Inter-Rater Analysis: Screening of Mental Health Innovations

Rationale for this analysis

The research group of the EA Human Welfare Project has agreed that in order to maximize our confidence in the results of the MHI screening, each intervention will be screened by at least three people. However, this is only helpful if all those who screen share a common understanding of how exactly they evaluate these interventions (i.e. which measures they extract from the available information/ how they estimate or judge them). It seems conceivable that to some extent there is disagreement and/or subjective components to this screening, which might cause (intended) differences in judgement. Nonetheless, these should be minimized to increase objectivity and validity of our analysis. This analysis is meant to quantify the degree of disagreement within the training set containing the first six interventions.

Measures to be obtained

1) Overall IRR: Screened "in" or "out"

- As defined in the screening document, interventions are screened "in" if their costs score multiplied with their central benefit estimate is greater than 6; otherwise, they are screened "out".
    - Add a new categorical variable to the document - screened "in" or screened "out"(1 vs 0)
    - On this basis, calculate e.g. Fleiss' Kappa (extension of Cohen's Kappa for multiple raters)
        - Assumes, however, that raters have been chosen at random - not entirely sure if that significantly affects our calculation

Also potentially interesting:

2) IRR of End/Continue evaluation → again, via e.g. Fleiss' Kappa

3) IRR of central benefit estimate → potentially via intraclass correlation coefficient, as variable is no longer categorical

4) IRR of cost estimate → also via intraclass correlation coefficient

5) IRR of intuitive score → intraclass correlation coefficient

6) IRR of upper bound 80% CI → intraclass correlation coefficient

Additionally relevant

For other aspects, calculating a IRR does not seem reasonable. Nonetheless, we should qualitatively look at differences in judgement and try to understand where they stem from. From my point of view, this would especially pertain to:

- Organization fundable/ intervention fundable as new organization?
- Percent considered beneficiaries and, relatedly,
- Costs per beneficiary

It seems reasonable to me to discuss this with two or three people and write up a clearer guidance based on the findings. This group should also take into account the notes on improving the evaluation scheme.

Additional thoughts

We should encourage raters to provide their calculations so we better understand roots of disagreement.

In general, do we encourage taking into account external information? E.g. StrongMinds has apparently estimated their cost-effectiveness in 2018. If number is at hand, should it be included?