PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	A unified multi-level model approach to assessing patient responsiveness including; return to normal, minimally important differences, and minimal clinical important improvement for patient reported outcome measures.
AUTHORS	Sayers, Adrian; Wylde, Vikki; Lenguerrand, Erik; Gooberman-Hill, Rachael; Dawson, Jill; Beard, David; Price, Andrew; Blom, Ashley

VERSION 1 - REVIEW

REVIEWER	Gillian Hawker University of Toronto, Canada I have published previously with Rachael Gooberman-Hill. No other disclosures.
REVIEW RETURNED	30-Sep-2016

GENERAL COMMENTS	This manuscript describes an alternative statistical approach to assessment of responsiveness to joint replacement surgery and compares this approach to currently used standards. The comparative effects of the different approaches to assessment of patient response are illustrated using data from an existing joint replacement cohort. The issue of how best to define a good outcome following joint replacement, and its determinants, is important. I have no methodological concerns. My major concern is the appropriateness of this paper for BMJ Open and its readability by a non-statistical audience. My specific comments are as follows: 1. If the audience is a general medical one, I think greater emphasis is needed on the following: a) what is the problem the authors are addressing? what are the problems with the current approaches? how will the MLM approach be used to advance patient care and outcomes? b) while this is a largely a statistical exercise / methodological paper, for a general readership it might be more reader-friendly to let the cohort data illustrate the points being made much more than is done in the current paper. For example, the tables are difficult to interpret - could they be formatted to better illustrate what effectively amounts to variability in the 'error' estimates across methods, which results in fewer or more patients being classified as 'responders'? 2. There is no mention of PASS - patient acceptable symptom state - as an indicator of a successful treatment outcome - might this not be considered the ultimate gold standard? Please discuss. 3. Page 17 first paragraph of results section: states that current

	methods overestimate SD of baseline and change scores but then states that SD estimates were greater using MLM - have I misunderstood here? Conceptually, I fully understand if we overestimate SDs we would underestimate the proportion of responders... 4. Re the point above, is the bottom line that MLM provides tighter estimates and thus will be less likely to misclassify a responder as a non-responder? If so, the currently used methods simply gives a worst case scenario? Given the complexity of the MLM approach, could you please comment on its usefulness in informing patient decision making? 5. As you know, we have previously examined % with good outcome using MID, MCID, OO (Arthritis Rheum. 2013 May;65(5):1243-52) - is there more that could be said about the current methods and comparisons with other papers using these metrics - in other words, do you have any recommendations regarding which of these current approaches provides results that are closest to MLM?

REVIEWER	Nicole Pratt University of South Australia, Australia
REVIEW RETURNED	16-Nov-2016

GENERAL COMMENTS	Thank you for the opportunity to review the paper "A unified multi-level model approach to assessing patient responsiveness including; return to normal, minimally important differences, and minimally clinical important differences for patient reported outcome measures". This is a well written paper that demonstrates a clear application to a common problem. The authors have clearly shown that there is an advantage to MSM over standard methods which ignore the time between first and last observation. The main finding appears to be an variance benefit and hence a potentially more accurate prediction. I have a few minor comments only. 1) The authors say that the standard approach 'overestimates' the variance. Do they have a basis for this statement? It may be that the MSM 'underestimates' variance. A simulation study would help to solve this issue. 2) In the discussion the authors state that model diagnostics are important however I do not believe they were presented for their models. 3) the discussion section is rather brief. For example different results were found between Hip and Knee analyses e.g. variance much lower for the knee analysis 4) the tables could be improved. It appears that the data presented are not in line and this makes the results difficult to read. e.g. is there an extra number in the last column of the first block in Table 1 (67.1) or too few numbers in the same position in Table 2?. It would be helpful to footnote what each of the acronyms means in the tables.

REVIEWER	Anne Thackeray University of Utah, USA
REVIEW RETURNED	16-Feb-2017

GENERAL COMMENTS	The authors present a well-written approach to better estimation of assessing patient responsiveness and provide detailed information on the research methods. My primary recommendations are directed at clarifying sections of the manuscript as outlined below: ABSTRACT: Objective would be clearer by indicating this is not only a review but a comparison of the techniques. METHODS: Participant consent, ethics approval: not explicitly stated (implied with reference to the APEX cohort study)- may want to explicitly state Indicate how satisfaction was measured and reference specifically anchor for pain measurement use with MCID estimates. It is a bit unclear how many time points were used in the MLM approach. Table and lines 21-33 (p16) suggest just baseline to 3 months. However, when referring to the MLM approach ,"all individuals are measured at exactly 0,3,6,12 months"– p.16 lines 40-41. Please clarify. Please indicate how the model with two splines (and know point at 3 months) was determined to be the best fit. (p16 lines 41-45) Reference results tables in results section (P 16, Lines 45-47) RESULTS: Clarify statements in the first paragraph (lines 12-17). You indicate MLM provides a better estimate of SD, but these statements suggest the SD is greater than conventional methods. To help determine the relative strength of the MLM approach, please include confidence intervals around % responders (Tables 1 and 2) and mean differences with confidence intervals between proportions of responders using current v. MLM approaches. Tables: write out P(Resp.)

