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.

One of the three reviewers who reviewed this paper declined to publish his comment alongside with the article.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Application of Minimal Important Differences in Degenerative Knee Disease Outcomes: A Systematic Review and Case Study to Inform BMJ Rapid Recommendations
AUTHORS	Devji, Tahira; Guyatt, Gordon; Lytvyn, Lyubov; Brignardello-Petersen, Romina; Foroutan, Farid; Sadeghirad, Behnam; Buchbinder, Rachelle; Poolman, Rudolf; Harris, Ian; Carrasco Labra, Alonso; Siemieniuk, Reed; Vandvik, Per

VERSION 1 - REVIEW

REVIEWER	Xingzhong Jin Menzies Institute for Medical Research
REVIEW RETURNED	01-Nov-2016

GENERAL COMMENTS	I would like to thank the BMJ editors for the opportunity to review the systematic review entitled "Application of Minimal Important Differences in Degenerative Knee Disease Outcomes: A Systematic Review and Case Study to Inform BMJ Rapid Recommendations" by Dr Tahira Devji. The systematic review included 14 studies and reported anchor-based minimal important differences (MIDs) for 8 patient reported outcomes in knee osteoarthritis, including pain, function and quality of life. The current review is an update from a previous systematic review published in October 2015 and specifically provides important reference MIDs for a number of outcome instruments commonly used in clinical trials of knee osteoarthritis, such as WOMAC, KOOS and SF-36. My specific comments are below: • The authors used the term 'degenerative knee disease', which actually refers to 'knee osteoarthritis'. I suggest changing to the term 'knee osteoarthritis' throughout the manuscript, as the term is more commonly used in the rheumatology research community. • Credibility assessment: the length of follow-up for the MID estimation was added as an additional criterion and the longer the follow-up, the less credible was rated. Could the authors please provide explanations on this? Because knee osteoarthritis is a chronic disease, clinical trials usually are designed to follow-up for more than 6 months. Also, placebo effects may play an important role in measuring short-term patient reported outcomes, this is particularly true in knee pain assessment. • Results: The systematic review included 14 studies but the meta-

	analysis only pooled MIDs from high credibility studies. I suggest that doing a meta-analysis with all available MIDs and a further sensitivity analysis stratified by different credibility estimates would provide a complete picture of the data and a more robust conclusion for the review. • Results: Could the authors please describe the assessment of heterogeneity among included studies and explore the causes of the heterogeneity?

REVIEWER	Roos, Ewa Sports intsitute
REVIEW RETURNED	09-Nov-2016

GENERAL COMMENTS	COIs: Deputy editor of Osteoarthritis and Cartilage. Developer of the KOOS, KOOS-Child, HOOS, HAGOS, FAOS, RAOS, NOOS. Instruments have been developed in an academic context. Instruments are freely available and no license is required for their use, neither for academic or commercial use. No funding from commercial parties or non-profit organizations has been received for their development. Reviewer’s comments The aim of this submission was “to identify the most credible anchor-based minimal important differences (MIDs) for patient important outcomes in patients with degenerative knee disease (osteoarthritis), to inform BMJ Rapid Recommendations for arthroscopic surgery versus non-operative management”. This is a very timely aim, considering the increasing focus on treatment of patients with knee pain and osteoarthritis. The effort is also timely since the area of PRO interpretation is under development but still full of confusion, ambiguities and lack of consensus when it comes to both definitions and methods applied. This submission is obviously linked to another submission, a SR/MA, which is not available to me. This somewhat hampers my understanding of the full context. Unfortunately I have some serious doubts that the current submission will help improve our understanding of, and facilitate, PRO interpretation in the area of knee arthroscopy. My major concerns relate to 1) the immaturity of the field of PRO interpretation including the challenges of current methods; 2) the few reports of established cut-offs available for PROs in the area of interest, and 3) the wide interpretation of ‘degenerative knee disease’ and suggestion for use of cut-offs established in patient groups vastly different from that of interest, namely those having knee arthroscopy. Overall, it can be considered if it is premature to publish definite cut-points for use in interpretation of knee arthroscopy trials. I have outlined my concerns in more depth below. PRO interpretation and MIC methodology Interpretation of PROMs is not straight forward, and many methodological problems still need to be resolved. The current submission has focused on the anchor-based method for which lack of consensus of wording of the anchor question itself, and the wording and number of response items, may impact the consistency of results achieved. Therefore applying additional methods such as Cohen’s effect sizes and percentage change in RCTs could be very

relevant when interested in one specific context, namely knee arthroscopy. It could also be argued that the effect size from surgery should be higher than the well-established effect size from exercise of about 0.5 for the effect from surgery to be clinically relevant. Likewise, it could be argued that the percentage improvement from surgery should be around 50 (large effect) as compared to around 20 (small effect), an approach used in rheumatology (ACR 20 and ACR 50 criteria). Adding these alternative perspectives of interpretation may improve our understanding of this specific field.

