Reviewer's report

Title: A New Implicit Review Instrument for Measuring Quality of Care Delivered in the Emergency Department

Version: 2 Date: 12 April 2007

Reviewer: David Thom

Reviewer's report:

General

This is a useful and generally well written article describing a new measure of quality of care provided in pediatric emergency departments to the sickest group of patients. The measures of inter-rater reliability appear to be appropriate; though with just 2 reviewer’s it is difficult to generalize the results. The measure of predictive validity, medication errors, is not ideal, as medication errors are probably included in the actual measure of quality being validated. In addition, the association between quality of care and medication errors is badly, an somewhat misleadingly, reported.

----------------------------------------------------------------------------------------------------------------------------------------

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

One issue that needs to be addressed is to what extent the validation measure used (medication errors) was actually part of the measure of quality. Presumably, medications were often part of the physician’s initial treatment plan and plan for disposition, two of the 4 principal domains used in assessing the quality of care. Is there any way to know to what extent errors in medication noted or suspected by the 2 physician reviewers contributed to their rating of the quality of care? In the worst case scenario, if the quality score depended mostly on physician medication errors, then validating it using medication errors would not be appropriate. One possibility would be to look at the variation (range and standard deviation) of quality of care among patients without medication errors, compared to all patients. If variation was the same, this would imply that medication errors did not contribute much to the variation in quality. At least this should be addressed as a limitation in the discussion.

In the results section (second to last paragraph) you report that 44.4% of children who received unacceptable care had a medication error compared to 25% of children who received acceptable care. While this difference looks impressive, you did not provide the actual numbers, which would have to be 4 of 9 children (44.4%) compared to 30 of 120 children (25%) with a p-value=0.42 for the difference. You then report the mean and median for the average overall quality score and for the summary scores for children with and without medication errors. The 95% CIs and p-values presumably refer to the medians in both cases (though you do not specify for the overall scores). However, the confidence intervals do not make sense here, since the difference in median overall scores is -1.0, the confidence interval cannot be from -0.9 to 0.000003). Likewise, the difference in median summary scores is -3.5, so the confidence interval cannot be from -3.4 to 0.000005. You should have chosen a single way to assess the association (probably the summary scores makes the most sense) and provide the correct CIs and p-values for the differences in means (using the t-test) and for differences in medians (using the Mann-Whitney-U test).

----------------------------------------------------------------------------------------------------------------------------------------

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

Title:
Should include Pediatric in the title

Abstract:
Please include the mean, median, and score range in the description of the instrument.

Please replace the sentence beginning “To demonstrate the validity of the instrument, the median of the reviewers’ summary score was higher..” to “The validity of the instrument is supported by the finding of a
higher score for children without medication errors compared to those with medication errors of borderline significance (median=29.5 vs. 26.0, p=.069). (Or could report mean scores – see below).

Methods:
I’m not sure you can’t use mean scores, since the mean and medians appear to be very similar, and the t-test is fairly robust with respect to at least moderate skewness in the data. Without looking at the data, I can’t make a judgment, and unfortunately there are no hard and fast rules about this, but since you talk about comparing means in calculating the ICC, it seems to me to make sense to use the mean scores throughout the paper and to report the t-test results for the validation using medication errors. The results using the more conservative Mann-Whitney-U test could be reported secondarily.

Who were the physician reviewers and how were they chosen? Were they study investigators, pediatric emergency care specialists?

It would be helpful to the reader to provide a couple examples of the types of clinical problems/presentations that were triaged to level 3 acuity.

How many physicians were in the study?

Results:
In the methods, you state that “we anticipate that the instrument will be used by at least 2 reviewers.." and "we sought to maximize reliability by averaging item-specific scores from the two reviewers". Yet internal reliability is reported in Table 2 for each reviewer separately. If the measure is intended to be the average of scores from 2 reviewers. what was the reliability for the averaged measure?

I would like to see the range of summary scores with some measure of distribution (mean and standard deviation or median and intraquartile range). Also, you should include a measure of skewness, and the percent of scores that were at the ceiling (35).

I would also like to see the mean scores among the 4 settings, since one of the stated uses for the instrument is to compare quality of care in different settings.

Discussion:
You mention in methods that you anticipate the instrument will be used by at least two reviewers. Is that your recommendation? Or is the reliability of the relative e rankings high enough that there is not much to be gained from using a second reviewer?

With only 2 reviewers, it is difficult to generalize the reliability findings. This should be noted as a limitation.

Discretionary Revisions (which the author can choose to ignore)

Figure 1 is not really needed if space is an issue.

You might consider giving a name to your scale, to make it easier for subsequent investigators to refer to (e.g., the Pediatric Emergency Department Quality Assessment Scale).

While outside the scope of the current study, it may be worth mentioning in the discussion that a future study could validate the measure against the diagnosis made by the ED physician and the diagnosis made upon review after hospitalization or observation.

What next?: Unable to decide on acceptance or rejection until the authors have responded to the major compulsory revisions

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.

**Declaration of competing interests:**

I declare I have no competing interests