

Reviewer's report

Title: Quality of chronic disease care in general practice: the development and validation of a quick measurement tool

Version: 1 **Date:** 2 November 2006

Reviewer: Mark Atkinson

Reviewer's report:

General

Most often approaches to quality of care assessment occur at the level of the individual, with individual results then being aggregated for group comparisons. The proposed interview assesses the characteristics of a practice (across all patients) and so differs from typical measurement approaches. However, the title leaves room for interpretation about who is being 'measured' and it would be helpful if phraseology clarified that the measure is designed for completion by a physician about their overall practice. Similarly, the 'Aims' or purpose of the measure should be sharpened (Abstract (pp 3, para 1), Introduction (pp 5, para 2) and Methods sections (pp 6, para 3)).

Major Compulsory Revisions (that the author must respond to before a decision on publication can be reached)

Methods:

Sample: Based on information gathered from chart reviews, did the practice case mix differ across participants? Were observed differences in the case mix of clinicians associated with any differences in item scores on the GPCCI? This is particularly important to know if the Interview is to be used to compare respondents, as results may need to be stratified or statistically controlled for patient type to standardize such comparisons.

Chart data coding and aggregation:

- 1) The number of items on the coding schedule differs from the number of items on the GPCCI and comparison of content coverage between the two is not possible without a table comparing the details of both, mapping the content of Interview and coding schedule to one another.
- 2) It is unclear how the data gleaned from patient charts were aggregated. For example, was a summary rating for each coding category derived by determining the proportion of patient files where the practice characteristics was observed?

Data collection and analysis:

1) Pp10, para 1 and pp12, para 2, describe how the 9 items found to reduce the internal consistency of the 'scale' were dropped. Which items were removed? Of note, striving for internal consistency between items may be most important when the resulting (true) score is best approximated by the statistical intersection of all items on the scale. This is not always the case, and precision of the estimate may be indexed by a number of independent items. Thus removal of items before determining whether they independently predict the validation criterion may be premature. I suggest that the original pool of interview items be entered as independent variables into regression models 'predicting' results of the clinical chart audit. Those variables that independently predict the chart results may provide an 'index' of care which is more useful and psychometrically convincing than a 'trait of care' approach to the construct of measure. (For a discussion of this topic, see: Atkinson MJ & Lennox. (2006). Extending Basic Principles of Measurement Models to the Design and Validation of Patient Reported Outcomes. Health and Quality of Life Outcomes. <http://www.hqlo.com/content/4/1/65>)

2) Pp 10, para 2, it is unclear how items in the Interview schedule were scaled and scored. (For example, were scores computed as a weighted mean across items? etc.). An appendix containing the interview schedule and instructions would be helpful.

3) Pp 11, para 1, it is unclear why items with the same scores across raters were removed from the inter-rater reliability calculations - couldn't a % agreement statistic be used instead. A better argument for their removal may be that the occurrence of 100% reliability across raters most often occurs for items in which the patient did not have the condition, in other words the item was seen as not relevant by all raters to a particular case*. In essence, by removing items, the authors were stratifying the statistical analyses by the presence/absence of the disease(s) (including comorbidities). If the authors are in agreement with this critique, the section needs reworking.

*note: problems of skew when working with particular items is suggestive that a Multiple Causal Indicator Model may provide a better fit between the measurement observations and the statistical estimate, see reference above

Results:

1) Without knowing how items on the GPCCI or coding scheme were scaled (e.g., categorical, ordinal or interval etc.) and/or aggregated it is not possible to evaluate whether the statistics used provide appropriate estimates of covariance or mean group difference. Please provide additional information on the Interview Schedule and coding scheme.

2) A table should be provided containing a description of the Interview items, their scaling/response options, and their distributional characteristics (mean, median, skew, SD etc).

Discussion and Conclusion:

1) Pp 14, para 2: The GPCCI is described as an acceptable tool... Acceptable for what purposes? The better performance of the overall scale compared to the disease specific subscales could be simply a function of the number of items (or possibly respondents) used when estimating scale reliability (due to elimination of persons without the condition).

2) The internal consistency of disease specific subscales differed, with asthma and heart disease schedules being substantially lower than the diabetes schedule. This observation should be discussed, although it is unclear if internal consistency and factorial coherence are desirable characteristics of this particular instrument.

3) Observations about inter-rater reliability and correlations between the coding and GPCCI results should be discussed in terms of the potential sources of error, with attention paid to where sources of error cannot easily be disentangled in the current study (for example coding error from chart error).

4) Pp 15, para 3: The current findings are presented as a strong demonstration of the validity of the GPCCI. However, there are a number of types of validation, and only one or two of them have been touched on by the current study. The 'preliminary' nature of this paper, the first in a series of publications on the validity of the measure, should be emphasized a couple of times in the abstract, body and discussion of the paper.

Minor Essential Revisions (such as missing labels on figures, or the wrong use of a term, which the author can be trusted to correct)

The term 'internal reliability' is used throughout the document, when perhaps 'internal consistency' is a more precise term. Also, assessment, measure, scale and interview are used interchangeably, when interview schedule and interview rating scale might be more precise terms.

Methods:

Development of the GPCCI:

It would be very helpful if a table were provided containing the (summary) clinical guidelines for the three disease states and the items on the GPCCI that address each of the guidelines. This would also help clarify how the clinical guidelines differed across the three conditions and how this was reflected in the GPCCI.

Were GPCCI items reviewed by colleagues involved in producing the clinical guidelines or piloted with a group of potential respondents? If so this should be stated. Were the final items assessed for clarity and comprehension?

What sorts of probing and clarification strategies were proposed for use within the Interviews?

The qualifications/experience and training of the five raters should be described.

Pp 13, para 1, sentence 3: "(diabetes r = 055..." should read "(diabetes r = 0.55..."

Discretionary Revisions (which the author can choose to ignore)

It would be helpful if some discussion were devoted to the content validity of the GPCCI to general practice in other countries. It is likely that the standards of care are similar across developed countries and addressing this point would make the article more interesting to an international audience.

Directions for research should include a discussion of plans for various forms of validation of the Interview Schedule over time.

(notes...) Pp 10, para 3, the inter-rater reliabilities between any two raters was moderate at best, using multiple raters to increase the reliability of a particular rating only improves the estimate (in this case the estimate of the validation criterion) if the larger number of raters are used to review all the files. In this way a highly reliable summary rating could be derived by aggregating across all raters. This was not done and the estimation of the inter-rater reliability using Spearman-Brown adjustments to present a case for the reliability of the criterion measure is not a particularly convincing.

What next?: Accept after minor essential revisions

Level of interest: An article of importance in its field

Quality of written English: Acceptable

Statistical review: No

Declaration of competing interests:

I declare that I have no competing interests.

Mark J. Atkinson, PhD