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:

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.

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.

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.

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.

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

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

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]).

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


4. Please, revise reference #39 in Appendix 2. I think the correct reference number is #40 (Demicheli et al)

5. Table 3 presents the characteristics of the 35 raters who assessed the algorithm. Characteristics of raters were presented according to their main experience with SRs; however, only 33 were reported. Please, clarify which was the role/experience of the two remaining raters.

Level of interest: An article of importance in its field

Quality of written English: Acceptable

Statistical review: No, the manuscript does not need to be seen by a statistician.

Declaration of competing interests: I declare that I have no competing interests