For the anchor-based method current methodological challenges include the low correlations consistently found between anchors and scores, most possibly due to response shift and maybe also due the poorer of reliability of a single item anchor (this single item anchor is considered the ‘gold standard’ towards which the usually multi-item more reliable PROM score is compared). A recently reported simulation study found that when the variability of the PRO scores was held consistent, stronger anchors reduced the variability within each anchor category. Correlations of 0.3 were difficult to interpret and cut-offs considered deflated, while a correlation of at least 0.5 produced more easily interpretable results (Coon, CD. Telling the interpretation story: the case for strong anchors and multiple methods. ISOQOL 2016). You have downgraded the quality when correlations were lower than 0.3, which I absolutely agree with, I however question if the bar should be set higher.

To further highlight this problem I below refer to the study reporting KOOS cut-offs by Mills et al. of which I was a co-author, and therefore have more profound knowledge.

Few available studies
In the study by Mills et al. we used a correlation of 0.4 as our cut-off. We however only found correlations of 0.4 in specific subgroups for which we consequently reported MIDs. We clearly reported the lack of robustness in the results section of the abstract: “The methods used to calculate the MID influenced the cut-point, however the type of anchor question only influenced the MID when analysed with the mean change method. Depending on patient and clinical characteristics, the subscale and analysis approach used, the MID for KOOS improvement ranged from an absolute change of -1.5 to 13 points and worsening ranged from -11 to 5.5 points.” Considering our limited findings and previously mentioned immaturity of the field in general I feel very uncomfortable with you stating very definite cut-points for the KOOS from this only available paper (pain: 16.6, activities of daily living 8.2).

You state the MID for WOMAC function to be 6.5 and the MID for KOOS ADL to be 8.2. This is yet another example to how problematic this area is; WOMAC Function and KOOS ADL are identical outcomes which you would have known if you had compared the content of the respective PROMs or accessed the multitude of information available from references and internet resources where this is clearly outlined.

Patient group
Degenerative knee disease and knee OA are not identical terms and cannot be used synonymously. Actually, knee OA itself can be defined differently but most often include structural findings on radiographs, alone or in combination with pain. The term degenerative knee disease is not properly defined but considered an

	umbrella term thought to include also (earlier) stages of disease at which structural findings on imaging are not present. Of the available RCTs studying knee arthroscopy, two of the recent ones (Sihvonen et al. and Kise et al.) included physically quite active patients with a mean age below 50 with no radiographic OA. This is vastly different from the population in the Mills et al. paper, which had more advanced disease with a large proportion being waitlisted for total knee replacement. Since it is well known that cut-points are context-dependent I question generalizing cut-points derived in one population to another quite different population. Additional comments: You downgrade studies using a longer than 3 mo follow-up time. The reason for this is not stated and unclear to me. The primary follow-up varies between 6 and 24 months in the available knee arthroscopy studies, with at least the longer time points considered clinically relevant. From unpublished data I know that cut-points vary with time to follow-up after knee surgery. Considering 12-24 months usually considered being clinically relevant time points in orthopedics and the variability of cut-points in relation to time to follow-up, I question your suggested quality criteria in the context of knee arthroscopy. You prefer the mean change method over the ROC method. I agree the mean change method is more robust but your recommendation goes against recommendations by the COSMIN group and needs argumentation. The writing is imprecise, actually the majority of patients in knee arthroscopy RCTs were included is studies using an additional design (surgery+non-surgical management vs. non-surgical management alone). Only two studies (Østerås et al. and Kise et al.) compared surgery and a non-surgical intervention directly. Considering the smallest detectable change (SDC) often can be quite large I suggest it would be appropriate to contrast the cut-points to the SDC for the respective instruments.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1

I would like to thank the BMJ editors for the opportunity to review the systematic review entitled “Application of Minimal Important Differences in Degenerative Knee Disease Outcomes: A Systematic Review and Case Study to Inform BMJ Rapid Recommendations” by Dr Tahira Devji.

