The level of a relative MID is biased by the baseline score, especially at both ends of a closed scale – see example and discussion in (1). Please comment.

RESPONSE: We only identify 4 MID estimates for the relative MIDs and the review of effectiveness and the guideline team did not use these relative in practice. Thus, our view is that further discussion of this point – although we agree with the reviewer's point well illustrated in the prior work – is tangential and would be distracting to include. Please note we have cited the reviewer's reference 1 at several places in the revised manuscript.

3.2. To me, it goes too far to discuss all possibilities and problems of bias of parameters that express MIDs in this paper (1). However, some short discussion has to be conducted and added about possibilities to reduce bias of MIDs in order to present a valuable and informative BMJ Rapid Recommendation in this theme, based on the information of (1,5) and possibly of other references.

RESPONSE: We agree and have highlighted this point prominently in the conclusions section as follows, including citations to the papers the reviewer suggests.

> "The review identified methodological limitations of the primary studies, future studies should strive for high precision of MID estimation, seek to identify difference between groups and reasons for those differences, and report correlations between the anchor and the PROM [8 10]."

Minor points:

4. The simple shoulder test (SST) has considerably low MIDs (Table 4, p. 25/50, line 17,45) when compared to all other reviewed scores: For example: SST: 1.8 score points, compared to the DASH: 10.2 and the Constant score: 8.3 etc. – all instruments are scaled by 0-100. Please comment and explain why. State hypotheses.

RESPONSE: Sorry, this is a mistake. The range of the SST should be 0-12. We corrected it accordingly.

5. The title needs the specification that only MIDs for improvement were reviewed.

RESPONSE: Revised as suggested.

New title: “Minimal important differences for improvement in shoulder condition patient-reported outcomes: a systematic review to inform a BMJ Rapid Recommendation”

6. There are two page numberings: n of 50 or m of 25. Please choose only one and number consistently.

RESPONSE: Revised as suggested.

7. At the end: Try to formulate a distinct take home message for the reader.

RESPONSE: The take home message, articulated in the conclusions section, highlights points that the reviewer has identified.

References
1 Angst F, Aeschlimann A, Angst J. The minimal clinically important difference (MCID) raised the significance of outcome effects above the statistical level, with methodological implications for future studies. J Clin Epidemiol 2017;82:128-36.
2 Revicki D, Hays RD, Cella DE, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. J Clin Epidemiol 2008;61:102-9.
3 Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. Controlled Clin Trials 1989;10:407-15.
4 Redelmeier DA, Lorig K. Assessing the clinical importance of symptomatic improvements. An illustration in rheumatology. Arch Intern Med 1993;153:1337-42.
5 Angst F, Benz T, Lehmann S, Aeschlimann A, Angst J. Multidimensional minimal clinically important differences (MCID) in knee osteoarthritis after comprehensive rehabilitation: a prospective evaluation from the Bad Zurzach Osteoarthritis Study. RMD open 2018. Oct 8;4(2):e000685. doi: 10.1136/rmdopen-2018-000685. eCollection 2018.

Reviewer: 3

Comments:
Thank you for the opportunity to review this interesting manuscript. The topic is of great interest and within the scope of the BMJ Rapid Recommendation concept. The manuscript is well-written. What I miss is a recommendation regarding the MID of the included PROMs or at least a detailed discussion of the heterogenicity of the results and the fact that there is no clearly defined MID for the PROMs in the field of shoulder surgery.

RESPONSE: We thank the reviewer for the positive evaluation of our paper. We see the median of the high credibility MIDs as the best estimate of the MIDs for each instrument and this is highlighted in Table 4. We have specified the possible sources of variability and noted our failure to identify clear associations between plausible determinants and the MID as a limitation of our work (highlighted in response to Reviewer 1 and 2). Given the limitations of the data, we are at a loss of how further discussion could shed additional light on this issue.

My specific concerns:

Abstract:

It is stated that the objective is “to identify credible anchor-based minimal important differences (MIDs) for patient reported outcome measures (PROMs) relevant to a BMJ Rapid Recommendations addressing subacromial decompression surgery for shoulder pain” and in the conclusion that “The MID estimates inform the interpretation for a linked systematic review and guideline addressing subacromial decompression surgery for shoulder pain. I get the impression that the authors only include studies of subacromial decompression and that the conclusion of the study is valid for this specific pathology only and not for cuff tear, glenohumeral osteoarthritis or humeral fractures. However, several eligible studies (table 3) is about shoulder arthroplasty and humeral fractures? Please clarify.

RESPONSE: We have clarified eligibility criteria (that were not restricted to subacromial compression in the revised manuscript) in the abstract as follows:

> “We included original studies of any intervention for shoulder conditions reporting estimates of anchor-based MIDs for relevant PROMs.”

We have further clarified that the results are valid beyond the subacromial compression syndrome in the abstract as follows:

> “The MID estimates inform the interpretation for a linked systematic review and guideline addressing subacromial decompression surgery for shoulder pain, and could also prove useful for authors addressing other interventions for shoulder problems.”

Background

The authors describe that “In this systematic survey, we (1) summarize MID estimates for the PROMs used in RCTs that investigate the effect of surgery on shoulder pain, and (2) assessed the credibility of these MID estimates.” However, some studies are not RCTs and not all studies include surgery? It is stated that “eligible studies used any design including retrospective and prospective observational studies…. (page 7 line 11). Please clarify.

