Author's response to reviews

Title: Adjustment of Nursing Home Quality Indicators

Authors:

Richard N Jones (jones@hrca.harvard.edu)
John P Hirdes (hirdes@healthy.uwaterloo.ca)
Jeffrey W Poss (jwposs@healthy.uwaterloo.ca)
Maureen Kelly (MKelly@cihi.ca)
Katherine Berg (Katherine.Berg@utoronto.ca)
Brant E Fries (btries@umich.edu)
John N Morris (jnm@hrca.harvard.edu)

Version: 2 Date: 15 July 2009

Author's response to reviews: see over
Dear Editors:

My co-authors and I are very grateful for the opportunity to revise and resubmit this manuscript. The reviews were very helpful, and we hope that our revisions have made this a stronger manuscript.

In this note, we address each of the reviewer's concerns and discuss how we have responded.

Referee #1.

Major Compulsory Revisions

1. The data presented are drawn from previous CMS funded contracts with Abt Associates. A general reader might not be familiar with these efforts and the various technical reports, some additional effort is needed to clarify the rationale for this ms and distinguish the analyses presented here from the Abt report. For example, the passage at the top of p. 8 referencing the methodology for the QI scores refers to prior work (regression coefficients) not presented here, so the reader cannot assess this important aspect of the risk adjustment.

We have added detail on the previous contracts that generated the data used in this work, and more detail on the generation of the regression coefficients used in the adjustment method. We have clarified that our new adjustment methodology does not utilize the regression coefficients developed in previous contract work. The new text...
It is important to note that although we use data collected under previous contract work with CMS to develop our new quality indicators, the results and methods we describe here are completely original. The regression coefficients and stratification weights used in our adjustment process are not part of the previous work. Interested readers may consult the technical reports from the Mega-QI study for additional details on the sample recruitment and characteristics, and the rationale and development procedure for second-generation quality indicators. Technical details on our adjustment procedure, including computer syntax with regression and stratification weights, are available upon request.

2. The authors do not present data on the impact of the new risk adjustment approach on the number of facilities that might be flagged for having poor quality (c.f. Berlowitz et al.). Please address this point, either with new analysis or justification.

We do not present data on the impact of the new risk adjustment approach on the number of facilities that might be flagged as having poor quality. We explain why in the text. The new text reads (page 21):

A limitation of our manuscript is that we do not present information comparing the relative proportion of facilities flagged as having poor quality under second and third generation adjustment procedures. Such an effort requires establishing a benchmark or threshold against which a facility may be judged as having poor quality of care. We are unaware of thresholds having been described for the first and second generation QIs, and they have not been described for third generation QIs. In the absence of clinically based thresholds, empirically based thresholds are typically used. For example basing judgments of poor quality on the basis of the mean, median or percentile of the adjusted QIs score. Use of empirically based thresholds would not be informative for comparing second and third generation QIs.

3. Table 2 is extremely long, and Tables 3, 4 and 5 summarize aspects of the adjustment in a way that obfuscates the approach. I strongly recommend combining the information on standardization and stratification into columns on Table 2, so the reader can see which approach goes with each QI. This will be a large table that could be turned into an appendix. One indicator from each domain should be retained in the main text, along with the total number of QIs per domain.

We have replaced tables 2-5 with a single appendix (Appendix A) describing each QI in detail.
4. Figure 3 should be reduced to a small (e.g., one per domain) number of exemplary graphs. The full can be presented as an appendix.

We have omitted Figure 3, and include all of the figures as an Appendix (Appendix C).

5. The discussion is brief given the length of the paper and amount of analysis presented. What are the implications of this new system for policy or practice? Do these results differ from what is reported in NHCompare (see #2)? Can they make the case that this new approach is sufficiently transparent to gain acceptance by the industry? Is the new method computationally intensive?

Our view is the fundamental issue regarding all quality indicators is that of validity, as we stated in our conclusion. Our assessment of validity in this analysis, and that which is common to first and second generation nursing home quality indicators, is very limited. Recent work (e.g., [1]) suggests nursing home QIs are influenced by extraneous and contextual influences. More work on this is certainly required. However, our goals in this manuscript do not include issuing a policy recommendation regarding whether or not nursing home QIs should be used for policy purposes (e.g., pay for performance). Our current research (not summarized in this manuscript) is attempting to demonstrate whether targeted clinical best practice protocols and continuous quality improvement activities can boost QI scores in nursing homes using a randomized controlled trail design. The results of that work will be extremely informative with regard to the validity of QIs, but is many years away at this point. Related work has been completed and is cited (page 22). The imperatives of policy initiatives follow a timeline not necessarily tied to the timeline of the accumulation of scientific evidence.

