Reviewer's report

Title: Horizontal equity and mental health care: A case vignette inter-rater reliability study

Version: 1 Date: 11 January 2012

Reviewer: Geir Pedersen

Reviewer's report:

General comments

The current manuscript focuses on a very important topic. Seldom are criteria for so important decisions within health care systems so ambiguous and hard to operationalize, combined with either consequences of high costs or unnecessary suffering from the patients if decisions are made wrong. It is therefore important to both disclose current practice of such procedures, as well as the level of equality of such evaluations across health care centers.

This is also a short, well written and focused manuscript. With some further improvements it would be suitable for publication, and a welcomed contribution of importance, both clinically and politically within health care systems.

Major Compulsory Revisions:

1:

This study does not apply the most proper method for data analysis: In this study Intra-Class Correlation (ICC) is applied to estimate rater agreement. This method is in line with classical test theory, which partitions observed score variance into two parts; one systematic, seen as true score variance, and one random, seen as error variance. If three or more variance components are confounded, ICC comes short since it cannot estimate them simultaneously in one analysis. Within the current study, ICC based on rater scores will hide possible variance due to systematic differences between centers. Likewise, ICC based on center scores will hide variance due to the rater effect. Since this study contains two object of measurement it has multiple sources of error. The most proper theory to address such issues is Generalizability Theory (G-theory).

Ideally the object of measurement should be both patients, represented by case vignettes, and health care centers. Patients are crossed with raters and centers, and raters are nested within centers. In G theory terms the design will be noted: “p x (r:c)”. Moreover, there were different numbers of raters within centers, making this a so called unbalanced nested two-facet random effect design. This is the same design as in the article number 10 in the reference list of the current manuscript. I would strongly suggest the authors to consider a re-analysis with a G-study, and subsequent D-studies to estimate the optimal number of observations necessary to obtain acceptable reliability.
The correct terms are ‘intraclass correlation’ as to name the measurement method and ‘intraclass correlation coefficient’ for the measure. Not ‘inter-class correlation’.

Ratings based on group consensus is a very special kind of measurement. Here, several persons are discussing their opinion, and bargains for their view. By this, such measurements might be highly biased, where the ‘winner’ usually is the most dominant and rhetorically skilled. This has to be discussed.

The best estimate for each vignettes ‘true score’, or expected score is the mean of all ratings. Mean of several ratings will always be more reliable than one rating. Thus, if the mean scores from persons within each center were used as the center score, the reliability between centers will be higher than for single raters, unless there is some systematic difference between centers in how they interpret and use the measurement scale (see point 5). This has to be discussed.

Minor Essential Revisions:

Make sure the reference list is formatted in line with the standard of the journal.

In the first paragraph of the results section, the authors write: “Our expectations of a strong association between number of referrals and number of staff members involved in referral assessment, was not confirmed…” This expectation should be stated and discussed earlier in the manuscript.

From table 2 it seems to be two groups of centers: One group (1, 9, and 13) with relatively high refusal rate and around 50% high priority, and one group (5, 7, and 8) with low refusal and high rate of high priority. Notice should also be given to group 5 and 13 since they seem to use the scale rather dichotomous. This should be given some attention.

Discretionary Revisions

Figure 1 might be some confusing. There are listed four legends, including ‘Missing data’, but the bars only contain three distinct ‘colors’. My guess is that vignette 5 has approx. 13% missing data and no ‘Low priority’, and further that only vignettes 8, 10 and 17 has missing data, with white bar-area at the top of the bars. That is, that white areas at the top of the bars indicates missing, and white areas within the bars indicate ‘Low priority’.

**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.