**Author’s response to reviews**

**Title:** Use of a self-rating scale to monitor depression severity in recurrent GP consultations in primary care - does it really make a difference? A randomised controlled study

**Authors:**

Cecilia Björkelund (cecilia.bjorkelund@allmed.gu.se)

Carl Wikberg (carl.wikberg@allmed.gu.se)

Jeanette Westman (jeanette.westman@ki.se)

Eva-Lisa Petersson (eva-lisa.petersson@vgregion.se)

Maria Larsson (maria.eh.larsson@vgregion.se)

Malin André (malin.andre@uu.se)

Robert Eggertsen (Robert.eggenrtsen@vgregion.se)

Jörgen Thorn (jorgen.thorn@vgregion.se)

Hans Ågren (Hans.agren@gu.se)

**Version: 1 Date: 23 Oct 2016**

**Author’s response to reviews:**

Dear Editor,

Thank you for reviewing our manuscript FAMP-D-16-00265 - Use of a self-rating scale to monitor depression severity in recurrent GP consultations in primary care - does it really make a difference? A randomised controlled study.

We have followed the recommendations for improvements of the article, and below we answer point by point the reviewers’ comments. We also submit a version of the new manuscript with track changes, as well as a clean copy new manuscript. We thank the reviewers for very competent comments and now hope that our revised copy of the manuscript will be accepted for publication.
Reviewer 1:

Summary: A randomized trial of depressed patients was conducted to determine if a self-assessment scale administered at a GP visit at least 4 times over 3mo as compared to usual care improved depressive symptoms, medication adherence, sick leave, and general health. 258 depressed patients from 91 GP practices were enrolled. Randomization to treatment or usual care was performed by GP and produced statistically valid results. The outcomes were measured at 3, 6, and 12months. The authors found no difference in the Beck Depression Index, nor general health as assessed by questionnaire. There was a slight difference in medication adherence favoring the treatment group.

Strengths:

* Clearly written
* Adequately explained methodology
* The authors used valid instruments to assess the outcome measures
* Randomization was adequate
* Conclusions are supported by the data

Weaknesses:

* Recurrent vs. first episode was not controlled. Patients in various stages of depression were enrolled. There is some evidence to suggest recurrent depression has slower remission (see reference #29)

Answer: All patients enrolled were diagnosed with a new episode of depression. We have extended information on recruitment in methods section, page 6, line 36: Inclusion of patients: “Study participants were patients aged 18 and up who visited the PHCCs and were identified and diagnosed by a GP with a new episode of mild/moderate depressive disorder [14].”

We did not exclude patients with recurrent depression, which is also shown by the fact that 22 % (intervention group) and 32 % (Treatment-as-Usual group) already were on an antidepressant as maintenance medication. The RCT design and its accomplishment in ordinary primary care context together with the similarity of antidepressant medication in the intervention and control
group makes it most likely that the proportion between first and recurrent episodes is the same in both groups.

We have added to discussion (Page 15, now 15-16):

“All patients included had a new episode of mild/moderate depression, but there was no selection based on recurrent or first episode. There is some evidence that recurrent depression has slower remission [29], but according to the similar frequency of maintenance antidepressant medication in both intervention and TAU groups at inclusion, we can anticipate that the rate of recurrent depression disorder was the same in both arms.”

“Randomization was performed by GP and the authors were very clear about this. However, it is less clear whether their power calculation results in the necessary number of patients or GPs per group.”

• Answer: We have changed the text in power calculation to clarify that we needed 105 patients in each group to obtain desirable power. We changed the word participants to patients on page 10 line 57 and page 11 line 1.

• “I suspect there may be some grouping effect on patients via the GP randomization. This was not accounted for in their logistic regression models. A stronger analysis would be to use generalized estimating equations or a generalized linear mixed model to account for the correlation between patients treated by the same GP. I would be satisfied with recognition of this and discussion as a limitation.

Answer: We have recognized this and added a statement in discussion that highlights the statistical method:

“A limitation is also that we did not use more advanced statistical models to account for the correlation between patients treated by the same GP and that handle all observations of the patient instead of pairwise comparisons (e.g. mixed models).” (Page 15 around line 47- now on page 16 in revised manuscript)

Comments:

• There was little difference in terms of the number of visits to GPs during the 3mo study period. This suggests the TAU group also received equal attention. There are numerous studies pointing to the fact that care coordination or collaborative care management produces improved results. One of the hallmarks of collaborative care management vs. usual care is more frequent contact with healthcare professionals. In this study we see the usual care group had almost the same number of visits as the study group that had 4 GP visits in 3mo as part of the protocol. I find this uncommon - most usual care won't have this
much attention (perhaps Sweden practices more like collaborative care management at baseline - a good thing!). This may limit the generalizability of their findings and should be discussed. Nevertheless, it supports their conclusion that simply including a self-assessment scale wasn't part of the "special sauce" of care management that produces good results.

