Reviewer's report

Title: Systematic review of disability weight studies: comparison of methodological choices and values

Version: 1 Date: 4 October 2013

Reviewer: Scott Grosse

Reviewer's report:

Major comments/compulsory revisions (itemized list follows)

This paper is said to have systematically reviewed all studies that had developed DALY disability weights (DWs) and were published from 1990 through 2012. What databases were searched and what search terms were used? This appears to be a non-systematic narrative review, and although the authors state that they had captured all studies that developed DWs, there is no documentation of attempts to check for completeness by testing alternative search strategies.

The inclusion criteria are not rigorously defined and do not appear to have been implemented in a consistent manner. For example, it is stated that only English-language articles were included, but according to an online database reference 34 is in Korean. Only studies that developed DW estimates were supposed to be included, but reference 38 (Jelsma et al. 2000), which was a brief letter to the editor, did not develop DW estimates. The letter reported on desirability rankings of 22 conditions among two panels in Zimbabwe but did not estimate DWs. That study is relevant to a discussion of the extent of agreement between professional and lay panels but did not develop DWs that could be used in calculating YLD.

The abstract states that the search was restricted to articles published in international peer-reviewed journals, but the text states that the grey literature was also searched. The authors were inconsistent in how they reference such studies. For the Dutch DW study, both the original grey literature report (Stouthard et al. 1997) and the journal article by Stouthard et al. (2000) summarizing the findings were cited, as references 39 and 54. It's understandable that the original report was cited as the source of DW estimates, since the 2000 article didn't report the detailed DW results. However, the same logic was not applied to the Australian Burden of Disease and Injury (BODI) study, for which the original 1999 report was not cited and the cited source of the DW estimates was a subsequent article by Mathers et al., reference 36. The two sources are not identical; the 1999 report noted that for some mental health conditions not included in the GBD 1990 or the Dutch DW study the authors used a VAS tool with disease experts to elicit DW estimates. Also, like reference 54, reference 36 did not report detailed DW estimates. Why were the two studies not treated equivalently? Likewise, the unpublished report of the Ontario burden of infectious disease study was cited as reference 35 (Kwong et al. 2010), but the
The sensitivity of the search strategy appears to be low, at least for studies that used MAUI weights to calculate DWs. I used Google Scholar to search on DALY disability weights and EQ-5D for articles published from 2009 through 2012. Two of the first 10 references that came up were cited, one of which, reference 40 (van Spijker et al. 2011), did not use the EQ-5D. The other, reference 23 (Haagsma et al. 2009), did use EQ-5D scores as DWs. Three other empirical studies calculated DWs using the EQ-5D as a MAUI. Lyons et al. (2011) used EQ-5D UK weights to estimate DWs in the UK Burden of Injury Study. Siddiqi et al. (2010) used EQ-5D US weights to estimate the number of DALYs from hemophilia in the USA. Jia et al. (2011) used the EQ-5D+-VAS to estimate DWs for Chinese patients with schistosomiasis japonica.

The authors correctly note that a major advantage of the DALY measure is the fact that because it assesses the overall burden of disease and injury at the population level it allows for comparability of impacts of different diseases and risk factors. However, that advantage assumes use of DWs that are comparable across disease areas. Regrettably, many of the DW studies that the authors reviewed do not have that attribute. The authors refer to studies with “alternative sets of disability weights”. In fact, many of those studies were intended not to develop alternative DW sets but instead to derive DWs for specific groups of diseases. Some of those studies may have been intended to boost the profile of a particular disease area by increasing its estimated burden of disease relative to other diseases, which is counter to the objective of comparable risk assessments at the population level for which the DALY was developed.

