Author’s response to reviews

Title: Systematic reviews need to consider applicability to disadvantaged populations: inter-rater agreement for an equity plausibility algorithm

Authors:

Vivian A Welch (vivian.welch@uottawa.ca)
Kevin Brand (kbrand@uottawa.ca)
Elizabeth Kristjansson (kristjan@uottawa.ca)
Janet Smylie (janet.smylie@utoronto.ca)
George A Wells (gawells@ottawaheart.ca)
Peter Tugwell (tugwellb@uottawa.ca)

Version: 2 Date: 13 November 2012

Author’s response to reviews: see over
Dear editors,

We greatly appreciate the detailed and thorough review of this manuscript by the two reviewers and their constructive comments. We have revised accordingly and responded to comments below.

We look forward to your decision. Please do not hesitate to contact us for further details or clarification.

Sincerely

Vivian Welch on behalf of the author team

Detailed Response to reviewer comments

Reviewer's report
Title: Systematic reviews need to consider applicability to disadvantaged populations: inter-rater agreement for an equity plausibility algorithm
Version: 1 Date: 17 February 2012
Reviewer: Maria Ospina
Reviewer's report:
Reviewer’s report
The manuscript by Welch et al is aimed at developing and assessing inter-rater agreement for an algorithm that will help systematic reviewers to predict whether differences in relative effect measures are likely for disadvantaged populations relative to advantaged populations. The background section describes in a clear and concise way the context in which this study is important. The research question and objectives are clearly stated and accompanying references appropriately support the relevance of the research question. The manuscript is well written, the study methodology is sound and the reporting of results is clear. This research constitutes an important contribution to the assessment of equity in systematic reviews.
I would like to provide a few essential and minor revisions and some comments that the authors may want to consider in the methods reporting and the discussion.

Authors reply: Many thanks for the careful and thorough review and important comments.
Major Compulsory Revisions:
1. My main concern regarding this research manuscript is related with the methodology to assess the construct validity of the equity algorithm and the use of the discussion around applicability and generalizability by SR authors across sex and socioeconomic status as an indicator of real differences in effects by these two PROGRESS dimensions. I will address these issues in comments #2 and 3 below.

2. I suggest the authors revise the definition of construct validity that was provided in this study. Typically, empirical validity relates to predictive validity, concurrent validity, content validity, and construct validity, noting that each involves a different emphasis on a criterion. The evaluation of construct validity in this study seems to correspond better with the definition of concurrent validity (i.e., an assessment of whether a new measure/algorithm correlates with some pre-specified criterion). Construct validity is traditionally defined as the empirical demonstration that a measure (in this case, an algorithm) measures the construct it claims to be measuring. I understand that recent developments in the clinimetrics field consider that all types of empirical validity are ultimately different facets of a single unified form of construct validity. However, I think it is important for the reader to have a more precise idea of what type of construct validity is under evaluation; in this case, concurrent validity.

    **Author reply:** Thank you for suggesting this clarification. We had suggested “construct validity” since the criterion of whether differences exist across sex or socioeconomic status is difficult to establish because systematic reviews rarely discuss or present results disaggregated by sex and socioeconomic status. In discussion with the author team, we think that readers may be more familiar with the term “construct validity”. We agree that concurrent validity is a helpful clarification, thus we have included a comment regarding this in the methods.

    **Changed Text page 9:**

    ....Comparison to this criterion would be considered concurrent validity. However, since we cannot know the real intervention effects, we used a proxy by considering the results and discussion of the systematic reviews as an indicator of real differences in effects and compared this with the health equity plausibility algorithm to assess construct validity.

3. The authors mention in the Methods section (Construct Validity, 1st paragraph) that some of the limitations of the approach to assess the “construct”/concurrent validity of the equity algorithm were related with the fact that SRs and underlying primary studies may have not been designed or powered to detect such differences. I suggest, however, that the authors provide more details in the discussion on how the use of this proxy indicator could have accounted for the low measures of concordance reported in the results, which is an important finding of this study. I respectfully would like to raise some issues that the authors may consider to improve the discussion of this important limitation of the study:
a. It is clear that the authors recognized that, overall, the proxy measure for “construct”/concurrent validity was imperfect. The proxy measure was based on “the judgments of applicability described by the authors of the systematic reviews”. I retrieved the 10 SR to verify what type of information SR reviewers provided (if any) in the methods, discussion and conclusions regarding the applicability of their findings. I found that in some instances, judgments regarding the plausibility of different effects for patient characteristics, intervention delivery or comparator were not based on the SR results per se but were mere hypothesis-generating exercises for future research. For example, the Mass media review (ref #35) mentioned potential differences in literate and non-literate in the discussion; however, the interpretation was not supported by their data nor was part of any subgroup analysis. Alternatively, another review (ref #38) reported the rate of female participants in the trials (with some of them including women only) but no subgroup analysis was conducted. Most importantly, there was no mention of the potential impact of sex on the tx effect in the conclusions or interpretation of the results.