The systematic review included 14 studies and reported anchor-based minimal important differences (MIDs) for 8 patient reported outcomes in knee osteoarthritis, including pain, function and quality of life. The current review is an update from a previous systematic review published in October 2015 and specifically provides important reference MIDs for a number of outcome instruments commonly used in clinical trials of knee osteoarthritis, such as WOMAC, KOOS and SF-36. My specific comments are below:

Comment #1: The authors used the term 'degenerative knee disease', which actually refers to 'knee osteoarthritis'. I suggest changing to the term 'knee osteoarthritis' throughout the manuscript, as the term is more commonly used in the rheumatology research community.

Response: We have kept the terminology consistent with the linked papers submitted to the BMJ: 1) SR of treatment effects and 2) guideline, which use degenerative knee disease. Members of the panel, which included orthopaedic surgeons, rheumatologists, general practitioners, and physiotherapists agreed that the term 'degenerative knee disease' is more inclusive of patients who may be candidates for arthroscopic surgery.

Comment #2: Credibility assessment: the length of follow-up for the MID estimation was added as an additional criterion and the longer the follow-up, the less credible was rated. Could the authors please provide explanations on this? Because knee osteoarthritis is a chronic disease, clinical trials usually are designed to follow-up for more than 6 months. Also, placebo effects may play an important role in measuring short-term patient reported outcomes, this is particularly true in knee pain assessment.

Response: Initially, we considered longer follow-up for MID estimation as less credible due to concerns that the longer the time between the initial assessment and the follow-up assessment the more difficulty patients will have recalling their previous state of health when making ratings of change. However, there is currently little evidence to inform the extent to which this is indeed a problem, or the hiatus between assessments that severely compromises the change ratings. We have therefore jettisoned this criterion from our credibility assessment and focused instead on the reported correlation coefficients between the rating of change anchor and change in the PRO instrument score. An acceptable correlation suggests that patients had at least some recollection of their prior state. Whether the degree of change is due to placebo or natural history should not bear, as far as we can tell, on the credibility of the MID estimate.

Comment #3: Results: The systematic review included 14 studies but the meta-analysis only pooled MIDs from high credibility studies. I suggest that doing a meta-analysis with all available MIDs and a further sensitivity analysis stratified by different credibility estimates would provide a complete picture of the data and a more robust conclusion for the review.

Response: The reviewer's suggestion is very reasonable, and we would have followed the suggestion if we did not have a criterion that we think is indisputably important (the equivalent to restricting eligibility for a therapy question to randomized trials). As Reviewer 3 points out, basing an MID estimate on results in which the correlation is very low is likely to lead to uninterpretable (that is, invalid or not credible) results. Given that this is the case, including MID estimates with unequivocal low credibility seems to us at best questionable, and certainly inefficient. We would therefore argue for retaining our current approach, and focusing exclusively on MIDs with correlations $\geq$ 0.4.

We have, however, followed the reviewer's suggestion and explored the extent to which the possible credibility criteria we have suggested are associated with the MIDs, and described the results as follows:

"We only performed subgroup analyses exploring potential sources of heterogeneity for the WOMAC pain and function domains, as estimates for the KOOS pain and ADL, and EQ-5D came from a single study. Type of intervention (i.e. total knee arthroplasty [TKA] versus conservative management) was significantly associated with magnitude of the MID for both WOMAC pain (p<0.00001; figure 2) and function (p<0.00001; figure 3). For pain, the weighted pooled MID for TKA was 25 (95% CI 24 to 27) in TKA and for conservative management 8 (95% CI 3 to 13). For function the weighted pooled MID for TKA was 28 (95% CI 27 to 29), and for conservative management 19 (95% CI 3 to 17). We found no association between the hiatus between initial and follow-up visits, nor between the analytic method (ROC or mean change) and the MID"

Comment #4: Results: Could the authors please describe the assessment of heterogeneity among included studies and explore the causes of the heterogeneity?

Response: As we have noted in the previous response, we have performed subgroup analyses to explore potential sources of heterogeneity and, in the revised manuscript, report results as described in the previous response.

