Reviewer's report

Title: Clinical Validation of the Portuguese Version of the CANFOR

Version: 2 Date: 20 December 2012

Reviewer: Miao-Yu Y Tsai

Reviewer's report:

Comments for Authors
Referee Report on “Clinical validation of the Portuguese version of the CANFOR”

The authors focus on assessing the validity and reliability of the Portuguese version of the CANFOR. Through the agreement measures and correlation coefficients, they have proved that this version can be a reliable tool for evaluating the needs of care in prisoners with mental disorders. The problem is interesting and the proposed solution has merit. However, I have the following concerns.

General concerns
1. About inter-rater reliability:
   (a) How many raters are there in all interviews? The authors mentioned about raters or researchers on page 9: “while one researcher conducted the interview and scored the different CANFOR domains, the second one sat silently in the room and scored CANFOR independently. These roles were alternated in every interview.” This paragraph seems to indicate that there are only two raters in all interviews. However, in Table 3 and Tables 6-8, all show that the raters in interviews include users and staff.
   (b) Why do the authors use service users as experimental subjects rather than prisoners?
   (c) The data include 96 individuals (service users) and each user is scored by two raters in each interview. According to above statements, are there 96*4=384 scores (rater 1 conducted the interview and rater 2 sat silently, and then the roles for two raters were alternated) for this analysis?
   (d) In Table 4, what is the agreement coefficient? Please cite the reference.
   (e) In Table 4, the scores are rated into two categories (“unmet need” and “other score”) as the responses in this study. Why are the two agreement measures (i.e. agreement and kappa coefficients) calculated based on separated data (unmet need score only and other score only)? Is there something wrong?

2. About test-retest reliability:
   (a) In the second paragraph on page 9, the authors mentioned that “a second CANFOR rating of service users was made between one to two weeks after the first interview. This was conducted by the same rater who performed the first
interview. A total of ninety-nine service users were interviewed twice. Is the first interview the same as that mentioned in the first paragraph on page 9? If so, why are there 99 service users rather than 96? If not, does it mean that all these 99*2 interviews are new conducted? Therefore, are there total 99*8 scores for the first and second interviews (i.e. there are 99*4 scores for each of the two interviews)? The first two paragraphs on page 9 are unclear and confusing.

(b) As Q1 (b), (d) and (e), why do the authors use service users as experimental subjects rather than prisoners? What is the agreement coefficient in Table 5? Why are the two agreement measures calculated under separated data (unmet need score only and other score only) in Table 5?

3. About convergent validity:

The Mann–Whitney U test is a non-parametric statistical hypothesis test for assessing whether two independent samples come from different distributions or whether the medians between two groups are different. However, the scores of CANFOR domains and the BPRS scales’ scores are based on the same observations, and therefore these scores of CANFOR and BPRS are not independent. Furthermore, in Tables 7 and 8, the significant result cannot imply the association between CANFOR and BPRS.

4. For crosstables without zero cells in some CANFOR domains, the weighted kappa (Cohen, 1968) for ordinal outcomes (“0-no need”, “1-met need” and “2-unmet need) is more appropriate.

Specific comments

1. The reference number [30] on page 9 and [29] on page 10 should be interchanged.

2. On page 12, -line 11. To use “type I error” instead of “error type I”.

3. On page 12. The sentence “In order to avoid the presentation of spurious significant associations (error type I) it will be displayed, in the Table 7 and 8, only the domains with high significant association between unmet need score and BPRS’s scales score.” is not correct. This strategy is only to show that a hypothesis testing is more conservative by using a significance level of 0.01 but cannot avoid spurious significant associations. Please modify it.

4. On page 25. In Table 6, it is fine to report p-values in parentheses below the values of Spearman’s rho directly and not necessary to put them in another cells. Besides, one p-value (value is 0.054) is greater than 0.05 and the authors don’t use the notation “ns” to replace it.

Level of interest: An article whose findings are important to those with closely related research interests

Quality of written English: Acceptable

Statistical review: Yes, and I have assessed the statistics in my report.