Author reply: Thank you for this detailed review. We noted this example on page 12 in the results about descriptions of differences in the systematic reviews.

Changed text, page 12:
It is important to note that some of the systematic reviews mentioned differences across PROGRESS-Plus factors, even though they were not supported by evidence. For example, the mass media review mentioned differences between literate and non-literate populations, but this interpretation was not supported by their data no part of any subgroup analyses.

b. The authors acknowledged that the ten SRs included a mix of plausibility of different effects across the three algorithm domains; however, a more precise definition of how SR judgments were obtained (e.g., ANY mention [supported or not by the evidence presented] of applicability, only applicability supported by results, etc) would have been useful.

Author reply: we clarified this method on page 10:

Changed Text:
We extracted these details from the discussion sections of each systematic review by assessing any mention of differences across PROGRESS Plus factors, whether they were supported by evidence or not (e.g. hypothesized differences not supported by evidence were included).

4. I noticed that the algorithm was applied to nine Cochrane reviews and one non-Cochrane review. Was the Cochrane “attribute” considered when selecting the reviews for assessment?

Author reply: Thank you for pointing this out. The attribute of Cochrane or non-Cochrane was not a criteria for selection because we did not feel this affected the issue of assessing subgroup differences. Rather, we sought to select a balance
of different types of interventions (policy, personal, population-based prevention) and disease burden.

Text change (page 9):
Candidate reviews were identified by searching MEDLINE and the Cochrane Library by VW. The ten reviews were chosen by discussion of two reviewers (VW, PT) based on the above criteria. They included nine Cochrane reviews and one non-Cochrane review. The attribute of being Cochrane or non-Cochrane reviews was not a criterion for selection since we did not feel there would be important differences in how systematic reviews approach the issue of differences across subgroups based on this attribute.

Was it relevant whether the SR included or not a meta-analysis. If not, why?

Author reply: We required the systematic review to include at least 5 studies. However, we did not require a meta-analysis since assessment of differences across PROGRESS-Plus factors does not require subgroup analyses.

Text change (page 9):
(a meta-analysis was not required since differences across populations can be assessed without a meta-analysis).

5. There are some other issues that remained a bit unclear and I think they are worth to be mention (briefly) in the manuscript: Who selected the SRs? (i.e., were the same three authors that developed the algorithm?);

Author reply:
We added how the reviews were selected by VW and PT.

Changed text (p. 9)
Candidate reviews were identified by searching MEDLINE and the Cochrane Library by VW. The ten reviews were chosen by discussion of two reviewers (VW, PT) based on the above criteria.

do they provide a reasonable representation of reviews that address (and do not address) the categories that the algorithm is aimed to assess?

Author reply:
We did not choose systematic reviews based on whether they did or did not assess differences in sex or socioeconomic status because we felt that searching for these differences might lead to a biased selection of reviews. As described in the article, systematic reviews were selected to balance the type of intervention (policy level, population-based prevention, personal level) and disease burden. We chose these because we thought that differences across sex and socioeconomic status might vary across these attributes (e.g. we expected more differences for policy level or population-based interventions). We feel the selected systematic reviews are balanced on these attributes.
Do you think that variations across the reviews in their criteria for study design (some of the reviews included studies with no control groups) and comparator groups (some of them included studies with placebo controls) may have impacted the assessment of Question #3 in the algorithm?

Author reply: Question #3 asked: “Are there differences in the comparator across patient, community or population that are likely to create important differences in magnitude of relative effects?”. In planning the study, we thought that study design and type of comparator (control vs placebo) might influence this assessment. However, we did not see a trend in the open-ended questions that reflected this and we did not think that we had sufficient variability to comment on this in the discussion.

Text change: we added a comment that we were unable to separate reasons for different responses in discussion.

Page 15:
Thus, we could not assess whether differences were due to some of the reasons that we expected differences such as the type of study design (randomized vs. observational) or type of comparator (placebo vs. control).

Minor essential revisions:
1. Please, consider reporting some guidelines for the interpretation of level of agreement and Fleiss kappa (i.e., A Kappa value in the range from 0.0 to 0.40 is considered poor agreement; 0.41 to 0.60 moderate agreement; and 0.61 to 0.80 substantial agreement [Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-74; Seigel DG, Podgor MJ, Remaley NA. Acceptable values of kappa for comparison of two groups. Am J Epidemiol 1992;135:571-8]).