REVIEWER	Kathryn Mills Macquarie University, Australia
REVIEW RETURNED	16-Feb-2017

GENERAL COMMENTS	Comments Thank you for the opportunity to review this manuscript. With the current debates over the optimal methods to define clinically meaningful improvement in patient reported outcomes, this manuscript provides another mathematical approach. In writing this manuscript, the authors have undertaken an ambitious task. Critiquing current methods for establishing patient important change and proposing another is a large task to undertake within a single manuscript. At a basic level, the manuscript achieves part of its objective: it demonstrates how MLM can be applied to four approaches that are utilised to determine clinical meaning in different contexts and study designs. However, the critique of the four approaches brief and the result is that the authors have under reported the complexities of the approaches including the different context within which they are typically applied and how error/confidence intervals are currently calculated within the

	approaches. In the methods, there is a paragraph that speaks to the benefits of MLM approaches (Pages 10-11, Lines 57-13), but the authors do not contrast these strengths with limitation of typical approaches to patient important change. The rationale why the MLM may be appealing to researchers is confused throughout the paper. The authors propose that the MLM may be used to predict the chances of a patient experiencing a treatment effect (which would be very useful, but not the same as retrospectively defining treatment effect/response). They also propose that it reduces the overestimation of the threshold the classify a “responder”. While it is possible that the MLM does both, jumping between these rationales and only demonstrating the changes on the clinically meaningful thresholds is confusing for the reader. Can the authors please clarify whether the use of the MLM is best for prediction of change or clarification that the change is meaningful? The authors need to provide a stronger rationale for why the MLM should be considered. In the results section the authors demonstrate how the application of MLM reduces the threshold of clinically meaningful change. While this demonstrates that more individuals may have improved, the authors may consider comparing their proposed threshold to the minimal detectible change for the ICOAP. A major issue, particularly in MID research, is when the MID is less than the MDC. This renders the MID somewhat useless in the clinical setting. It would interesting to see the authors consider this in the current manuscript. As the authors acknowledge, MLMs are complex and require multiple alterations to cope with (1) non-linear trajectories (2) heterogeneity within the trajectory classes or (3) multiple time points. Both of these factors are not only expected, but they are common in the OA process. Could the authors please comment further on the clinical utility of their proposed methods? Specific comments Page 5: Line 39; While the argument the authors put forward on regarding the limitation of statistical models being used to quantify patient change is accurate, there have been far more elegant and thorough explanations of this limitation published that the authors may want to substitute e.g. King, M. T. (2011). A point of minimal important difference (MID): a critique of terminology and methods. Expert Review of Pharmacoeconomics & Outcomes Research, 11(2), 171–184. http://doi.org/10.1586/erp.11.9 Page 5-6: Lines 57-7: Could the authors please provide a rationale the distribution based methods have been identified as “minimal important difference” and the anchor-based methods have been identified as “minimal clinically important differences”? Recommendations by numerous authors (e.g., Revicki, D., Hays, R. D., Cella, D., & Sloan, J. (2008). Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. Journal of Clinical Epidemiology, 61(2), 102–109. http://doi.org/10.1016/j.jclinepi.2007.03.012) indicate that both methods are called MID. Page 7: Line 13 onwards: This section does not fit in the methodology section as it is not outlining a process of

	tasks/calculation etc undertaken. In the previous paragraph, the authors state that as part of the methods they will outline common methods used to describe response. This would imply that the explanation belongs in the results section. However, on reading this explanation it is not the result of a systematic search strategy and includes a critique of the literature, not just an outline, that is integral for the authors rationale for the new proposed method. Thus, this would suggest that this section of the manuscript belongs in the introduction. Could the authors please move the explanation section of the manuscript to a more appropriate subheading? Page7: Line 42-48; First, can the authors please provide a reference for this explanation of the RCI. Second, a reader could find the last sentences in this paragraph confusing. Particularly the phrase "reliability values in the spirit of a sensitivity analysis". Given the precise nature of reliability studies, using "spirit" is an inappropriate word choice here. Could the authors either remove this part of the sentence, or provide a brief example to the reader on how where such values could be drawn from? Page 7: Line 55; The authors state that the 0.5 SD is of the pre-post-surgery change scores. Revicki et al., (2008) is then cited to support this. However, the 0.5 SD is of baseline score only. In their 2008 article, Revicki et al., cite Normal et al., (Norman, G. R., Sloan, J. A., & Wyrwich, K. W. (2003). Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. Medical Care, 41(5), 582–592. http://doi.org/10.1097/01.MLR.0000062554.74615.4C) This article clearly states that only baseline scores contribute to the distribution model. Can the authors please change their paragraph to reflect this? Page 8: Line 12-26; There are several issues in the paragraph where the authors describe what they call an anchor-based MCID approach. First, if a patient's change score is the subject of interest, it is more appropriately called the MIC (Minimal important change) rather than M(C)ID, which infers difference between groups. Second, the method explained by the author (corresponding to citation 35) is a description for the PASS (patient acceptable state score), which is not the same thing as an M(C)ID. Could the authors please revise this paragraph to ensure that the appropriate terminology and methods are described? Page 9: Lines 12-15; the authors state that the MCID approach assumes that the response trajectory of those who report improvement is distinct from those who are unsatisfied. The MID/MIC/MCID are measured at a single point, typically the conclusion, of the intervention period. While it is a measure of change through time, it is not a measure of the pattern of change and therefore makes no assumptions about change trajectory. As such, could the authors please explain why and how a statement regarding assumptions of the trajectory of change is relevant to a critique regarding the MID/MIC/MCID? Page 10: Line 19-21; Could the authors please clarify the following sentence: "the sum of the B0 + u0J is the estimated individual baseline response"? What is meant by "baseline response" Isn't this the average plus the jth individual's difference from that average? Page 10-11: Lines 57-13; This paragraph belongs in the introduction