Reviewer: 2

The aim of this submission was "to identify the most credible anchor-based minimal important differences (MIDs) for patient important outcomes in patients with degenerative knee disease (osteoarthritis), to inform BMJ Rapid Recommendations for arthroscopic surgery versus non-operative management".

This is a very timely aim, considering the increasing focus on treatment of patients with knee pain and osteoarthritis. The effort is also timely since the area of PRO interpretation is under development but still full of confusion, ambiguities and lack of consensus when it comes to both definitions and methods applied.

This submission is obviously linked to another submission, a SR/MA, which is not available to me. This somewhat hampers my understanding of the full context.

Unfortunately I have some serious doubts that the current submission will help improve our understanding of, and facilitate, PRO interpretation in the area of knee arthroscopy. My major concerns relate to 1) the immaturity of the field of PRO interpretation including the challenges of current methods; 2) the few reports of established cut-offs available for PROs in the area of interest, and 3) the wide interpretation of ‘degenerative knee disease’ and suggestion for use of cut-offs established in patient groups vastly different from that of interest, namely those having knee arthroscopy. Overall, it can be considered if it is premature to publish definite cut-points for use in interpretation of knee arthroscopy trials. I have outlined my concerns in more depth below.

Comment #1: Interpretation of PROMs is not straight forward, and many methodological problems still need to be resolved. The current submission has focused on the anchor-based method for which lack of consensus of wording of the anchor question itself, and the wording and number of response items, may impact the consistency of results achieved. Therefore applying additional methods such as Cohen’s effect sizes and percentage change in RCTs could be very relevant when interested in one specific context, namely knee arthroscopy. It could also be argued that the effect size from surgery should be higher than the well-established effect size from exercise of about 0.5 for the effect from surgery to be clinically relevant. Likewise, it could be argued that the percentage improvement from surgery should be around 50 (large effect) as compared to around 20 (small effect), an approach used in rheumatology (ACR 20 and ACR 50 criteria). Adding these alternative perspectives of interpretation may improve our understanding of this specific field.

Response: We apologize that the linked systematic review and the BMJ Rapid Recommendation article was not available to the peer-reviewers. The BMJ Editorial team responsible for the peer-review of this package of information has also recognized this as a limitation. For future Rapid Recommendations linked papers will be available for peer-reviewers to view if they would like to.

The reviewer makes two good points. The first has to do with alternative ways of enhancing the interpretability of the instruments used as outcomes in the relevant studies. We have dealt with this issue in responding to comments #1 and #6 from the editorial team and the following paragraph in the discussion.

“The following considerations mitigate the concerns regarding the credibility of the MID estimates that guided the panel’s recommendation. First, our best estimates of the MID approximate 10% of the instruments total range, a value that is both intuitive and consistent with MID estimates for other instruments. Second, our best estimates of the MID are consistent with the experience of clinicians who have used the instruments as part of their clinical practice. Third, estimates for the risk difference in proportion improved with arthroscopy from the sensitivity analyses in the linked systematic review show that using the upper and lower boundaries of the MID that we have suggested, and a value based on the standardized mean difference, approximate those using our best estimate of the MID[4] [12].”

The reviewer’s second point has to do with whether the MID may differ depending on the intervention. We have dealt with this issue in response to Reviewer 2’s 8th comment.

Comment #2: For the anchor-based method current methodological challenges include the low correlations consistently found between anchors and scores, most possibly due to response shift and maybe also due the poorer of reliability of a single item anchor (this single item anchor is considered the ‘gold standard’ towards which the usually multi-item more reliable PROM score is compared). A recently reported simulation study found that when the variability of the PRO scores was held consistent, stronger anchors reduced the variability within each anchor category. Correlations of 0.3 were difficult to interpret and cut-offs considered deflated, while a correlation of at least 0.5 produced more easily interpretable results (Coon, CD. Telling the interpretation story: the case for strong anchors and multiple methods. ISOQOL 2016). You have downgraded the quality when correlations were lower than 0.3, which I absolutely agree with, I however question if the bar should be set higher.

To further highlight this problem I below refer to the study reporting KOOS cut-offs by Mills et al. of which I was a co-author, and therefore have more profound knowledge.

Response: We have taken note of these excellent points. As described in responses to previous comments we have implemented the approach implied in the reviewer’s comments, and have revised our credibility assessments accordingly to set the minimum acceptable correlation coefficient threshold that the reviewer suggests in comment #3, 0.4.