Author reply: Thank you for this suggestion. Reference 45 describes the interpretation of kappa as follows: “k<0.2 considered slight, 0.2<k<0.4 as fair; 0.4<k<0.6 as moderate, 0.6<k<0.8 as substantial and k>0.8 as almost perfect”. We have added this to the paper.

Text changed (page 12)
A kappa of less than 0.2 is considered slight agreement [45], with kappa of 0.2<k<0.4 considered fair; 0.4<k<0.6 moderate, 0.6<k<0.8 substantial and k>0.8 almost perfect agreement.

1. Introduction (1st paragraph): Please, change to: ‘World Health Organization Commission on Social Determinants of Health (CSDH), as the acronym is used

Author reply: added

Reviewer's report
Title: Systematic reviews need to consider applicability to disadvantaged populations: inter-rater agreement for an equity plausibility algorithm
Version: 1 Date: 17 September 2012
Reviewer: Donna Helene Odierna
Reviewer's report:
This innovative paper by Welch and colleagues represents an effort to develop and test a tool for determining whether or not the applicability of results of systematic reviews are likely to differ according to population disadvantage. Such a tool would be valuable; it could assist systematic review authors in a priori specification of subgroup analyses, increasing users’ confidence in the findings. Therefore, this paper, even with its somewhat inconclusive results, is an important step in the development of a usable and valid equity plausibility algorithm.

Author reply: Thank you.

My area of research includes generalizability of research to diverse populations, and I do not have expertise in the development of clinimetric scales. Therefore, I concentrate my remarks on the applicability of the equity plausibility algorithm, rather than on the specific items it comprises. I enjoyed reading this paper; It has many strengths and few flaws, most of them minor.

General comments:
• The paper is well-written, and adheres to relevant standards.
• The research question is clear.
• Title - If “equity plausibility algorithm” is a term that is being newly introduced in this literature, using it in the title might baffle some readers. (I believe the term “equity algorithm” is used in matters of trade.)

Author reply: Thank you for pointing us to this literature. In order to reduce confusion, we have changed the title to “health equity plausibility algorithm” throughout the paper. We felt that the concept of “algorithm” is important to reflect the decision under assessment.

• The methods are well-suited to the research question, and are described in adequate detail.

Author reply: Thank you.

• The PICO format provides a strong and coherent framework. Furthermore, the addition of the question about respondents’ reasons for their answers adds valuable insight into the process, and potentially into the utility of the instrument. I will say more about that later in my review.
• The limitations are clearly stated, perhaps somewhat overstated - I was sorry to see that the discussion section focused more on the low kappa scores and their possible meaning than on the raters’ agreement on likely differences by sex and SES. I think these findings deserve prominence. For example, the 79% agreement about intervention delivery and SES is particularly important, given that many individual studies and systematic reviews don’t adequately report on or analyze by SES, while at the same time we know that poverty and other
SES-related social health determinants have profound effects on health.

Author reply: We appreciate your concern for lack of emphasis on these results. However, given the low kappa values, we felt that it was inappropriate to place too much emphasis on these results in the discussion. We hope that this paper will provide a stimulus for future studies designed to avoid some of the limitations that we highlight.

Major (but not compulsory) revisions:
1 - Introduction, p.6; Consistency, p.8, Discussion: The study was described as evaluating the algorithm across PROGRESS-Plus categories, but it only assessed gender and SES. This should be made clear at the outset, and include an explanation of why these categories were chosen. Results might be different for other categories, but this remains unknown because they were not assessed. For example, the authors could briefly speculate about which PROGRESS-Plus categories might be substantially different and why. Along with race, gender and SES are among the more common sub-groups discussed in the context of health equity even absent the PROGRESS-Plus framework. Speculation about the other, less commonly-recognized elements would strengthen the authors' decision to apply the framework in their paper.

Author reply: Thank you for this important point. We chose sex and socioeconomic status because we felt that the respondents would feel most informed and comfortable with making this decision, partly because they are the most commonly characteristics that are assessed for (and reported) differences in the literature.

Text change (pages 6-7):
We chose to focus on only two PROGRESS-Plus characteristics because we felt that this would be a difficult task for respondents, and multiple characteristics might lead to lack of attention to additional characteristics. We selected sex and socioeconomic status because they are amongst the most commonly reported differences in effects and burden of disease in the literature, and we felt respondents would feel the most able to make decisions about these characteristics. We recognize that differences across other characteristics such as ethnicity, educational attainment and occupation are important but these have not been assessed in this study.