RESPONSE: We have clarified in the background as follows:

> “A variety of study designs could inform MIDs for PROMs chosen by investigators for the RCTs. Therefore, in this systematic review, we (1) summarize MID estimate that come largely from observational studies for the PROMs chosen by the triallists in RCTs that investigated the effect of surgery on shoulder pain, and (2) assessed the credibility of these MID estimates.”

Method:

The Constant score is not exactly a patient-reported outcome, so why is this measurement tool included (It is recognized by the authors, page 6 line 47)? If the Constant score is included (the measurement tool with the highest number of papers) the paper is more about outcome measurement of shoulder pathologies and not patient-reported outcomes.

RESPONSE: The reviewer is quite right – the Constant score has components that are patient-reported and components that are not. We raised the same issue with our panel. However, the panel point out that the Constant score was the most common used tools in clinical trials on shoulder conditions. The guideline panel therefore asked us to address MIDs for the Constant score to interpret the magnitude of effects of interventions.

Discussion:

The authors acknowledge that the range of reported MIDs was wide and use the Constant Score as an example. For the absolute Constant score the Median MCI was 8.3 range 3 to 16.6, n=10 (high credibility). Furthermore, there is a wide range of pathologies from nonoperative treatment of impingement to shoulder arthroplasty for complex humeral fractures. The MIDs for these pathologies are, from a clinical perspective, not the same. Patients with impingement can often regain full range of motion and no pain whereas the aim of surgical treatment of complex humeral fractures is pain-relief (which only accounts for 15% of the maximum Constant Score). Thus, I am not convinced that an unweighted median value should be used to define the MID. Wouldn't it be better to use the MID that has been reported for specific treatment of specific pathologies such as shoulder arthroplasty for humeral fractures? I believe this should be discussed in more detail.

RESPONSE: Ideally, we should use the MIDs that have been reported for specific treatment of specific disease. However, the strict criteria would result in most of the studies being ineligible. We have included this point in the limitations section of the document as follows:

> "For others, MIDs for shoulder conditions closely related to subacromial syndrome, or for shoulder conditions at all, were not available, and we therefore relied on estimates from any upper extremity problem population."

Conclusion:
The authors state that "The MID estimates inform the interpretation for a linked systemtatic review and guideline on arthroscopy for shoulder pain." Again, the results are not specific for patients treated with arthroscopy. Please clarify.

RESPONSE: We have clarified, at several points, that the results may be applicable to a wide variety of shoulder conditions. In the conclusion, the statement is:

> "Researchers addressing a wide variety of shoulder conditions can in future make use of our summary MIDs to inform sample size and aid in interpretation of results."

References:

1. Roy JS, Macdermid JC, Woodhouse LJ. Measuring shoulder function: A systematic review of four questionnaires. *Arthritis Care and Research* 2009;61(5):623-32. doi: http://dx.doi.org/10.1002/art.24396
2. Tashjian RZ, Deloach J, Porucznik CA, et al. Minimal clinically important differences (MCID) and patient acceptable symptomatic state (PASS) for visual analog scales (VAS) measuring pain in patients treated for rotator cuff disease. *Journal of Shoulder and Elbow Surgery* 2009;18(6):927-32. doi: http://dx.doi.org/10.1016/j.jse.2009.03.021
3. St-Pierre C, Desmeules F, Dionne CE, et al. Psychometric properties of self-reported questionnaires for the evaluation of symptoms and functional limitations in individuals with rotator cuff disorders: a systematic review. *Disability and rehabilitation* 2016;38(2):103-22. doi: http://dx.doi.org/10.3109/09638288.2015.1027004
4. Copay AG, Chung AS, Eyberg B, et al. Minimum Clinically Important Difference: Current Trends in the Orthopaedic Literature, Part I: Upper Extremity: A Systematic Review. *JBJS Rev* 2018 doi: 10.2106/JBJS.RVW.17.00159 [published Online First: 2018/09/05]
5. Frahm Olsen M, Bjerre E, Hansen MD, et al. Minimum clinically important differences in chronic pain vary considerably by baseline pain and methodological factors: systematic review of empirical studies. *J Clin Epidemiol* 2018;101:87-106 e2. doi: 10.1016/j.jclinepi.2018.05.007 [published Online First: 2018/05/25]
6. Buchbinder R, Page MJ, Huang H, et al. A Preliminary Core Domain Set for Clinical Trials of Shoulder Disorders: A Report from the OMERACT 2016 Shoulder Core Outcome Set Special Interest Group. *J Rheumatol* 2017;44(12):1880-83. doi: 10.3899/jrheum.161123 [published Online First: 2017/01/17]
7. Guyatt GH. Making sense of quality-of-life data. *Med Care* 2000;38(9 Suppl):II175-9. [published Online First: 2000/09/12]
8. Angst F, Aeschlimann A, Angst J. The minimal clinically important difference raised the significance of outcome effects above the statistical level, with methodological implications for future studies. *J Clin Epidemiol* 2017;82:128-36. doi: 10.1016/j.jclinepi.2016.11.016 [published Online First: 2016/12/18]
9. Revicki D, Hays RD, Cella D, et al. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *J Clin Epidemiol* 2008;61(2):102-9. doi: 10.1016/j.jclinepi.2007.03.012 [published Online First: 2008/01/08]
10. Angst F, Benz T, Lehmann S, et al. Multidimensional minimal clinically important differences in knee osteoarthritis after comprehensive rehabilitation: a prospective evaluation from the Bad Zurzach Osteoarthritis Study. *RMD Open* 2018;4(2):e000685. doi: 10.1136/rmdopen-2018-000685 [published Online First: 2018/11/08]