	as it speaks directly to the rationale for why MLM are of interest. Page 11: Lines 37-41; Can the authors please provide greater details regarding how the researcher is to estimate the values that will replace B0, B1, u0j and u1j? If these measurements are specific to the individual, how can they be made prospectively? Page 12: Line 34; When calculating the MID based off a mean-difference approach, only the individuals who have reported they have “slightly” improved or who have not changed are included in the model. The limitation of this is that it often results in very small sample sizes. Can the authors please clarify whether this is also the case for the MLM or whether all participants are needed to be included? Does the model cope with small samples sizes? Page 13: Line 27; Please see my previous comment about the calculation of the PASS. Using the 75th percentile method is not a common method for calculating anchor-based methods. Page 15: Line 10; If the threshold for response using the Omeract-OARSI criteria must still be arbitrarily chosen, then can the authors please comment on how the application of the MLM makes the criteria more precise? Page 17: Lines 10-17; Can the authors please explain why overestimates of the baseline and change SD were included in the model?

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1
Reviewer Name: Gillian Hawker
Institution and Country: University of Toronto, Canada
Competing Interests: I have published previously with Rachael Gooberman-Hill. No other disclosures.

This manuscript describes an alternative statistical approach to assessment of responsiveness to joint replacement surgery and compares this approach to currently used standards. The comparative effects of the different approaches to assessment of patient response are illustrated using data from an existing joint replacement cohort. The issue of how best to define a good outcome following joint replacement, and its determinants, is important. I have no methodological concerns. My major concern is the appropriateness of this paper for BMJ Open and its readability by a non-statistical audience.

My specific comments are as follows:

1. If the audience is a general medical one, I think greater emphasis is needed on the following: a) what is the problem the authors are addressing? what are the problems with the current approaches? how will the MLM approach be used to advance patient care and outcomes? b) while this is a largely a statistical exercise / methodological paper, for a general readership it might be more reader-friendly to let the cohort data illustrate the points being made much more than is done in the current paper. For example, the tables are difficult to interpret - could they be formatted to better illustrate what effectively amounts to variability in the 'error' estimates across methods, which results in fewer or more patients being classified as 'responders'?

Thank you for raising this point. While we agree that further emphasis would be important for a general medical audience, this paper is aimed at quantitative scientists using PROMs in clinical settings. There have been a number of articles reviewing and critiquing responsiveness measures, but none that have attempted to explicitly demonstrate how different responsiveness measure can be estimated from a single underlying statistical model, or clarify a number of the implicit assumptions in the existing methods. Given that BMJ Open welcomes studies on research methods, and papers that address patient outcomes we believe this is within the scope of the BMJ Open.

We have also refined and added tables 3 & 4 to highlight the differences between the current and MLM methods, and compared the proportions defined as responders using a McNemar test explicitly

2. There is no mention of PASS - patient acceptable symptom state - as an indicator of a successful treatment outcome - might this not be considered the ultimate gold standard? Please discuss.

We thank Professor Hawker for identifying the PASS metric, we have now extended the discussion of Anchor based methods to also include the PASS (Page 8, Paragraph 3). The method of construction is similar to that of the MCID. The PASS as described by Tubach et al. "was estimated by constructing a curve of cumulative percentages of patients as a function of the score of interest at the final visit among patients who considered their state satisfactory."[1] Despite the noble intention of focussing on a patient's reported state post-operatively, opposed to change, the threshold is based on an arbitrary statistical definition of 75th Centile of the final score in those who are satisfied. Therefore, much of the critique levied against the anchor based MCII can also be levied at the PASS. In addition, the PASS is conceptually aligned with the RTN criteria, also described, and therefore some of the criticism levied against RTN also applies to PASS.

3. Page 17 first paragraph of results section: states that current methods overestimate SD of baseline and change scores but then states that SD estimates were greater using MLM - have I misunderstood here? Conceptually, I fully understand if we overestimate SDs we would underestimate the proportion of responders...

