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.

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

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.

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.

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.

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.
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.

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

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.

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.

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.)

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.

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 am a member of the Cochrane Collaboration Equity Methods Group. I have no financial competing interests.