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.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Minimally important difference estimates and methods: A protocol
AUTHORS	Johnston, Bradley; Ebrahim, Shanil; Carrasco Labra, Alonso; Furukawa, Toshiaki; Patrick, Donald; Crawford, Mark; Hemmelgarn, Brenda; Schunemann, Holger; Guyatt, Gordon; Nesrallah, Gihad

VERSION 1 - REVIEW

REVIEWER	Pablo Martinez-Martin National Center of Epidemiology Carlos III Institute of Health
REVIEW RETURNED	27-May-2015

GENERAL COMMENTS	This manuscript is aimed at presenting the protocol of a study focused on anchor-based methods for estimation of the Minimally Important Difference (MID) for Patient-Reported Outcomes (PROs) included as variables of interest in longitudinal and intervention studies. The authors justify appropriately the effort and this project is very welcome as the methodology in this field is varied and not standardized, with different methods offering different results. This is a serious attempt –the first one, to my knowledge – to introduce order and rationality into the topic and the authors will use evidence-based tools to this purpose. Several objectives are specified, being the first a “systematic survey addressing the anchor-based methods used to estimate MIDs”, followed by the development of an instrument to assess the credibility of the MIDs and its application to those MIDs obtained a systematic survey of published anchor-based MIDs from chronic medical and psychiatric disorders. Finally, the inter-rater reliability of this instrument will be tested. The methods and procedure to achieve the objectives planned are displayed in this manuscript and are adequate. Some of the authors have an extensive recognized experience in the development and aplication of this kind of methodology for this sort of studies. The Discussion section refers to expectations and some foreseen difficulties, as well as the advanced contibutions this project can furnish. Therefore, the protocol seems satisfactory for reaching appropriate objectives related with an important issue reflected in clinical research and practice. There is point probably needing rewording to be more understandable. In the Page 5, lines 21 to 26, it is read: “Knowledge of the MID allows interpretation of the magnitude of effect and thus the trade-off between beneficial and harmful outcomes.” and would be more comprehensible as: “Knowledge of the MID allows interpretation of the magnitude of effect and thus the trade-off between beneficial or harmful outcomes and negligible outcomes.” Finally, although is true that anchor-based methods are the preferred way to determine the MID as they incorporate the

	"patients' preferences and values" (page 5-6), it would convenient here to indicate that are not free of problems adding some examples (e.g., adequacy of the selected anchor; recall of the baseline condition,etc.).

REVIEWER	Caroline Terwee Department of Epidemiology and Biostatistics VU University Medical Center Amsterdam the Netherlands
REVIEW RETURNED	02-Jun-2015

GENERAL COMMENTS	This is a protocol of a very interesting study with multiple aims: 1) to review the methods used for estimating Minimal Important Differences (MID) values of Patient-Reported Outcomes (PROs), to develop as instrument to assess the credibility of MID studies. 2) to document published MID values of PRO instruments. The study is likely to yield important and useful results for a wide audience. However, I do have some concerns. Comments: • In the introduction, an example is given to illustrate the importance of interpretability of PRO instruments. The example refers to an improvement after an intervention by 3 points relative to control. The example is a bit confusing because this difference of 3 points is not the same as the MID, as the MID refers to a change within a group, not a difference between groups. This may be confusing for readers. • The remark that "the MID also provides a metric for clinical trialists planning sample sizes for their studies" needs further explanation because in a clinical trial the focus is on a difference in change between groups, while the MID refers to a change within a group. It would be helpful to add an explanation of how the MID can be used in sample size calculations for clinical trials. • My main concern regarding this study is whether it will be possible to develop a valid standard for MID studies, based on a review of the literature. The search will yield a number of publications on MID methodology, but all of these methods are opinion-based, rather than evidence-based. If strengths and limitations of the methods will be described in the papers, they will reflect the opinion of the authors of the papers. We don't know whether these strengths are associated with more valid MID estimates, not that the limitations are associated with biased MID estimates. So it will not be possible to develop an evidence-based standard for MID studies. The alternative would be to develop a consensus-based standard. That would require for example a Delphi study like the COSMIN group has done for developing standards for studies on measurement properties. However, the COSMIN group decided at that time (around 2005) that there was not enough consensus on MID methodology to develop such a standard. I doubt whether there is such consensus at this time. Nevertheless, I recommend to include somehow input from authors who published on MID methods in the development of the standard. • Another important concern is the sensitivity and comprehensiveness of the search strategy. The search strategy seems not logical. It is very unclear what each block of search terms represents, why the search is built as described, and whether the total search is comprehensive. A more structured approach would

