Author's response to reviews

Title: Examining intra-rater and inter-rater response agreement: A medical chart abstraction study of a community-based asthma care program

Authors:

Teresa To (teresa.to@sickkids.ca)
Eileen Estrabillo (eileen.estrabillo@utoronto.ca)
Chengning Wang (chengning.wang@sickkids.ca)
Lisa Cicutto (lisa.cicutto@utoronto.ca)

Version: 2 Date: 27 March 2008

Author's response to reviews: see over
March 27, 2008

Dr. Maria Kowalczuk
Senior Assistant Editor, BMC-Medical Research Methodology
editorial@biomedcentral.com

RE: Manuscript ID MS 5336978461620330: Examining Intra-Rater and Inter-Rater Response Agreement: A Medical Chart Abstraction Study of a Community-Based Asthma Care Program

Dear Dr Kowalczuk,

Thank you for reviewing our manuscript entitled “Examining Intra-Rater and Inter-Rater Response Agreement: A Medical Chart Abstraction Study of a Community-Based Asthma Care Program”. We were very interested to read the comments of the reviewers, who made a number of excellent points and suggestions. We have addressed and incorporated all of the reviewer’s comments in this revised version and would like to resubmit it for consideration for publication in BMC-Medical Research Methodology. The following are a listing of our responses to the reviewer’s comments.

**Comments from Review 1: Timonthy Lash**

**Major Compulsory Revisions:**

1) *Under definition of chart abstraction items, the authors defer analysis of medication information such as strength, dose, and duration. This information, however, is usually crucial to an analysis of pharmaceutical effects. They are difficult to abstract accurately, which is exactly the reason that the uncertainty in such abstractions ought to be quantified. If this information is available, then I would ask the authors to reconsider their decision to exclude presentation of these agreement results.*

   We acknowledge the importance of abstracting medication information accurately. During the initial abstraction (Time 1), the data collection form and database were set as text fields with no specification for abstractors to indicate the strength, dose and duration of any medications that appeared in the patient charts. Thus the information was not provided consistently. However, reference to this issue has been added to the discussion section (see page 13, paragraph 2), and we will consider this topic separately pending further coding and analysis of our current dataset.

2) *Under the inter-rater reliability section, the charts were created synthetically, so the information included in the chart is known with perfect accuracy. Thus, there is a true gold standard, which allows calculation of sensitivity and specificity rather than agreement.*

   We agreed and per the reviewer’s suggestion, in addition to presenting % agreement and kappas, we calculated sensitivity and specificity with corresponding 95% CIs. Comparing the assessments by all raters against the gold standard, the results showed high levels of sensitivity...
and specificity across all the simulated charts and across all the categories studied. The results were added to Tables 4 and 5.

3) The section on data quality and control ought to appear earlier in the methods, under data collection.

As suggested, the section now appears earlier in the methods section (see page 7, paragraph 2).

4) I strongly recommend against using the categorization of kappa, for the very reasons cited by authors in the discussion. There are many shortcomings to this measure of agreement, which have been well reviewed and ought to be discussed at greater length by the authors. One important shortcoming is that it is impossible to use kappa to adjust for categorical misclassification in other studies, where this study's agreement data might be useful for sensitivity analysis.

We agreed with the reviewer that the kappa statistic have shortcomings and we discussed this extensively in our discussion section. In the presence of a reference or a gold standard, test statistics such as sensitivity, specificity, predictive values, and likelihood ratios are more often used than the simple kappa statistic. As indicated in our response above that in this study, since we had a gold standard, we were able to compute sensitivity and specificity estimates across all the categories and for all the charts created. The overall sensitivity and specificity measures in this study were in the order of 90% indicating good validity. While we recognized the limitations of the kappa statistic, it provides a simple measure to assess intra-rater reliability (precision) and also inter-rater reliability in studies that involve multiple raters. While there are other methods of assessing interobserver agreement, kappa remains by far the most commonly reported measure in the medical literature. We expanded our discussion section to include this view (see page 13, paragraph 3).

5) Please include in the methods a statement about whether abstractors knew that a record was being re-abstracted for the validity study. Given the proximity in time between initial and re-abstraction, it seems certain they did. What are the implications (for example, might they have worked harder on re-abstractions or recalled their initial recall results, thereby artificially inflating agreement?).