The description of the use of MAUI in the first paragraph in the Review section is incomplete. A MAUI such as the EQ-5D has two components – a health state classification system and a tariff or set of valuation weights for each health state based on preference-based exercises conducted among members of the general population. Numerous tariffs exist for calculating EQ-5D index scores and one should always identify the tariff used to generate health utilities. There are two other ways to derive HRQL scores using the EQ-5D or its variants that do not involve calculation of health utilities and hence do not qualify as a MAUI. One is to use the VAS that accompanies the EQ-5D, also known as the EQ-5D-VAS to distinguish it from the preference-based EQ-5D index. Finally, one can use the health description system of the EQ-5D or other instrument and apply valuations derived independently, from a panel of experts or laypeople, to calculate DWs. For example, the Dutch DW study, described in references 39 and 54, used the
EQ-5D+ together with PTO for valuation. Haagsma et al. (reference 23) used the EQ-5D in both ways, as a conventional MAUI using the Dutch tariff of QALY weights, and using TTO weights derived from a panel of 143 members of the Dutch population, previously published as reference 32.

The Results section on page 8 states that four studies used the EQ-5D or EQ-5D+C (references 23, 29, 30, 36). That list is problematic for reasons other than that it left out studies missed by the authors’ search strategy. In particular, reference 36 (the Australian BODI study) did not directly derive DWs using the EQ-5D or any MAUI and should not be included in this list. Instead, Mathers et al. used DW estimates taken from both the GBD1996 and the Dutch DW study and used a regression model based on the Dutch DW study to calculate DW for additional conditions not included in that study and to adjust annualized estimates for duration for acute conditions. If the Australian BODI study is included, the Dutch DW study and reference 32 (Haagsma et al. 2008) should be listed as well. Of course, none of those studies used the EQ-5D as a MAUI. References 29 and 30, from the same investigators and both addressing periodontal conditions, differed in one important respect. Reference 29 (Brennan and Spencer 2004) asked dentists to apply the EQ-5D classification to hypothetical patients, which is not consistent with the intended use of MAUIs, whereas reference 30 (Brennan et al. 2007) asked dental patients to use the EQ-5D to self-report health states. Not surprisingly, reference 30 reported lower DWs and acknowledged that reference 29 “may have” overstated the DWs.

Major Compulsory Revisions:
1. Consistently apply inclusion criteria, excluding those studies that do not meet the criteria
2. Either redo the search to include relevant studies that were omitted or acknowledge that this was not a systematic review and that the studies cited do not include all relevant studies.
3. Consistently report DW studies using the first peer-reviewed journal article that summarizes the methods and DW estimates (the SAMJ article cited does not meet that criterion).
4. Consistently and accurately classify studies according to whether they used a MAUI to estimate DWs or instead used valuations from special panels in combination with a generic health state description system.
5. Clarify the difference between studies that produced alternative DW sets for national or global YLD calculations from those that only focused on specific groups of conditions.

Minor Essential Revisions:
1. The statement that the approach using MAUI to calculate QALY weights is similar to DW other than an additional step needs clarification. The additional step should be succinctly defined.
2. The “pairwise comparison” method used for valuation in reference 25 (the GBD 2010 study) needs some explanation and clarification. Pairwise comparison
alone is useless for estimating DWs, because it only indicates relative desirability. In fact, the GBD 2010 study used additional information on population health equivalence to anchor results to a DW scale from 0 to 1.

3. The phrase “are represented” in the last paragraph of the Review section should be replaced by “represent”.

4. In the section on Comparison of disability weights, the DW for severe depression is said to be 0.047 in the GBD 2010 study (reference 25), which is absurd. Table 3 shows the correct value to be 0.655.

5. Table 3 reports the DW for depression in reference 27 as 0.619. That should be checked. According to reference 43 (Kruijshaar et al. 2005), the GBD 1990 study reported an overall DW of 0.47 for major depression.

6. The statement that the DW for paraplegia varies from 0.047 in GBD 2010 (reference 25) to 0.671 in GBD 1990 (reference 27) is misleading. Table 3 makes it clear that reference 25 reported separate DW estimates for treated and untreated paraplegia, unlike reference 27. This could be corrected either by presenting both DW estimates from reference 25 or by dropping the comparison.