Comment #3: <u>Few available studies</u>

In the study by Mills et al. we used a correlation of 0.4 as our cut-off. We however only found correlations of 0.4 in specific subgroups for which we consequently reported MIDs. We clearly reported the lack of robustness in the results section of the abstract: “The methods used to calculate the MID influenced the cut-point, however the type of anchor question only influenced the MID when analysed with the mean change method. Depending on patient and clinical characteristics, the subscale and analysis approach used, the MID for KOOS improvement ranged from an absolute change of -1.5 to 13 points and worsening ranged from -11 to 5.5 points.” Considering our limited findings and previously mentioned immaturity of the field in general I feel very uncomfortable with you stating very definite cut-points for the KOOS from this only available paper (pain: 16.6, activities of daily living 8.2).

Response: We have taken the reviewer’s concerns into consideration and modified our approach to identifying credible MIDs. Instead of selecting a definitive cut-point for the candidate PROs, we present the median absolute MID estimate, along with the minimum and maximum values across a range of plausible trustworthy MIDs. We provided these estimates to the systematic review team conducting the linked meta-analysis of treatment effects for arthroscopy. Sensitivity analyses demonstrated that the results were robust even after accounting for potential uncertainties in the MID (indeed, the ranges led to lower estimates of the proportion benefiting from surgery, reinforcing our inference that effects are small or very small).

Comment #4: You state the MID for WOMAC function to be 6.5 and the MID for KOOS ADL to be 8.2. This is yet another example to how problematic this area is; WOMAC Function and KOOS ADL are identical outcomes which you would have known if you had compared the content of the respective PROMs or accessed the multitude of information available from references and internet resources where this is clearly outlined.

Response: In this case we would make a different inference from the reviewer: 6.5 and 8.2 strike us as very similar. In any case, we still agree with the reviewer's fundamental point that focusing on a single estimate of the MID is problematic, that suggesting a range in which the MID might actually lie is preferable, and acknowledging that the range then mandates the sensitivity analysis conducted in the linked review.

Comment #5: Patient group

Degenerative knee disease and knee OA are not identical terms and cannot be used synonymously. Actually, knee OA itself can be defined differently but most often include structural findings on radiographs, alone or in combination with pain. The term degenerative knee disease is not properly defined but considered an umbrella term thought to include also (earlier) stages of disease at which structural findings on imaging are not present. Of the available RCTs studying knee arthroscopy, two of the recent ones (Sihvonen et al. and Kise et al.) included physically quite active patients with a mean age below 50 with no radiographic OA. This is vastly different from the population in the Mills et al. paper, which had more advanced disease with a large proportion being waitlisted for total knee replacement. Since it is well known that cut-points are context-dependent I question generalizing cut-points derived in one population to another quite different population.

Response: We have addressed this in response to reviewer 1's comment #1 and the editors' comment #5, as well as in response to this reviewer's comment #3.

Additional comments:

Comment #6: You downgrade studies using a longer than 3 mo follow-up time. The reason for this is not stated and unclear to me. The primary follow-up varies between 6 and 24 months in the available knee arthroscopy studies, with at least the longer time points considered clinically relevant. From unpublished data I know that cut-points vary with time to follow-up after knee surgery. Considering 12-24 months usually considered being clinically rele vant time points in orthopedics and the variability of cut-points in relation to time to follow-up, I question your suggested quality criteria in the context of knee arthroscopy.

Response: Addressed above in reviewer 1's second comment.

Comment #7: You prefer the mean change method over the ROC method. I agree the mean change method is more robust but your recommendation goes against recommendations by the COSMIN group and needs argumentation.

Response: Addressed above in reviewer 2's comment #4.

Comment #8: The writing is imprecise, actually the majority of patients in knee arthroscopy RCTs were included is studies using an additional design (surgery+non-surgical management vs. non-surgical management alone). Only two studies (Østerås et al. and Kise et al.) compared surgery and a non-surgical intervention directly.

Response: We have revised this throughout the manuscript

Comment #9: Considering the smallest detectable change (SDC) often can be quite large I suggest it would be appropriate to contrast the cut-points to the SDC for the respective instruments.

Response: Although of potential interest, we believe this particular exploration would be tangential to the purpose of the current manuscript.