Discretionary revisions:
2 - Purpose, p7. Example of relative risk of TB for immigrants and refugees compared to high country-born populations. It seems that the 2% absolute risk reduction is only for the country-born, rather than for an entire population including refugees, and is thus somewhat misleading and should be clarified.

Author reply: Thank you, we have clarified as follows (addition is in bold).

Text changed
the difference of 33 per 100 for refugees or immigrants compared to 2 per 100 for high-income country born reflects a difference in baseline rates alone.
3 - Methods, Consistency, p.8, Limitations p.15: The forced yes/no choice is of concern. The authors justify their decision, but don’t adequately discuss possible consequences. Did this result in skipped items? The authors say that raters may have resorted to judgment or guesses (possibly the 4 guesses referred to in Table 6?), but they don’t speculate how it might affect results, or how this could be addressed in future research.

Author reply: Thank you for this comment. No items were skipped. We added this to the results (page 12). Future research could use more categories to reflect strong agreement to strong disagreement, or use a visual analogue scale. This might allow a better assessment of the confidence of raters for each response.

Text changed: page 12
No questions were skipped.

Text changed: page 15
This might be addressed in future research by using more categories to reflect strong agreement to strong disagreement or by using a visual analogue scale; both of these methods would allow an assessment of confidence of raters in the importance of the difference.

4 - Construct validity: As the authors acknowledge, use of the discussion sections of SRs as a proxy is problematic. This could likely be a bigger problem with less-commonly acknowledged PROGRESS-Plus categories; how often do investigators mention, say, undocumented immigrants, people with disabilities, or minorities that may be subject to stigma and even persecution, depending on where they are located? This might merit mention, given the emphasis on the PROGRESS-Plus framework.

Author reply. We agree with this reviewer that the likelihood of mentioning other characteristics such as immigrants or people with disabilities is lower. However, since we decided to focus on only two characteristics, we only extracted details about these characteristics, and we feel that discussion beyond these characteristics would be overstated given our decision to focus.

5 - Inter-rater consistency, p.11-2. Also Table 6. Almost a third of the reasons for the raters’ answers were based on personal experience, more than any other category. This is valuable information in itself, and it raises important, intriguing questions. How does personal experience fit in? That is, who is included in the process, whose voice remains unheard? From the description of the raters in the text, I am guessing that women were well represented among the raters, but people with low SES were probably not. Does shared experience translate into high agreement and kappa scores? What are the implications for other PROGRESS-Plus variables? How can members of target groups be incorporated into the process? Should they be, even though they may lack academic training? I don’t expect the authors to address all of these in their paper, but I think that
mention of some of the issues would be appropriate if the authors agree that they are important.

Author reply: we agree that this is important. We acknowledge this on page 16.

Text change: Page 16

Also, given that personal experience was the most common reason for ratings, the raters individual characteristics may be important (e.g. whether they are women, low income, immigrants), and this could be explored in future studies.

6 - Agreement with conclusions p.13. I would like to know more about the nature of the differences between the raters and the SR authors.

Author reply: We agree that these differences are perhaps the most interesting result of the paper. The details are in Tables 4 and 5. However, since the kappa was low, we felt that discussing these differences in detail would be misleading.

7 - Tables and Appendices:
7a - Table 1: Provides background information and could be moved to the appendix. (Discretionary revision)

Author reply: Thank you, we agree and have done this.

7b - Construct validity, p.12, also Appendix 2. The authors give examples from two reviews that appear in the appendix. It would be helpful if these reviews were at the top of the list in the appendix so that readers don’t have to search for them.

Author reply: we have done this.

Minor essential revisions:
7c - Table 3: The total in the “experience with systematic review” cell is 33, whereas there are 35 raters. Two raters are unaccounted for.

Author reply: We have corrected this.

7d - Table 4: MMR is not explained (while most people might know this, PSTD and SES are also commonly used and are explained in the footnote. HIV is of course fine as it is.)

Author reply: we have added: MMR: Measles, mumps and rubella vaccine

7e - Appendix 3 Superscripts should be explained in a note so the reader doesn’t have to figure it out from the text. Also, this appendix might possibly be more appropriate as a table in the main section of the paper.

Author reply: thank you for suggesting this. As above, we felt that, with the low kappa, we did not want to overstate the results for agreement and disagreement with the systematic reviews. We have added the following note:
**Note:** Bolded items indicate greater than 70% agreement between raters. Superscripts indicate whether results agree with the discussion sections of the systematic reviews: NK: not known, Y: agree with discussion section of systematic review, X: disagree with discussion section of systematic review.