As for transparency, we will not attempt to make the case that the adjustment method is transparent enough for acceptance by the industry. This is a qualitative question that we have not investigated. Moreover, not all of our author group accept the premise that transparency is required for acceptance by the industry, citing numerous examples in daily life of complex numerical composites that are computed in an opaque fashion but have broad acceptance by industry and the public (e.g., achievement tests, personal credit ratings, market indices). What we agree upon is that the industry will accept a set of QIs that are reasonably and appropriately risk-adjusted. Recurrent calls in the scientific literature can be heard for further refinement of risk adjustment (e.g., [2]). Further refinement of risk adjustment models for nursing home QIs is the main goal of this manuscript.

The third generation method is not appreciably more computationally intensive than that of second generation QIs, in that essentially only two additional steps – stratification and then averaging across strata – are added to the adjustment procedure. We have added this detail to the manuscript (page 19).
6. In the conclusion, the comments about appropriate reference facilities are not supported by this manuscript, as the authors did not appear to analyze the impact of different reference groups on the results. The suggestion for peer nominated facilities is interesting, but off topic. Please revise this section after responding to #5 above regarding the Discussion.

We have revised this section, and clarified that our suggested for peer nominated facilities is future work and not work supported by this manuscript, and that it specifically addresses the implementation of QIs for policy purposes.

Minor Essential Revisions:

1-2. P. 4; 1st full para, 2nd sentence “First, differences in types of residents…”, is unclear what is meant by care sectors. Same para, “NQ” seems to be out of place.

We have removed the phrase “care sectors” and the misplaced “NQ”

3. The phrase, “must be carefully considered” does not add to the general critique of deriving QIs from resident assessment data. The authors should be more precise about the issues they are raising.

We have added an example that we hope clarifies our meaning.

4. The sentence beginning “Third, many relevant NH outcome measures are rare…” is an important point, but it is not clear that the new methods will alleviate this problem. The words ‘prevalent’ and ‘rate’ should be deleted as they do not add to the thought.

The reviewer is correct, and we have added as a limitation section in our discussion that the third generation adjustment method does not address the rare event problem (page 20). The suggested rewording to the introduction passage has been incorporated.

5. 3rd para, sentence that begins “Disadvantages include…” should be revised to drop the parenthetical, which is overly technical. Change “coarse adjustment” to “coarseness of the adjustment and creation of strata with small denominators….?” Delete the word ‘indicator’.

Rather than drop the parenthetical, we have expanded and explained. We feel the point about residual confounding is important enough to retain. Other suggested rewording has been adopted.

6. P. 5, 1st full para, 2nd sentence, end after the word ‘facilities’ and start a new sentence with “Expected rates...”
The suggested rewording to the introduction passage has been incorporated.

7. P. 5; 3nd full para, “(variability in valence...)” This phrase is dense and technical. Please expand and explain

The sentence has been removed.

8-12. P. 7; 2nd sentence, “QI” should be plural. P. 8; 2nd full para, 2nd sentence, fix typo in word ‘using’: “improved through using”. P. 10; 2nd para, sentence that begins “Responsive strategies...” there is an extra word “a” to delete. P. 15; the word “indict” seems a bit strong. Nevertheless, please change “is” to “as” to read “indict a QI as invalid.” P. 17, last line on the page is a fragment.

We have made the indicated changes and are very grateful to the reviewer for the careful reading.
Discretionary Revisions

1. On page 7 first paragraph second line, the authors say that these 79 QIs were chronic care measures. Can you briefly discuss why the post-acute care QIs were not included in this study?

We have developed new chronic care QIs. We are working on developing new post-acute care (home care, and acute care) quality indicators, but are not ready to report on these.

2. It would be very informative to show the reliability and validity of newly developed (or enhanced) risk adjustments for current CMS QMs.

We have added reference to and included two new appendices, one detailing the numerical results for the reliability (Appendix D) and validity (Appendix E) results.

3-4. The last sentence on page 17 needs a period at the end. Page 7 Methods section, first paragraph, fourth line, take out “June” between “second quarters of” and “2002”.

The corrections have been made, and we are grateful to the reviewer for drawing these errors to our attention.
Referee #3

Major Compulsory Revisions

1. An overall goal of the study is not clearly stated. Also clear and explicit research questions are missing.

We have added to the introduction of the overall study goal and research questions to the (page 5-6).

2. Please justify the need for and the practical value of a new list of adjusted quality indicators.

We have added a justification to the introduction (page 6).

3-4. There is no (well-grounded) definition of “quality of care in Nursing homes”. The list of QIs could also be used for describing the health of the residents (which is not necessarily a consequence of nursing interventions). The list of 79 QIs mainly obtains resident’s conditions (often unclear if diagnose or outcome), also changes in conditions (which could be interpret as outcome, but without linkages to nursing interventions) as well as obtained aids like “new indwelling catheter” or “feeding tube” (with no information if this is appropriate/indicated). So it is not comprehensible how these conditions reflect care quality. (e.g. a high rate of residents who are bladder or bowel incontinent?). Is there any evidence that all these indicators are preventable or capable of being influenced by nursing activities?

