Author's response to reviews

Title: Creating scenarios of the impact of COPD and their relationship to COPD assessment test (CAT) scores

Authors:

Paul W Jones (pjones@sgul.ac.uk)
Maggie Tabberer (margaret.x.tabberer@gsk.com)
Wen-Hung Chen (wen-hung.chen@unitedbiosource.com)

Version: 3 Date: 13 May 2011

Author's response to reviews: see over
13 May 2011

Timothy Shipley PhD
In-House Editor
BMC PULMONARY MEDICINE

Dear Dr Shipley

Thank you for providing the reviewers comments for our manuscript “Creating scenarios of the impact of copd and their relationship to copd assessment test (CAT) scores". The manuscript has been revised in consideration of the comments and below a point-by-point response is provided to each of the comments, in red text.

We hope you that you find our revisions to the manuscript suitable to warrant publication, and please do not hesitate to contact me should you have any queries

Sincerely,

Paul Jones PhD, FRCP
Reviewer 1: Raffaele Antonelli-Incalzi

Reviewer’s report:
The suggested three level comment does not fit my view of the study and related suggestions. Then, I will provide a single comment. The elegant study by Jones et al aims at performing a correspondence analysis. My view of this study between the CAT questionnaire and the parent questionnaire SGRQ. Collaterally, but not marginally, it sheds light on the effects of COPD on the health status.

The study uses a sophisticated methodology and demonstrates an excellent agreement between CAT and SGRQ. However, this is to some extent an expected finding given that CAT has been derived from the SGRQ.

The reviewer is not correct on this point – the CAT is an entirely new questionnaire developed independently of the SGRQ.

Indeed, the study does not add to the reader’s knowledge of the rationale for using CAT, which is the object of the previously described development and validation studies. The added value of this study really is in the demonstration that not uncommonly even mild COPD severely impacts health status. This observation complements that on the relationship between COPD severity and impairment of functional capabilities rated by a multidimensional assessment (ERJ supplement 2009).

1. Physicians in charge of respiratory patients need some practical information about how the rating of health status affects their daily practice. Such information cannot be easily drawn from the present study which lies more in the field of methodology than in that of clinical application.
   We have included more information in both the introduction and the discussion which we hope will allow physicians to more readily appreciate the value of descriptive scenarios when considering health status findings.

2. The prospective assessment of the four CAT categories vs. clinical outcomes and needs of care will clarify this issue.
   We have addressed this comment in the discussion.

3. In other terms, these data suggest that the CAT has classificatory properties, but this does not automatically mean that it is clinically useful. To sum up, the study is of interest, but it might be converted in a short methodological report.
   We have tried to shorten, but we feel it is important to describe our methodology transparently. We feel that our manuscript is very well suited to the readership of BMC Pulmonary Medicine.
Level of interest: An article whose findings are important to those with closely related research interests

Quality of written English: Acceptable

Statistical review: Yes, but I do not feel adequately qualified to assess the statistics.

Reviewer 2: Julie Morris

Reviewer’s report:

Statistical Review:

This paper describes the derivation of ‘qualitative’ labels for categories of CAT scores based on a mapping of CAT scores to SGRQ-C scores. There are some aspects of the statistical methodology used, their interpretation and the presentation of results that need to be addressed. The mapping of CAT scores to SGRQ scores using logits is appropriate. However, a more comprehensive application of the Bland and Altman technique should be carried out, and a more detailed interpretation of these results would be advised.

Major Compulsory Revisions:

1. A Bland and Altman plot is constructed, but the mean bias and 95% limits of agreement are not calculated. The latter are important because they indicate the ‘extremes’ of any discrepancy between the two scoring systems. A clinical judgment should then be made as to whether these limits correspond to acceptable discrepancies. From Figure 3 I would guess the 95% limits of agreement are approximately +/- 20. This means that the two scoring systems could differ by around 20%. It is this percentage that then needs to be assessed as being clinically acceptable or not.

We thank the reviewer for asking us to add the limits of agreement – it was an omission. We give a more detailed response to the inferences drawn from the B&A plot below (in italics), but point out that it was not the purpose of this plot to demonstrate direct equivalence, but to justify that, at a population level, this method of mapping should be reliable across the range of CAT/SGRQ scores. We have provided more text to illustrate the point made immediately below.

**Bland and Altman plot**

The purpose of the Bland and Altman plot is to assess the agreement of two instruments. As stated in the manuscript, the difference should be small across the range mean score and have no or very small correlation between the difference and the mean. Based on the two criteria, we believe Figure 3 has shown evidences of
stable agreement between the two instruments. The mean bias is 1.22 out of a 0-100 score scale. The lower 95% limit is -22.5 and the upper 95% limit is 25.2. These are the extremes of any discrepancy. We believe that the score agreement between the two instruments should not be judged and determined by few extreme discrepancies, but rather by overall performance. Between CAT and SGRQ, 31% of the score differences are less than 5 points (difference of 5%), 60% are less than 10 points (difference of 10%), and 90% are less than 20 points. Based on these numbers, we believe there is a substantial agreement between the CAT and SGRQ.

Minor essential revisions:

2. Methods. Comparison of CAT and SGRQ scores. ‘The correlation between SGRQ and CAT scores is good (r=0.82) [3]’. It is not clear why this correlation value is quoted. Reference 3 quotes correlations of 0.8 in stable patients and 0.78 in acute patients, ie does not quote a correlation of r=0.82. We apologise for the typo. We have corrected the paper to explain that the correlation in stable patients is 0.80.

