Author's response to reviews

Title: Estimating Number Needed to Treat from continuous outcomes in randomised controlled trials: Methodological challenges and worked example using data from the UK Back Pain Exercise and Manipulation (BEAM) trial

Authors:

Robert J Froud (r.j.froud@qmul.ac.uk)
Sandra Eldridge (s.eldridge@qmul.ac.uk)
Ranjit Lall (r.lall@warwick.ac.uk)
Martin R Underwood (m.underwood@warwick.ac.uk)

Version: 4 Date: 8 May 2009

Author's response to reviews: see over
Dear Mark,

Please find below, a point-by-point response (bold text) to Professor Guyatt’s criticisms (italic text) and the changes we have made to our manuscript. In addition, please be advised that we have now typeset our manuscript in LaTeX, using the BMC template. We are given to understand (from the BMC website) that this will help save time with regards to proofing and extracting references, etc (http://www.biomedcentral.com/info/ifora/tex).

Also, we have revised our author contributions section, which now conforms to the BMC MRM house style. Please contact us if there is anything else that requires our attention.

We look forward to hearing from you.

Regards,

Robert Froud (On behalf of all authors)

Quoting Professor Guyatt

“1. Validity of transition ratings. Use of the transition rating depends on its validity. Thus, in both the methods and results, the analyses related to the validity should precede the presentation of other analyses.

“Unfortunately, the results of the exploration of the validity of the transition ratings are problematic. First, it seems to me the authors must have sign wrong with either the correlation between the transition rating and the change in RMDQ (which the authors report as negative) and the correlation between the follow-up scores and the RMDQ (which the authors report as positive). It seems to me that these must be going in the same direction.

“Second, assuming this is the case (or even if it isn't, and I've missed something, and it is the absolute value of the correlations on which we should focus) it is a big problem that the correlation between the post rating and the transition rating are greater than the post – pre and the transition rating. This suggests that the respondents have essentially ignored the pre rating and their transition response is being driven largely by the post rating. Further evidence of this is that the correlation with the pre rating is going in the same direction as the correlation with the post rating and that the pre rating explains a trivial amount of variance in the transition rating in relation to the post rating.

“A harsh (but possibly the most accurate) assessment of these results is that the transition rating isn't working at all the way that it should be and that any analyses based on the transition rating should be abandoned. If the authors still wish to use the transition rating in the face of these problematic results they should at least acknowledge, in the discussion, the major questions these results raise about the validity of the transition rating.”
Our response:

We agree with Professor Guyatt’s assessment that the transition question did not work the way it was intended to work. These results do indeed suggest that the main driver of the transition response was the post RMDQ rating and the transition rating did not measure change per se. We had acknowledged this in our manuscript in the discussion section (page 17 of our previous submission). However, we have now attended to making this point much clearer (the relevant paragraph is reproduced below). Critically, after considering Professor Guyatt’s argument, we have been persuaded to abandon those NNT analyses that were based solely on the transition rating and report NNTs based solely on the consensus thresholds.

As we now report only those NNT analyses that were informed by the consensus ratings, we feel that in the methods section the material pertaining to the transition question is now correctly placed: after the introduction to the consensus threshold and before the description of methods used to calculate NNT. However, in the results section we have placed results pertaining to transition question validity before the results of our NNT analyses, as per Professor Guyatt’s instruction. We have also edited our manuscript for continuity, removing all references to the previous NNT analyses that were based on the transition rating or the associated sensitivity analysis.

We note that we have interpreted Professor Guyatt’s suggestion that ‘any analyses based on the transition rating should be abandoned’ to refer to our NNT analyses. The main purpose of our using the transition rating was to verify the appropriateness of applying the 5 point consensus threshold to the UK BEAM population. Whilst the validity of the transition rating clearly affects both our estimates of MIC, and minimal detectable change (MDC), our estimates of these properties were well within the ranges reported in other studies, and also within those reported in studies reviewed by the consensus team. This is encouraging from the point of view of assessing the appropriateness of applying the consensus threshold to the UK BEAM population; the generalisability of MIC / MDC thresholds to different populations is something that is often assumed, and not explored empirically at all. Notwithstanding the similarity of our estimates with those reviewed by the consensus team, we have emphasised the limitations of our estimates given the transition question’s questionable validity (paragraph reproduced below).

We are grateful to Professor Guyatt for pointing out our sign error. This was due to the generation of change variables by subtracting post scores from pre (this made our ROC analyses more straightforward), rather than pre from post, which of course gives the correct sign. We have amended this in the manuscript. We
have also corrected two minor errors we detected following our previous submission: the regression coefficient for the baseline score at three months was actually -0.056, rather than -0.067 as was reported previously; and the RMDQ MIC at 12 months was 5.0 rather than 4.3 as was reported previously. Neither of these changes affect our conclusions in any way.

We look forward to hearing from you.

Reproduced paragraph:

“It is not ideal that our transition question ratings correlate moderately with follow-up scores, and slightly but in the same direction with the baseline score; nevertheless this is not an unusual finding.[1, 2] The baseline RMDQ score significantly explained 2% of the residual variance in transition rating in the regression models we fitted. However, this is a trivial proportion. In addition, we found the correlation between the follow-up score and the transition question was greater than the correlation between the change score and the transition question. These findings suggest that participants’ health status at the time of follow-up may have been the prime driver of their response to the transition question.

The poor performance of the transition question may have led to inaccurate estimates of MIC and minimal detectable change, as both of these rely upon the transition rating to identify improved or stable patients. However, our estimated MIC value of 4.0 points, falls within the 3.0 to 5.0 range of values reported in other studies using similar methods;[3-9] and our minimal detectable change estimate of 7.6 points, falls between the 5.4 to 12.1 ranges seen in other studies.[3-5, 7, 10, 11] Moreover, both our MIC and minimal detectable change estimates fall within the 2.0 to 8.6 point range considered by the consensus study team.[12] Therefore, notwithstanding the questionable performance of our transition question, we applied the 5 point RMDQ consensus threshold to our population.”

Quoting Gordon Guyatt

“2. The term "clinically important improvements" in the abstract is problematic. See: Guyatt GH, Montori VM, Devereaux PJ, Schünenmann HJ, Bhandari M. Patients at the center: In our practice, and in our use of language (Editorial). ACP Journal Club 2004;140:A11.”

We have now changed this to “individual responders to treatment (e.g., those who reach a particular threshold of change).”
Whilst we would encourage the use of the term “patient-important”, which Professor Guyatt suggests in the paper cited above, we are unable to use it in this in this instance. The consensus thresholds are essentially a hybrid of 1) anchor-based studies, which produce patient-important estimates, and 2) distribution-based studies, which estimate what magnitudes, are detectable beyond the measurement error of the instrument; it would be incorrect to describe these as patient-important. We believe that both properties are important: it is desirable to know a change is both patient-important and detectable beyond measurement error. Thus, when describing a hybrid of these properties, we prefer to use the term ‘responder’. Notwithstanding this, we have added the term ‘patient-important’ to a sentence under the subsection “Individual improvement” (reproduced below), where we specifically discussed MIC, and referenced Guyatt’s paper on practice and use of language.

“Trials can be designed so that the minimal detectable change, is less than the threshold of minimally important change ((MIC) i.e., a magnitude of change that may be considered patient-important).[3, 11, 13]”

References in reproduced text:


