Reviewer's report

Title: Profiling Quality of Care: Is there a role for peer review?

Version: 1 Date: 27 February 2004

Reviewer: Vincent Mor

Reviewer's report:

General
This is a very well done paper addressing a critical issue in the whole field of health care quality assessment. The paper addresses fundamental issues of the sources of measurement error when the data are based upon an “implicit” review framework in which “experts” judge the quality of health care provided a patient in order to determine whether care by that hospital or physician is of high or even adequate quality. The authors present the results of a carefully designed, multi-site inter-rater reliability trial which was done as a part of a larger study designed to assess care quality provided to patients with 4 different conditions longitudinally. In all instances a set of “implicit” review criteria were established and the physician raters were trained in how to interpret and code the medical condition(s) of each patient.

The manuscript introduces a relatively novel application of hierarchical modeling to deal with how variance in ratings can be partitioned into that associated with the patient, the rater and “noise”. This approach takes advantage of the available data than merely reporting “Kappa” statistics. However, the intricate aspects of interpreting the results, although relatively simply and elegantly expressed, does, for this reviewer, raise the question of whether this is a “methods” paper about an important substantive topic or a substantive paper using methods to make a point. Currently, it is written much more as the latter but it raises questions that beg to be addressed with a more comprehensive explanation about reliability of measures.

The questions and suggestions for revisions posed below are dependent upon the manner in which the authors chose to address this fundamental question about the nature of the paper. The current paper would be quite adequate and would make an important contribution as it stands with some necessary clarifications, but might be far more complete were there more emphasis on the implications of the methods and not just the implications for the burgeoning substantive field of provider profiling. Indeed, since the latter is somewhat dependent upon the former, particularly for implicit review, such an expanded paper would be complementary.

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

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

Needed Revisions:

Most of the suggested revisions are minor and should help clarify the points being made.

In the Design section, it would be helpful to add a clarifying sentence about the number of conditions per patient and therefore per rater. I had to go to the table footnotes to figures our where the “n” of 400+ ratings came from.
The use of the Hierarchical Bayes approach to deal with the modeling and cross-classification of effects nonetheless assumes that the rating data are “normally” distributed since variance partition is structured along the lines of a MANOVA. It would be helpful to know that the data are in fact distributed thusly. The data in Figure 1 appear to justify this, but a sentence indicating as such in the methods is warranted.

In the Appendix, the authors note that they end up using equation 2 for the analyses. Many readers might be more familiar with weighted Kappa rather than the ICC; an explanation of the relationship between the two in the context of this study would be helpful.

It is not indicated whether the data in Figure 1 is based upon the sum of all raters or not (multiple observations per resident).

Figure 2 presents a graphical presentation of the variance levels in Table 2. It is not clear that Figure 2 add anything. Since the overall variance of each outcome variable differs across conditions, having a proportionate bar chart can actually be misleading. If anything the Figure may be best based upon the proportion of providers.

On page 7 the authors note that findings around one condition may be more “clinically interesting” than another. The reason for this use of words is unclear in this context.

---

Discretionary Revisions (which the author can choose to ignore)

Potential revisions.

The authors work on reliability in relation to implicit review is very important because implicit review is based upon judgment. A key finding which is correctly highlighted is that in the face of more consensus about what is right and wrong, raters have higher rates of agreement about what is right and wrong. This is proper because the ambiguity that clouds judgment evaporates in the face of increasing consensus, particularly if raters are trained to all have the most current knowledge. However, this relates to reliability theory in a number of important ways. First, as the authors point out, if the pattern of care begins to conform increasingly to that shared view of quality, then there will be less and less variance in the quality measured, effectively reducing the reliability. While this reviewer appreciates this apparent anomaly, it is unlikely that many do. Additional explanation on this point is particularly important. Since true variance in quality may systematically differ between providers, it is theoretically possible to observe lower reliability in higher quality providers using this kind of implicit review rating of quality. It may be worth mentioning this in the short discussion of use of provider profiling based upon implicit review.

Another technical point that has implications for the interpretation of the discussion points is where the authors note that the reliability of an individual rating on hypertension management is .46 (ICC), if 5 raters are used, the measurement instrument would have a reliability of .80, good by most standards. Unless the authors meant that individual patients are actually the replicates and can increase reliability (not really since they are a separate effect), the authors must know that in most real world circumstances, only one rater is used to determine the quality of care in a single case. They mention the notorious unreliability of satisfaction scales in physicians’ practices, but neglect to point out that we are really dealing with different kinds of reliability. Adding more items (questions) drawn from the same domain can greatly improve the reliability of the resulting additive scale. What that means is that were the scale administered twice (without learning) to the same patient, their scale score would be within the reliability measure confidence intervals around the actual scale. In terms of the profiling example discussed, the “reliability” of the scale within a practice is actually more a function of variation in the mix of patients than the ICC of the multi-item scale score. In the real world it is very rare to have multiple raters judge the quality of care based upon each looking at
the same cases, so the pertinent question is what is the “reliability” of the measure in real world applications where only one rater will be judging each case record? In some sense, reliability used in assessing the reliability of a multi-item scale and a panel of raters is the same thing. However, it is rare that a single item in a scale is asked to “stand in” for all the other items, but that is what we always want to happen with raters because it is so costly to have multiple raters review each case.

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:

None