	be to build three blocks of search terms: (1) search terms for MID; (2) search terms for PRO instruments; and (3) search terms for the time interval; and then combine these blocks with AND. For MID, I recommend to use part of the validated search filter for finding studies on measurement properties (Terwee et al, Qual Life Res 2009). The relevant terms from this filter are: (interpretab*[tiab] OR ((minimal[tiab] OR minimally[tiab] OR clinical[tiab] OR clinically[tiab]) AND (important[tiab] OR significant[tiab] OR detectable[tiab]) AND (change[tiab] OR difference[tiab])) OR (small*[tiab] AND (real[tiab] OR detectable[tiab]) AND (change[tiab] OR difference[tiab])) OR "meaningful change"[tiab]). For PRO instruments I recommend to use a search filter developed by the University of Oxford, specifically for this purpose. The filter can be found on the COSMIN website: http://www.cosmin.nl/images/upload/files/PROM%20Gp%20filtersOCTOBER%202010FINAL.pdf • I have some detailed comments on the data abstraction form: o Item 8 is also relevant for cohort studies, not just experimental studies o Item 20: you also need to know if the scores are calculated per domain or for the instrument in total o Item 22 seems irrelevant because it is an inclusion criterion o Item 23 is vague: what exactly do you want to know? E.g. mean change in 1 group, ROC method, etc.. o Item 24a seems irrelevant because it is an inclusion criterion o Item24b: relative to what? o Item 26-27: why is it relevant to estimate MID values for males and females separately? o Item 28 is vague: what information about the anchor are you extracting? o Item 31a: also the number of response options of the anchor is relevant o Item 32: you also need to know whether the anchor was domain-specific or generic o Item34 is not clear: the PRO measures will probably be administered multiple times and the anchor only at follow-up o Overall, I think the form needs more consideration and perhaps input from MID experts. • For assessing the inter-rater reliability of the credibility instrument, I recommend to also calculate a parameter of measurement error, e.g. the limits of agreement, to evaluate the magnitude of differences between raters.

VERSION 1 – AUTHOR RESPONSE

Reviewer 1:

This manuscript is aimed at presenting the protocol of a study focused on anchor-based methods for estimation of the Minimally Important Difference (MID) for Patient-Reported Outcomes (PROs) included as variables of interest in longitudinal and intervention studies. The authors justify appropriately the effort and this project is very welcome as the methodology in this field is varied and not standardized, with different methods offering different results. This is a serious attempt –the first one, to my knowledge – to introduce order and rationality into the topic and the authors will use evidence-based tools to this purpose. Several objectives are specified, being the first a “systematic survey addressing the anchor-based methods used to estimate MIDs”, followed by the development of an instrument to assess the credibility of the MIDs and its application to those MIDs obtained a systematic survey of published anchor-based MIDs from chronic medical and psychiatric disorders.

Finally, the inter-rater reliability of this instrument will be tested.

The methods and procedure to achieve the objectives planned are displayed in this manuscript and are adequate. Some of the authors have an extensive recognized experience in the development and application of this kind of methodology for this sort of studies. The Discussion section refers to expectations and some foreseen difficulties, as well as the advanced contributions this project can furnish.

Therefore, the protocol seems satisfactory for reaching appropriate objectives related with an important issue reflected in clinical research and practice.

[1] There is point probably needing rewording to be more understandable. In the Page 5, lines 21 to 26, it is read: “Knowledge of the MID allows interpretation of the magnitude of effect and thus the trade-off between beneficial and harmful outcomes.” and would be more comprehensible as: “Knowledge of the MID allows interpretation of the magnitude of effect and thus the trade-off between beneficial or harmful outcomes and negligible outcomes.”