We apologise for the confusion, we have revised the tense paragraph to ensure clarity and transparency. Page 19, paragraph 1

4. Re the point above, is the bottom line that MLM provides tighter estimates and thus will be less likely to misclassify a responder as a non-responder? If so, the currently used methods simply gives a worst case scenario? Given the complexity of the MLM approach, could you please comment on its usefulness in informing patient decision making?

This is loosely the correct interpretation for RTN and MID definitions of responsiveness, as the definition of response is made based on the estimates of the standard deviation or baseline

scores. However, due to correction for measurement error and regression to the mean, the MLM result in a fundamentally different classification, see cross tabulation of existing and multilevel methods in Tables 3, 4 and Figure 3.

However, the threshold for response defined using MCII is estimated from the MLM. In the estimation of a patient's response, the estimates are adjusted for measurement error, regression to the mean, and allow for heterogeneity in the timing of responses post surgery. This tends to reduce the variability in change scores and shrink the threshold for determining responsiveness towards the mean i.e. it makes it larger. But similar to RTN and MID, the use of MLM provides a fundamental different classification of patients through the incorporation of measurement error and regression to the mean. Therefore, the existing approaches to MCID could be described as providing a best case scenario as they do not account for measurement error or regression to the mean.

The lack of consistency between existing methods of providing either a best or worst case scenario is a property of the specific method. However, we hypothesize that results based on MLM are more likely to be replicated in external populations and reflect a patient's likely response. We have updated the results to reflect this on Page 22, paragraph 2.

5. As you know, we have previously examined % with good outcome using MID, MCID, OO (Arthritis Rheum. 2013 May;65(5):1243-52) - is there more that could be said about the current methods and comparisons with other papers using these metrics - in other words, do you have any recommendations regarding which of these current approaches provides results that are closest to MLM?

Thank you for this suggestion. The similarity of results in comparison to MLM will be very dependent on the data being modelled, i.e. the time between measurements, within subject variability and between subject variability. The use of MLM does not only change the threshold at which patients are deemed to respond or not, it also changes their trajectory due to incorporation of measurement error and regression to the mean.

We do not believe that MLM provides a panacea to patient responsiveness, however we feel that it emphasize the similarity of the underlying models used to estimate the different responsiveness classifications, and draw attention to the assumptions under pinning the model, explicitly emphasize what each model is estimating and how. We have updated our discussion accordingly on Page 22, Paragraph 3.

Reviewer: 2
Reviewer Name: Nicole Pratt
Institution and Country: University of South Australia, Australia
Competing Interests: None declared

Thank you for the opportunity to review the paper "A unified multi-level model approach to assessing patient responsiveness including; return to normal, minimally important differences, and minimally clinical important differences for patient reported outcome measures". This is a well written paper that demonstrates a clear application to a common problem.

The authors have clearly shown that there is an advantage to MSM over standard methods which ignore the time between first and last observation. The main finding appears to be a variance benefit and hence a potentially more accurate prediction.

I have a few minor comments only.

1) The authors say that the standard approach 'overestimates' the variance. Do they have a basis for this statement? It may be that the MSM 'underestimates' variance. A simulation study would help to solve this issue.

We would like to thank Professor Pratt for taking the time to review our manuscript and for her helpful suggestions.

Assuming MLM provide a reasonable approximation to the true underlying data generating process, there have been a number of articles that examine the performance of MLM in terms of bias and coverage of both fixed and random effects. Browne et al. 2002 discuss a number of issues in MLM performance including a comparison between Frequentist and Bayesian estimators. Browne demonstrates that when the number of level 2 units becomes large the performance of IGLS is unbiased, and even when the samples are small RIGLS can correct much of this bias. We now include a reference and have updated our introduction (Page 6, paragraph 2).

2) In the discussion the authors state that model diagnostics are important however I do not believe they were presented for their models.

Thank you for this suggestion. We have now include an example graphical model diagnostics plot for THR patients using the ICOAP total scale in the RTN / MID MLM (Figure 2). The mean function was checked using a ladder plot and normal plots were used to assess the distribution of residuals.

3) the discussion section is rather brief. For example different results were found between Hip and Knee analyses e.g. variance much lower for the knee analysis

We agree that the discussion would benefit from expansion with focus on the differences between THR and TKR patients. We had principally focused the discussion around describing the differences in approaches, opposed to the substantive variability between THR and TKR patients. Professor Pratt correctly notes the variability in TKR is lower than THR patients, and similarly the average improvement post surgery is also lower in TKR patients.

From a clinical perspective the outcome from THR and TKR are known to vary, this analysis supports assertion that hip and knee osteoarthritis are different conditions and should be considered separately. We have now expanded the discussion to reflect the differences in THR and TKR results more fully (Page 22, paragraph 4).

4) the tables could be improved. It appears that the data presented are not in line and this makes the results difficult to read. e.g. is there an extra number in the last column of the first block in Table 1 (67.1) or too few numbers in the same position in Table 2?. It would be helpful to footnote what each of the acronyms means in the tables.

