Reviewer's report

Title: The Improved Clinical Global Impression Scale (iCGI): development and validation in depression

Version: 3 Date: 16 August 2006

Reviewer: Herbert Fliege

Reviewer's report:

The authors made a thorough revision of their paper and addressed almost all issues. I think it is principally acceptable for publication now, but I have one final point of critique.

With regard to my first suggestion (including one more rater in order to have the same n of raters for all groups under comparison) they bring forward three arguments against it. The first and obviously relevant one is economical. That is quite understandable. As the authors addressed all other issues I can well accept that argument. I also I would not expect any great changes in the results. (I am afraid I don’t get the message of the second argument. Why would the test have to be repeated if the rater were an additional person?). However, my main concern lies with the third argument. It is now even included in the methods section and I cannot quite agree with it.

As far as I can see the reliability design indicates a two-way random effects model of ICC calculation, treating both, patients and raters, as a random. In each of the study groups, the same rater rates each case. The raters here should be conceived as a random sample of possible raters. Then ICC estimates the reliability of the larger population of possible raters, e.g. future users of the rating scale. The basis of ICC is the variance between subjects by the overall variance. I think we agree so far. Do correct me if I am wrong.

Now the authors state that it makes no difference whether the variances are divided by 3 or 4 or any other number of raters. Indeed some studies yielded that increasing the number of raters had little impact on ICC reliability (Hamdy et al, Medical Education 2003). However, statistical simulation studies showed that a larger number of raters could enhance precision when the ratio of between-rater-variance / error-variance was great. Furthermore, the use of identical numbers of raters may have other advantages (Saito et al: Effective number of subjects and number of raters for inter-rater reliability studies. Stat Med 2006, 25, 1547-1560.).

So if the authors decide not to include any further raters, that is all right with me, but I suggest not emphasizing the statistical argument in the current form, as there may be good arguments against it.

What next?: Accept after minor essential revisions

Level of interest: An article of importance in its field

Quality of written English: Needs some language corrections before being published

Statistical review: No

Declaration of competing interests:

I declare that I have no competing interests.