Response: Thank you for your suggestion. Our current statement states the trade-off between one extreme of benefit and the other extreme of harm. Negligible or non-important outcomes are inherently in between these two outcomes. We have revised our statement as follows, “Knowledge of the MID allows decision-makers to better interpret the magnitude of treatment effect and assess the trade-off between beneficial and harmful outcomes.”

[2] Finally, although is true that anchor-based methods are the preferred way to determine the MID as they incorporate the “patients’ preferences and values” (page 5-6), it would convenient here to indicate that are not free of problems adding some examples (e.g., adequacy of the selected anchor; recall of the baseline condition,etc.).

Response: Thank you for your suggestion. We have revised it as follows, “It is generally agreed that the patient-reported anchor-based approach is the optimal way to determine the MID because it directly captures the patients' preferences and values, although it can still be problematic if the credibility of the anchor is in question (e.g. is the anchor itself interpretable and are responses on the anchor independent of responses on the PRO).”

Reviewer 2:

This is a protocol of a very interesting study with multiple aims: 1) to review the methods used for estimating Minimal Important Differences (MID) values of Patient-Reported Outcomes (PROs), to develop as instrument to assess the credibility of MID studies. 2) to document published MID values of PRO instruments. The study is likely to yield important and useful results for a wide audience. However, I do have some concerns.

[3] In the introduction, an example is given to illustrate the importance of interpretability of PRO instruments. The example refers to an improvement after an intervention by 3 points relative to control. The example is a bit confusing because this difference of 3 points is not the same as the MID, as the MID refers to a change within a group, not a difference between groups. This may be confusing for readers.

Response: Thank you. Although the MID does refer to within-individual change (or an average thereof), the estimate is then applied to between-group differences.

[4] The remark that “the MID also provides a metric for clinical trialists planning sample sizes for their studies” needs further explanation because in a clinical trial the focus is on a difference in change between groups, while the MID refers to a change within a group. It would be helpful to add an explanation of how the MID can be used in sample size calculations for clinical trials.

Response: Thank you for your suggestion. We have added the following statement in our manuscript, “The MID also provides a metric for clinical trialists planning sample sizes for their studies. This is accomplished by first calculating the proportion of patients achieving an MID or greater change, and subsequently determining the difference in the proportion of responders trialists would like to examine between the treatment and control that would constitute a clinically important difference.”

[5] My main concern regarding this study is whether it will be possible to develop a valid standard for MID studies, based on a review of the literature. The search will yield a number of publications on MID methodology, but all of these methods are opinion-based, rather than evidence-based. If strengths and limitations of the methods will be described in the papers, they will reflect the opinion of the authors of the papers. We don’t know whether these strengths are associated with more valid MID estimates, not that the limitations are associated with biased MID estimates. So it will not be possible to develop an evidence-based standard for MID studies. The alternative would be to develop a consensus-based standard. That would require for example a Delphi study like the COSMIN group has done for developing standards for studies on measurement properties. However, the COSMIN group decided at that time (around 2005) that there was not enough consensus on MID methodology to develop such a standard. I doubt whether there is such consensus at this time. Nevertheless, I recommend to include somehow input from authors who published on MID methods in the development of the standard.

Response: Thank you for your suggestion. Our group consists of individuals who have led multiple methods studies on MIDs published widely in the field, in fact co-developed the concept of the MID, and are considered experts in this domain. In developing our credibility tool, we will use both the survey of the methods literature as well as input from the experts in our group. We have used this approach previously. We have added the following statement to our manuscript to reflect this, “Based on the survey of the methods literature and our group’s experience with methods of ascertaining MIDs, we will develop initial criteria for evaluating the credibility of anchor-based MID determinations”