We thank you for spotting this typo, we have now added gridlines to emphasize this phenomenon more clearly (Tables 1 and 2). We have also now included definitions of acronyms within the footnote of the table. The primary reason numbers are not in line for the baseline and change estimates is to emphasize the model used to estimate RTN and MID responsiveness are the same. Similarly, the absolute threshold and proportion of individuals responding are defined equally across satisfied and unsatisfied patients in the MCII method.

Reviewer: 3
Reviewer Name: Anne Thackeray
Institution and Country: University of Utah, USA
Competing Interests: None declared

The authors present a well-written approach to better estimation of assessing patient responsiveness and provide detailed information on the research methods.

My primary recommendations are directed at clarifying sections of the manuscript as outlined below:

We would like to thank Professor Thackeray for taking the time to review our manuscript and for providing constructive comments.

ABSTRACT:
Objective would be clearer by indicating this is not only a review but a comparison of the techniques.

We have amended the objectives to make this clear (Page 2, paragraph 1).

METHODS:
Participant consent, ethics approval: not explicitly stated (implied with reference to the APEX cohort study)- may want to explicitly state

We have now provided an explicit ethical approval statement Page 18, paragraph 2.

Indicate how satisfaction was measured and reference specifically anchor for pain measurement use with MCID estimates.

We now include a brief description of how satisfaction was measured in our cohort (page 17, paragraph 2).

It is a bit unclear how many time points were used in the MLM approach. Table and lines 21-33 (p16) suggest just baseline to 3 months. However, when referring to the MLM approach ,"all individuals are measured at exactly 0,3,6,12 months"– p.16 lines 40-41. Please clarify.

We now clarify that change between 0 and 3 months were estimated using both the existing approaches and MLM approach. However, the MLM approach utilises measurements from 3, 6 and 12 months to refine the estimation (Page 17, paragraph 4).

Please indicate how the model with two splines (and know point at 3 months) was determined to be the best fit. (p16 lines 41-45)

Like most PROMs there is a ceiling effect, and the majority of change occurs within the first 3 months following surgery. We therefore simply visually inspected the data to determine that 3 months was the appropriate placement of the knot point. In more complex responses ensuring knots are placed appropriately is important. We now emphasize that an iterative process is required in complex patterns of response. We have updated the methods to reflect this on page 18, paragraph 1.

Reference results tables in results section (P 16, Lines 45-47)
RESULTS:
Clarify statements in the first paragraph (lines 12-17). You indicate MLM provides a better estimate of SD, but these statements suggest the SD is greater than conventional methods.

We have now amended this paragraph to ensure clarity (Page 19, paragraph 1).

To help determine the relative strength of the MLM approach, please include confidence intervals around % responders (Tables 1 and 2) and mean differences with confidence intervals between proportions of responders using current v. MLM approaches.

We have now added confidence intervals around the proportion defined as responding. However, we feel it is inappropriate to simply calculate a mean difference in the number of individuals responding, as this does not reflect the paired nature of the classification. Therefore, we have now included tables 3 and 4 which demonstrate the cross classification of responder status by each method, and conducted a McNemar test. We believe this more clearly shows the variability in classification of responder status. We now comment on this in the results on page 20, paragraph 3.

Tables: write out P(Resp.)

We have now included a footnote to indicate P(Resp) in Tables 1 & 2

Reviewer: 4
Reviewer Name: Kathryn Mills
Institution and Country: Macquarie University, Australia
Competing Interests: None declared

Comments
Thank you for the opportunity to review this manuscript. With the current debates over the optimal methods to define clinically meaningful improvement in patient reported outcomes, this manuscript provides another mathematical approach. In writing this manuscript, the authors have undertaken an ambitious task. Critiquing current methods for establishing patient important change and proposing another is a large task to undertake within a single manuscript. At a basic level, the manuscript achieves part of its objective: it demonstrates how MLM can be applied to four approaches that are utilised to determine clinical meaning in different contexts and study designs. However, the critique of the four approaches brief and the result is that the authors have under reported the complexities of the approaches including the different context within which they are typically applied and how error/confidence intervals are currently calculated within the approaches.

We thank Dr Mills for taking the time to review our manuscript and providing us with constructive comments. We are grateful that you recognise that this is an ambitious manuscript. We do not believe our intentions are to propose another set of methods to describe patient responsiveness. Instead, we hope to illustrate how current methods of patient responsiveness can be unified into a single statistical modelling framework, which highlights the assumptions and limitations more formally.

We therefore believe it is necessary to provide a brief review of current methods in order to motivate the MLM framework. We are aware there have been a number of extensive reviews on many different aspects of patient responsiveness. However, we tried to keep our review of existing methods brief. We now provide references to a number of review and discussion articles on patient responsiveness (Page 6, paragraph 3).

In the methods, there is a paragraph that speaks to the benefits of MLM approaches (Pages 10-11, Lines 57-13), but the authors do not contrast these strengths with limitation of typical approaches to patient important change.