We added in the methods section more details regarding the process (see page 6, paragraph 2). In brief, the initial abstraction was conducted during the study was in active patient recruitment and follow-up assessments (between September 2003 and June 2005). The re-abstraction study was designed and implemented between July 2005 and February 2006. During the initial abstraction phase, the abstractors were unaware of the subsequent re-abstraction study. To further minimize potential bias, the patients/charts included in the re-abstraction study were randomly selected (110 charts out of a total of 1,433). Neither the investigators nor the abstractors knew or could predict which charts would be included in the subsequent re-abstraction study.
Minor essential revisions:
1) The sample for reabstraction is in one place called a convenience sample and in another place called a random sample. Please explain which is the case.

The re-abstracted charts were selected randomly. Use of the term “convenience sample” was intended to describe the way in which the sample size was determined. In the revised manuscript, the term “convenience” has been omitted. For justification of the sample size of 110 for this study, readers are directed to the article by Sim and Wright (cited within the text).

2) The presentation of the results were difficult to plow through in the text. Please review and consider whether they could be presented more clearly.

The results section was revised to present results more succinctly.

Comments from Reviewer 2: Danielle van der Windt
Minor essential revisions:
1) Methods: I understand from the text (page 5, and later, in the discussion) that time between 1st and 2nd assessment varied. This could be indicated more clearly.

More details were added to the methods section to clarify the varying times for the abstraction.

2. Methods: Summary kappa scores were calculated, for example for each category. How were these pooled kappa’s computed? Simply as a mean kappa? Or were summary scores calculated per category (summation of scores on variables per category) with kappa subsequently computed over these summary scores? Or was another method used to pool kappa’s?

The overall kappa score was calculated by summation of scores within each category, and subsequently using these category-specific kappa values to compute an overall kappa summary statistic. Clarification regarding the details of this calculation has been added to the methods section (see page 9, last paragraph).

3. Results: For intra-rater agreement, kappa’s per category ranged between 0.44-0.90 (page 9), which is summarized by the authors as substantial to ‘excellent agreement’. Shouldn’t this be ‘moderate to excellent’?

The results have been corrected to reflect the appropriate description of the resulting kappa values. Some text has also been removed from the results section so that it can read more succinctly.

4. Results: The heterogeneity in kappa’s is emphasized, which is good. The most relevant results of the paper possibly concern the categories which show either very good, or relatively poor agreement, e.g. asthma education. This is addressed in the discussion paragraph, but it would be good to also mention this in the results.
Modifications to the results section were made in light of this comment (see page 10, paragraph 2, and page 11, paragraph 1).

5. Discussion: Agreement was certainly good, although there will still be misclassification during data collection for the intervention study. The discussion section might briefly address the implications of the results for the trial.

The discussion section has been modified to reflect this suggestion (see page 14, paragraph 3).

6. Discussion: In this study, chart review was only used to obtain information on the process of care, i.e. whether aspects of care had been documented or not. Agreement was quite good, but this information is relatively straightforward. In general, chart review may more often be used to extract more detailed information, such as results of diagnostic tests, decisions regarding diagnostic procedures and content of treatment (type of treatment, frequency, dose or duration), etc. The authors could address this issue in the discussion section, emphasizing that agreement may be less good when more detailed or complex information needs to be abstracted, and that studies on agreement are all the more important in this area.

We acknowledge the importance of abstracting clinically pertinent information accurately. Reference to this issue has been added to the discussion section (see page 13, paragraph 2) and we will consider this topic separately pending further coding and analysis of our current dataset.

We hope that the responses to the reviewers’ comments and resulting revisions to the manuscript will be to the satisfaction of the Editors and the reviewers. Please do not hesitate to contact me should you have any questions related to this manuscript.

On behalf of the co-authors I would like to thank you for your consideration.

Sincerely yours,

Teresa To, BSc, MSc, PhD
Senior Scientist and Head, Child Health Evaluative Sciences, Hospital for Sick Children
Associate Professor, Department of Public Health Sciences, U of Toronto
Investigator, CIHR and MOHLTC
Telephone: (416)813-8498
Fax: (416)813-5979
Email: teresa.to@sickkids.ca

555 University Avenue
Toronto, Ontario
M5G 1X8
www.sickkids.ca