7. Table 3 reports that reference 27 had a DW for stroke of 0.266. It should be clarified that all estimates in the column are for chronic sequelae of stroke rather than for acute stroke, for which the GBD 1990 DW is 0.920 (see Hong and Saver 2009 for clarification).

8. The DW estimates for stroke from reference 28 (Hong and Saver 2009) are not the same as those reported in the publication.

9. References 22 and 47 are the same and one of them should be deleted.

10. Tables 1 and 3 list 2010 as the year for reference 25 (Salomon et al. 2012), 2004 for reference 43 (Kruijshaar et al. 2005), 2009 for reference 28 (Hong and Saver 2010), and 2010 for reference 44 (Lai et al. 2009). The other references in the tables appear have the correct years of publication listed.

Discretionary Revisions

If the search is redone, it would be a good idea to include articles published in the first nine months of 2013. Rehm and Frick (2010) described a protocol using the CLAMES health description system in combination with expert valuations to estimate DWs as part of a new US DALY study. The DW estimates from that study were reported in a 2013 publication by Rehm and Frick. The article would be more useful to readers if that study were to be referenced.

The section on Comparison of disability weights document the wide range of DW estimates for specific conditions. Following sections provide some information about reasons for these differences, but not as much as is desirable. In particular, it would be helpful to compare mean DW estimates using different valuation methods. It is well established in the literature, for example, that DWs using VAS methods tend to be much larger in magnitude than the equivalent weights calculated using choice-based valuation methods, as a result of respondents stretching out weights to fill in the range from 0 to 1. Looking at the
findings from reference 40, which used VAS, appears to support that generalization. The Valuation methods section acknowledges that the VAS estimates tell one nothing about trade-offs people are willing to make between health states and death and only indicate relative desirability among health states. Those statements, although accurate, do not inform readers about the direction or magnitude of bias that can result from using unadjusted VAS scores to estimate DWs.

An important contribution of this paper is the side-by-side comparison in Table 3 of DWs for up to 8 conditions from the GBD 2010 and GBD 1990, as well as other studies. It would be easier to follow if the studies were grouped by type of study rather than chronology, listing all of the comprehensive studies first, including both GBD studies, followed by the narrower studies. It would also be helpful if the authors were to explicitly compare the weights from the two GBD studies and address the reasons for the differences. The GBD studies, along with their extensions such as the Dutch DW and Australian BODI studies, are the most common source of DWs for estimation of DALYs and it would be appropriate to give those studies greater attention than the minor, condition-specific studies. However, as noted above under the Minor Essential Revisions, some of the DW estimates presented for the two GBD studies are either erroneous or misleading. It would be useful to have a separate paragraph comparing and contrasting the findings from the two GBD studies along with the Dutch and Australian studies. The differences in estimates for a number of conditions, such as stroke and depression, appear quite large. What are the implications of these differences?

The use of unadjusted EQ-5D utility weights as DWs may be inconsistent with the objective of DALYs. DWs in GBD and other population-based studies are calibrated to ensure that YLD from all causes add up to the total YLD in the population. There is no equivalent constraint on utility weights. Consequently, it is not surprising that DWs estimated based on EQ-5D weights are often much larger in magnitude than the equivalent GBD DWs. Disease specialists often explain such differences in terms of the GBD weights understating the importance of “their” conditions. However, another contributing explanation to such differences is that GBD DW estimates are required to be mutually consistent, whereas DW estimates from narrow studies do not. Thus, although it is tempting and convenient for disease specialists to use EQ-5D or other utility weights as DWs, the resulting estimates may not necessarily be commensurate with existing BOD estimates for other conditions. It should be noted that this is not a necessary feature of disorder-specific DW studies. Indeed, reference 28 (Hong and Saver 2010) ensured consistency of their DW estimates for post-stroke health states with GBD DW estimates through a normalization process.

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 I have no competing interests