Thank you. We have now indicated the general benefits of MLM are beyond existing approaches. Our implication is that existing methods do not directly model measurement error, allow for regression to the mean when predicting individual scores, and cannot be extended to multivariate outcomes that appropriately model the correlation between the responses, or allow for variability in the timing of measurements (Page 11, paragraph 2).

The rationale why the MLM may be appealing to researchers is confused throughout the paper. The authors propose that the MLM may be used to predict the chances of a patient experiencing a treatment effect (which would be very useful, but not the same as retrospectively defining treatment effect/response).

We believe some of the confusion has arisen around common terminology used in methodological literature. Specifically, we believe the true patient's response is never observed i.e. Latent, and we only observe an error bound measurement, which requires the underlying patient response to be estimated. Therefore, prediction of a patient's response following surgery and classification of response to treatment are inextricably linked, yet not formally recognised in current approaches to responsiveness (Page 9, paragraph 1).

They also propose that it reduces the overestimation of the threshold the classify a "responder". While it is possible that the MLM does both, jumping between these rationales and only demonstrating the changes on the clinically meaningful thresholds is confusing for the reader. Can the authors please clarify whether the use of the MLM is best for prediction of change or clarification that the change is meaningful? The authors need to provide a stronger rationale for why the MLM should be considered.

The definition of responsiveness is a two stage approach. First, you must predict (estimate) an individuals change, and second, define how to categorise individuals. Estimating change is the primary function of the MLM, defining change in a responsiveness framework is typically dependent on the variability of some feature of the MLM, whether that be variance in the baseline, change or predicted response. Therefore the two components must be considered jointly. We hope the clarification with regard the two stage process is helpful (Page 5, paragraph 3).

In the results section the authors demonstrate how the application of MLM reduces the threshold of clinically meaningful change. While this demonstrates that more individuals may have improved, the authors may consider comparing their proposed threshold to the minimal detectible change for the ICOAP. A major issue, particularly in MID research, is when the MID is less than the MDC. This renders the MID somewhat useless in the clinical setting. It would interesting to see the authors consider this in the current manuscript.

We thank Dr Mills for her observation. We now clarify that a test similar to the RCI could be conducted to determine if change was beyond the measurement error of the instrument. We note that The RCI can be considered analogous to the MDC[2] (Page 8, paragraph 1).

In addition, we had already discussed the problem of ensuring that the definition of response is beyond what is possible by chance alone in the context of MLM, and ensuring the test is suitably generated from the model at hand is an additional benefit of the MLM approach. This is on page 12, paragraph 3.

As the authors acknowledge, MLMs are complex and require multiple alterations to cope with (1) non-linear trajectories (2) heterogeneity within the trajectory classes or (3) multiple time points. Both of these factors are not only expected, but they are common in the OA process. Could the authors please comment further on the clinical utility of their proposed methods?

We agree with Dr Mills that non-linearity and heterogeneity are expected phenomena, and we have edited the text so that we now state this more clearly. However, existing methods do not recognise that these problems occur and are likely to influence the analysis. Whereas MLM ensure these considerations are explicitly considered (Page 9, paragraph 3).

We comment that the clinical utility of refined definitions are currently unclear, and more research is required to assess if the refined definitions of short term patient responsiveness correlated with long term self-reported outcomes or hard end points such as mortality and revision (Page 24, paragraph 1).

Specific comments

Page 5: Line 39; While the argument the authors put forward on regarding the limitation of statistical models being used to quantify patient change is accurate, there have been far more elegant and thorough explanations of this limitation published that the authors may want to substitute e.g. King, M. T. (2011). A point of minimal important difference (MID): a critique of terminology and methods. Expert Review of Pharmacoeconomics & Outcomes Research, 11(2), 171–184. http://doi.org/10.1586/erp.11.9

We thank Dr Mills for highlighting this review, and we agree with many of the sentiments, and concur the Professor King provides a very through critique of MID, and have added citation of this article (Page 6, paragraph 3).

However, we do note that Professor King states that "MID is not an immutable characteristic, which may vary depending on population and context". Therefore, the use of MLM adeptly illustrates the estimation of measurement error within the context the PROM is being used, and confers many benefits. Furthermore Professor King highlights that future directions of MID research should include multiple methods to determine MIDs, and a consolidation of methods would be useful. We believe that primary aims of this manuscript are in tune with Professor King's recommendations. Specifically, illustrating how common measures of responsiveness can be unified in a MLM framework goes someway to providing a consolidation of existing approaches.

Page 5-6: Lines 57-7: Could the authors please provide a rationale the distribution based methods have been identified as "minimal important difference" and the anchor-based methods have been identified as "minimal clinically important differences"? Recommendations by numerous authors (e.g., Revicki, D., Hays, R. D., Cella, D., & Sloan, J. (2008). Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. Journal of Clinical Epidemiology, 61(2), 102–109. http://doi.org/10.1016/j.jclinepi.2007.03.012) indicate that both methods are called MID.