[6] Another important concern is the sensitivity and comprehensiveness of the search strategy. The search strategy seems not logical. It is very unclear what each block of search terms represents, why the search is built as described, and whether the total search is comprehensive. A more structured approach would be to build three blocks of search terms: (1) search terms for MID; (2) search terms for PRO instruments; and (3) search terms for the time interval; and then combine these blocks with AND. For MID, I recommend to use part of the validated search filter for finding studies on measurement properties (Terwee et al, Qual Life Res 2009). The relevant terms from this filter are: (interpretab*[tiab] OR ((minimal[tiab] OR minimally[tiab] OR clinical[tiab] OR clinically[tiab]) AND (important[tiab] OR significant[tiab] OR detectable[tiab]) AND (change[tiab] OR difference[tiab])) OR (small*[tiab] AND (real[tiab] OR detectable[tiab]) AND (change[tiab] OR difference[tiab])) OR "meaningful change"[tiab]). For PRO instruments I recommend to use a search filter developed by the University of Oxford, specifically for this purpose. The filter can be found on the COSMIN website: http://www.cosmin.nl/images/upload/files/PROM%20Gp%20filtersOCTOBER%202010FINAL.pdf

Response: Thank you. In consultation with a group of experienced methodologists in PRO development and interpretation, our search criteria were developed by an experienced academic librarian. As well, our search strategy was reviewed and validated by an independent academic librarian with expertise in methodological studies. Further, we selected a random sample of studies

and cross-checked these studies with our search results to ensure that it was comprehensive and sensitive in identifying potentially eligible studies. We were able to identify all studies in our random sample.

[7] I have some detailed comments on the data abstraction form:
• Item 8 is also relevant for cohort studies, not just experimental studies
• Item 20: you also need to know if the scores are calculated per domain or for the instrument in total
• Item 22 seems irrelevant because it is an inclusion criterion
• Item 23 is vague: what exactly do you want to know? E.g. mean change in 1 group, ROC method, etc..
• Item 24a seems irrelevant because it is an inclusion criterion
• Item24b: relative to what?
• Item 26-27: why is it relevant to estimate MID values for males and females separately?
• Item 28 is vague: what information about the anchor are you extracting?
• Item 31a: also the number of response options of the anchor is relevant
• Item 32: you also need to know whether the anchor was domain-specific or generic
• Item34 is not clear: the PRO measures will probably be administered multiple times and the anchor only at follow-up

Overall, I think the form needs more consideration and perhaps input from MID experts.

Response: Thank you for your suggestions. We have made the revisions listed below. In addition, before beginning our study, we will pilot our data extraction forms, and forms for assessing the credibility if MIDs, with a group of independent experts in MID methods (see page 11: Using a sample of eligible studies, we will pilot the draft instrument with 4 target users, specifically, researchers interested in the credibility of MID estimates, who will be identified within our international network of knowledge users (please see 'Knowledge Translation' section below). The data collected at this stage will inform item modification and reduction. This iterative process will be conducted until we achieve consensus for the final version of the instrument.)

• Item 8: Please state the study interventions (if applicable)
• We have not revised item 20, but instead item 25 as follows: "Please report the type of precision estimate of the MID reported (Please list domain and total, if applicable)"
• Item 22: We have removed this item.
• Item 23: "Describe the methods used to determine the anchor-based MID? For example, ROC method, change in scores corresponding to range or specific scores on the anchor"
• Item 24a: We have removed this item.
• Item 24b: "Indicate whether the authors report a relative change or absolute change of the MID or both?"
• Item 26-27: Some studies have shown important differences in MID between men and women, we have included this to capture this gender effect. This is also a requirement from our sponsor, the Canadian Institute of Health Research, to explore the effect of gender in studies when feasible.
• Item 28: "Please specify the anchor-based instrument that was used?" Otherwise, the characteristics the anchor-based instruments are addressed in question 29 to 33 (e.g. number of items, scoring, type of response options).
• With respect to the reviewer's comment for 31a, this is captured in item 30 as follows, "What is the range of scores or the response options on the anchor instrument?"
• Item 22 is now "Is the anchor instrument domain-specific or generic?"
• Item 34: "Was the anchor instrument administered at the same time as the administration of the PRO?"

[8] For assessing the inter-rater reliability of the credibility instrument, I recommend to also calculate a parameter of measurement error, e.g. the limits of agreement, to evaluate the magnitude of differences between raters.

Response: Thank you for your suggestion. We have revised this as follows, “We will conduct a reliability study of our instrument to measure the credibility of MIDs calculating the inter-rater reliability and associated 95% confidence interval as measured by weighted kappa with quadratic weights.”