3. In the authors’ description of the Bland and Altman plot (paragraph ‘Correlation with SGRQ’), the over- and under- estimation is presented in terms of CAT original units. It would seem more appropriate to describe these differences in terms of adjCat, ie comparable SGRQ and CAT units. We can see the reviewer’s point, but a CAT user, doesn’t see adjCAT scores, only CAT scores. We have addressed this comment in the “Correlation with SGRQ” section of the results, and have included both the CAT and adjCAT values.

4. Table 2. It would help the reader if the CAT severity bands were labelled with the corresponding CAT score ranges (eg. Low impact = 1 to 10) etc). We have added the severity bands to the table.

5. Figure 1. It is not clear how the co-ordinates of the dotted lines were chosen. It would see more appropriate to place them at the thresholds of the CAT score severity bands. We apologise, for some reason the dotted lines were moved in the version seen by the reviewer. The figure should have been in the form the reviewer recommended and has been corrected.
6. Abstract. ‘A Bland and Altman plot showed a consistent relationship (r=0.82) between CAT scores and scores obtained with the St George’s Respiratory Questionnaire for COPD……’. But the quoted correlation is not connected to the Bland and Altman analyses (see above).

   Yes we agree. It’s confusing and we have removed the reference to r value from the abstract.

   **Level of interest:** An article of importance in its field

   **Quality of written English:** Acceptable

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

---

**Reviewer 3:** David Michael Meads

**Reviewer's report:**

**General comments**

The paper describes a novel method of attributing clinical meaning to scores on a health status questionnaire. The report describes an interesting combination of issues surrounding clinical meaning/importance of questionnaire scores, Rasch analysis and mapping between measures. However, as a consequence, clinicians may find it a challenge to grasp the overall aims. Since the research describes novel methodological work with outcomes questionnaires it might be better appreciated in an outcomes-focussed journal.

**Discretionary Revisions**

1. More explanation perhaps required on the clinical implications of the work and its practical use

   We have included more information in both the introduction and the discussion which we hope will allow physicians to more readily appreciate the value of descriptive scenarios when considering health status findings.

**Minor Revisions**

1. "The key property of this type of scale is the assumption that, for an item of given severity, a patient will have a high probability of responding positively to items that indicate lesser severity than the item in question and a lower probability of responding to items that reflect greater severity." (Page 4). A caveat should be added that this is the case when positive responses denote the presence of an impairment or disability.

   Agreed, we have added a caveat to the sentence
2. “A conversion table allows CAT scores to be converted to logits or vice versa.” Please state if this is available with the original publication.

We have provided the full conversion table as an appendix and retained the abbreviated version of the conversion table as Table 1.

3. Description of the data should be moved from Results to Methods.

Done

4. Can the authors justify the combination of multinational data? Would the scenarios be the same if they were derived using individual country data – aside from sample size issues?

In the CAT development we used tests of differential item functioning (DIF) within the Rasch model to test for any differences in response to each item that was dependent on country. There was no DIF in any of the items that make up the CAT. We also know from development work with the SGRQ, that the empirically derived weights for SGRQ did not differ between the 6 countries in which this was tested (Quirk et al. Eur Respir J. 1991;4:167-71)

5. A statement (and reference) is needed about sample independence of Rasch to help justify use of two different data samples for the mapping.

We have added a statement to this effect, and a reference, to the discussion. As the reviewer knows the literature can be very mathematical and not easily accessible for the average reader. We hope we have chosen one that is comprehensible (and explains other rasch terms).

6. Can the authors offer an explanation for the low correlation (R=0.16) between the measures? Include this in the discussion.

Ideally the correlation would be zero, indicating that the relationship between the CAT and SGRQ remained exactly the same across the scaling range if the two instruments. In this there was a small significant slope, showing (as we say in the paper) very slight differences between the two instruments at the extreme ends of their scaling ranges.

7. Mention should be made of the likely significant uncertainty surrounding the categorisation of patients into scenarios since this combines uncertainty around the original thresholds devised for the CAT (Mild – Severe) and then uncertainty in the mapping of items from the SGRQ-C to the CAT.

We have added a statement to this effect to the discussion
8. References 7 and 10 are the same. In general, I think the paper would benefit by referencing a wider range of studies.

*We have addressed this comment, and a number of additional references have been cited in line with other comments received*

**Major revisions**

1. References should be provided on the validity of the methods described, and on the validity of the mapping process using Rasch. What previous work has been conducted in this area? There are examples of the use of Rasch analysis to map between measures:


   *We have added a statement to this effect, and the suggested references, to the discussion.*

2. More details are required on the mapping process. It’s not very clear whether the mapping process took place on the basis of item logit positions or item response category positions. Presumably it’s possible that the mild categories of some questions may map to the mild CAT level while the severe response category for the same item may map to the severe CAT level?

   *We have clarified that point in Methods.*

3. The categorisation of patients into scenarios should be tested and validated – for example by reverse checking patient CAT level to SGRQ-C responses and level of agreement or prediction error. If this is beyond the scope of the current work then it should be noted that future research should seek test the validity and robustness of the mapping and categorisation of patients using data where SGRQ-C and CAT are collected concurrently.

   *Yes we agree, but it was beyond the scope of this work. We have added a statement to this effect to the discussion.*

**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.