We agree with Dr Mills that language used in the responsiveness field is rather heterogeneous. Our use of MID and MCID was consistent with a recent article by Judge et al. [3] In hindsight, it would have been more sensible for our work to be consistent with the first description of the methods and we have made changes to this effect throughout the manuscript. We now refer to MCID as MCII as per the original description by Tubach et al.[4]

Page 7: Line 13 onwards: This section does not fit in the methodology section as it is not outlining a process of tasks/calculation etc undertaken. In the previous paragraph, the authors state that as part of the methods they will outline common methods used to describe response. This would imply that the explanation belongs in the results section. However, on reading this explanation it is not the result of a systematic search strategy and includes a critique of the literature, not just an outline, that is integral for the authors rationale for the new proposed method. Thus, this would suggest that this section of the manuscript belongs in the introduction. Could the authors please move the explanation section of the manuscript to a more appropriate subheading?

We thank Dr Mills for her comment. However, we respectfully disagree and believe the structure of the manuscript is appropriate given its methodological nature. The methods section contains 3 sub-sections which clearly outline their purpose: "Review of existing approaches to responsiveness", "Multi-level modelling approach to responsiveness and "Example: The APEX cohort Study".

Page7: Line 42-48; First, can the authors please provide a reference for this explanation of the RCI.

References for the RCI are provided on page 7, these are Christensen et al. 1986 and Jacobson et al 1991 (Page 7, paragraph 4).

Second, a reader could find the last sentences in this paragraph confusing. Particularly the phrase "reliability values in the spirit of a sensitivity analysis". Given the precise nature of reliability studies, using "spirit" is an inappropriate word choice here. Could the authors either remove this part of the sentence, or provide a brief example to the reader on how where such values could be drawn from?

Thank you for raising this. However, we would like to keep these sentences in the manuscript. All statistics even those generated from reliability studies are estimates from an underlying sampling distribution and subject to error and sampling variability. Sensitivity analyses generally test models with regards to the underlying assumptions made. Therefore, we believe "a range of reliability values in the spirit of a sensitivity analysis" is appropriate given the context.

Page 7: Line 55; The authors state that the 0.5 SD is of the pre- post-surgery change scores. Revicki

et al., (2008) is then cited to support this. However, the 0.5 SD is of baseline score only. In their 2008 article, Revicki et al., cite Normal et al., (Norman, G. R., Sloan, J. A., & Wyrwich, K. W. (2003). Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. Medical Care, 41(5), 582–592. http://doi.org/10.1097/01.MLR.0000062554.74615.4C) This article clearly states that only baseline scores contribute to the distribution model. Can the authors please change their paragraph to reflect this?

We thank Dr Mill for spotting this error and have amended the text and results accordingly. (Page 7, paragraph 5).

Page 8: Line 12-26; There are several issues in the paragraph where the authors describe what they call an anchor-based MCID approach. First, if a patient's change score is the subject of interest, it is more appropriately called the MIC (Minimal important change) rather than M(C)ID, which infers difference between groups.

We have changed the terminology to be consistent with the original publication by Tubach et al. so we now refer to this as MCII. This has been changed throughout the manuscript.

Second, the method explained by the author (corresponding to citation 35) is a description for the PASS (patient acceptable state score), which is not the same thing as an M(C)ID. Could the authors please revise this paragraph to ensure that the appropriate terminology and methods are described?

We apologise for this error in referencing, we intended to cite the companion article by Tubach et al. We have now corrected this.

Page 9: Lines 12-15; the authors state that the MCID approach assumes that the response trajectory of those who report improvement is distinct from those who are unsatisfied. The MID/MIC/MCID are measured at a single point, typically the conclusion, of the intervention period. While it is a measure of change through time, it is not a measure of the pattern of change and therefore makes no assumptions about change trajectory. As such, could the authors please explain why and how a statement regarding assumptions of the trajectory of change is relevant to a critique regarding the MID/MIC/MCID?

We respectfully disagree with Dr Mills. The assumption of linear change is implicit in existing pre-post analyses. As Dr Mills has already pointed out, an individual's response to a therapy may not be linear, heterogeneous, nor measured at exactly the same time post intervention. The use of MLM makes these assumptions explicit, and when individuals are not measured at exactly the same point in time the MLM adjusts for this phenomenon by creating the appropriate prediction. We now comment on this implicit assumption on page 22, paragraph 3.

Page 10: Line 19-21; Could the authors please clarify the following sentence: “the sum of the B0 + u0J is the estimated individual baseline response”? What is meant by “baseline response” Isn’t this the average plus the jth individual’s difference from that average?

Baseline response is simply the estimated response to a PROM at baseline. We now have clarified the notation, and now indicate that y is the measured response to ensure the meaning of response is defined (Page 10, paragraph 1).

Page 10-11: Lines 57-13; This paragraph belongs in the introduction as it speaks directly to the rationale for why MLM are of interest.

We respectfully disagree, and believe the description of MLM is best presented prior to the description of MLM. We hope that Dr Mills finds this acceptable.

Page 11: Lines 37-41; Can the authors please provide greater details regarding how the researcher is to estimate the values that will replace B0, B1, u0j and u1j? If these measurements are specific to the individual, how can they be made prospectively?