Answer: We interpret the high number of visits also in the TAU group as a result of the nowadays widely conducted person centered care in Swedish primary care, well known among Swedish GPs. The contacts are a result of agreement between doctor and patient in every individual case. We have added a sentence on page 15, bottom section: “We interpret this as an indication of the person centred care presently widely conducted in Swedish primary care”.

If looking to save space, I'm not sure that Table 3 contributes much. The same holds true for Table #2 (although there was one significant difference - it could just be commented on in the results section).

Answer: For now we are happy if the two tables remain, but if space is sparse, we naturally are willing to exclude them.

Overall, I recommend this study for publication. The finding that simply including a self-assessment instrument didn't change remission of depressive symptoms is useful information.

Reviewer 2:

Depression is highly prevalent in primary care. GP's have a central role in the recognition, diagnosis and management of depression. There have been only a few studies that had evaluated the effect of recurrent use of self-rating scales regarding depression so that's why the final result of this study could be interesting for the readers of the BMC Family Practice. It was very interesting (probably because of the not appropriate statistical analysis) that there were no significant differences between the intervention and control group in depression severity reduction or remission rate, change in quality of life, psychological wellbeing, sedative prescriptions, or sick leave during the whole 12-month follow-up.

• “I think it would be better if the employment status would be categorised as working/studying, unemployed or retired. I do not think that unemployed and retired patients should be in the same group.”

Answer: The reason for this categorization is that we were interested of the sick-leave outcome, and since retired patients and unemployed cannot be on a sick leave, the groups were structured on this ground. If we would evaluate epidemiological issues it would have been more appropriate with the proposed categorization.
If we look at the educational level my suggestion that it is better if they make 3 groups as lower educational level (primary school or vocational school), middle level (high school) and higher educational level (college or university).”

Answer: We have made 3 groups in accordance to the reviewer’s comment (Table 1).

If we look at the statistical analysis my first and foremost problem is the multiple comparisons situation encountered in the paper (more precisely, the lack of its handling). Table 2 represents 6 comparisons, Table 3 contains 3 and Table 4 represents 12. In addition to that, all three figures (from Figure 2 to 4) represent 3 comparisons, bringing the total to 30. When deciding at 5% significance level, 30 independent comparisons mean an expected 1.5 significant results even if there is no difference at all in any of the comparisons, and the probability of having three or more significant results purely by chance is 18.8%. Thus, even that is quite possible that all three significant results obtained by the authors are all Type I errors. The authors made no attempt to provide protection against multiplicity (even though low p-values indicate that at least a part of the significant findings would remain significant even after correction.) A far more serious problem is that they haven’t mentioned it, not even in limitations: while it is not per se unacceptable to perform so high number of comparisons (especially if they were prespecified), it should be made very clear to the reader that this is the situation, and results can be only considered exploratory, which need further confirmation.

Answer: We have added in discussion (page 15, line around 47):” Further, the high number of outcomes, assessed at baseline and at 3, 6 and 12 months, represents a multiple comparisons situation with possibility of type I errors. Although the low p-values indicate that the few significant findings not are by chance, the results need further confirmation in other studies.”

Figures 2 to 4 represent very basic analysis, which should be made more sophisticated. It has two aspects, first, the visualization needs improvement so that we can obtain more information from the figures (258 patient even allows a spaghetti plot, or at least boxplots should have been provided instead of means etc.).

Answer: We have included box plots in the figures (Fig 2, 3, and 4).

More important is the analytical approach: simple paired t-tests should be replaced with more advanced methods, possibly such that handle all observations of the patients (and not pairwise). Mixed effects models [1] seem to be a logical choice in this situation. Locascio and Atri [2] and
Gibbons et al [3] provide an accessible introduction to this topic (using examples from a close fields) in addition to the well-known textbooks from Diggle [4] and Fritzmaurice [5].

Answer: We are aware that more sophisticated statistical methods could be applied. However, in the study protocol, we specified the statistical methods that we would use, based on discussions with academic statisticians. At that time more conservative methods were proposed. Considering the limited time for re-submission, we chose the alternative presented by the reviewer 1, and we have added in the discussion (page 16): “A limitation is also that we did not use more advanced statistical models to account for the correlation between patients treated by the same GP and to handle all observations of the patient instead of pairwise comparisons (e.g. mixed models)”.

My opinion is that the submitted article needs major review but it is a good and well-designed study to be published in the BMC Family Practice.

We now hope the manuscript will be accepted for publication.

Best regards

For the authors

Cecilia Björkelund

Carl Wikberg