The values of B0, B1, u0j and u1j are estimated with respects to the sample characteristics by the MLM, they cannot be made prospectively. We believe that this query is a result from the confusion in terminology that we have addressed previously.

Page 12: Line 34; When calculating the MID based off a mean-difference approach, only the individuals who have reported they have “slightly” improved or who have not changed are included in the model. The limitation of this is that it often results in very small sample sizes. Can the authors please clarify whether this is also the case for the MLM or whether all participants are needed to be included? Does the model cope with small samples sizes?

The tendency of researchers to include patients that report themselves as “slightly” improved is very much at the discretion of the analyst, and whilst it is likely to result in small sample sizes, this is likely to depend on the specific disease state and treatment being investigated. However, MLM are known to work with small samples with minor adjustments to the method of estimation to some form of restricted maximum likelihood, restricted generalised least squares, and or adjustment to the denominator degrees of freedom.

Despite any issues of small sample estimation, the classification of the remaining cohort of patients experiencing a MCII depends on the application of an appropriate statistical model of change, whether this is done jointly with those who are satisfied is debatable. However, we present this in a modelling framework as it makes the assumptions relating to change explicit. We have added to the description of MLM MCII to make this explicit on page 13, paragraph 1.

Page 13: Line 27; Please see my previous comment about the calculation of the PASS. Using the 75th percentile method is not a common method for calculating anchor-based methods.

Thank you for raising this, however we are following the published recommendations of Tubach et al. in their definition of MCII.

Page 15: Line 10; If the threshold for response using the Omeract-OARSI criteria must still be arbitrarily chosen, then can the authors please comment on how the application of the MLM makes the criteria more precise?

We describe on page 15 that the response(s) are jointly estimated allowing for correlation between the different outcomes and also correlation within the measurement error. Therefore, the predicted estimates of change or adjusted for correlation between pain and function, measurement error and regression to the mean (Page 15, paragraph 1).

Page 17: Lines 10-17; Can the authors please explain why overestimates of the baseline and change SD were included in the model?

We have revised this paragraph 1 on page 15 to ensure that our intentions are clear.

References

1. Tubach F, Ravaud P, Baron G, et al. Evaluation of clinically relevant states in patient reported outcomes in knee and hip osteoarthritis: the patient acceptable symptom state. *Ann Rheum Dis* 2005; 64(1):34-7.
2. Schuck P, Zwingmann C. The 'smallest real difference' as a measure of sensitivity to change: a critical analysis. *Int J Rehabil Res* 2003; 26(2):85-91.
3. Judge A, Cooper C, Williams S, et al. Patient-reported outcomes one year after primary hip replacement in a European Collaborative Cohort. *Arthritis Care Res (Hoboken)* 2010; 62(4):480-8.
4. Tubach F, Ravaud P, Baron G, et al. Evaluation of clinically relevant changes in patient reported outcomes in knee and hip osteoarthritis: the minimal clinically important improvement. *Ann Rheum Dis* 2005; 64(1):29-33.

VERSION 2 – REVIEW

REVIEWER	Nicole Pratt University of South Australia, Australia
REVIEW RETURNED	10-Apr-2017

GENERAL COMMENTS	The authors have adequately addressed my concerns. I have no further comments.

REVIEWER	Kathryn Mills Macquarie University, Australia
REVIEW RETURNED	11-Apr-2017

GENERAL COMMENTS	I have no further comments regarding my initial queries. My only concern with the manuscript is the utility of the method explained. The authors responses to reviewer 1's comments regarding its usefulness in informing patient decision making makes valid points i.e. it permits heterogeneity, adjusts for measurement error and regression to the mean. They also clearly state that the manuscript is aimed at quantitative scientists using PROMS in clinical settings. However, assessing responsiveness is also a pertinent issue in clinical practice and one of the benefits of the methods the authors rightly criticise is that they are easily calculable (if also mis-applied) in clinics. I believe that the authors need to acknowledge that this method is confined to research scenarios either in their introduction or as part of the limitations in the study design.

VERSION 2 – AUTHOR RESPONSE

Reviewer: 4
Reviewer Name: Kathryn Mills
Institution and Country: Macquarie University, Australia
Competing Interests: None declared

I have no further comments regarding my initial queries. My only concern with the manuscript is the utility of the method explained. The authors responses to reviewer 1's comments regarding its usefulness in informing patient decision making makes valid points i.e. it permits heterogeneity, adjusts for measurement error and regression to the mean. They also clearly state that the manuscript is aimed at quantitative scientists using PROMS in clinical settings. However, assessing responsiveness is also a pertinent issue in clinical practice and one of the benefits of the methods the authors rightly criticise is that they are easily calculable (if also mis-applied) in clinics. I believe that the authors need to acknowledge that this method is confined to research scenarios either in their introduction or as part of the limitations in the study design.

We now include the following sentence in the limitations.

In addition, the use of multiple measurements in MLM primarily restricts the method to a research setting.