The reviewer as astutely identified the most salient issue in nursing home quality research and measurement. It is true that first, second and our new third generation quality indicators are ultimately measures of aggregate clinical state or course of a facility’s residents (with few exceptions as noted by the reviewer). As such, nursing home QIs reflect only one leg of the Donabedian structure-process-outcome model for quality of medical care [3]. Whether and how nursing home QIs are risk adjusted will not remedy this. A potential result of the exclusive focus on resident outcomes, taken in the context of the unknown validity of nursing home QIs for use in decision making by consumers and policy makers, can lead to nihilism with respect to the prospects of developing and using measures of nursing home quality. Perhaps no better expression of the frustration of this state of affairs can be found in Charles Phillips et al.’s review, Where should Momma go? (BMC Health Services Research. 2007;7(1):93. [4]).

To address this point in the manuscript, we have added reference to published studies indicating that interventions can be put in place that can help facilities improve their standing on QIs (page 22). In addition, we have added a paragraph to the limitations section that reads:
Beyond the issue of validity, there are limitations in the conceptualization and operationalization of nursing home quality indicators that require additional research. With few exceptions, the first, second and our new third generation quality indicators are ultimately measures of aggregate clinical state or course of a facility’s residents. As such, nursing home QIs reflect only one leg of the Donabedian structure-process-outcome model for quality of medical care [3]. Whether and how nursing home QIs are risk adjusted will not remedy this. A potential consequence of this framework is nihilism with respect to the prospects of developing and using measures of nursing home quality. Perhaps no better expression of the frustration of this state of affairs can be found in Charles Phillips et al.’s review, Where should Momma go? [4].

5. The reader gets no information about the risk adjustment for each QI. Please describe more detailed the direction and evidence for the underlying associations between the covariates and the QIs. (Some covariates are also QIs.)

We have replaced tables 2-5 with a single appendix (Appendix A) describing each QI in detail. This appendix does not include parameter estimates for individual covariates used in the adjustment process. Such data are available upon request (and this is mentioned in the manuscript), embedded in the QI scoring algorithms. This is not an attempt to be obfuscatory, and we note that CMS’s regression weights used in second generation QIs have also never been published. We feel that what is more important is how QIs variously adjusted perform with respect to measures of reliability and validity.

6. Why is “resident is comatose” an exclusion criteria to judge the care quality e.g. preventing pressure ulcers? Please explain how the list of exclusion criteria was developed.

Resident is comatose is an exclusion criteria for many QIs, but not for pressure ulcers. The rationale is that there is a skip pattern in the MDS for comatose residents. MDS assessment data will be appropriately skipped for residents who are comatose. This information is now detailed in Appendix A.

7. Please describe the data quality. How complete were the data of the medical charts? How correct were these data? Who made the documentation (qualification? Competence in using scales? How long/ how well do they know the residents?) How was the data quality checked? If scales/ instruments (e.g. pain scale, RUG-III, PSI, depression rating scale, etc.) were used: how were their psychometric properties (references)? How were the data of the medical charts collected (observation, asking the resident/ relative, measurement, etc.)

We have added additional detail on the data collection and data quality for the field
Field Data Collection

As part of the Mega QI study, data collection in each facility included independent data describing the process of care at facility and individual resident level. Details on the validation field study are presented elsewhere [5]. We summarize aspects of the field study here. Data collection (2001-2002). Teams of trained nurse researchers visited sampled facilities, and completed Medical Record Review of the charts of residents representing the 30 most recently completed MDS assessments. Nurse researchers also completed a partial MDS 2.0 assessment on each sampled resident, and Environmental Walk Through and Resident Observation survey. In addition, an Administrative Questionnaire was delivered to administrators and/or directors of nursing. The purpose of the Medical Record Review (MRR) was to obtain information regarding the care processes and types of patient/resident assessments performed by sampled facilities on select areas. Twenty-one care areas (or quality dimensions) were reviewed during the MRR (cognitive impairment, communication, delirium, depression/mood, behavior problems, ADL improvement, ADL decline, mobility/walking, falls, anti-psychotic drugs, pain, physical restraints, feeding tubes, undernutrition / low BMI / weight loss, indwelling urinary catheter, bladder incontinence, bowel incontinence, infections, pressure sores / potential for skin breakdown, burns, abrasions, skin tears, and little or no involvement in activities). For each of these domains, nurse assessors reviewed the medical record (including nursing progress notes, physician orders and progress notes, care plans, therapy consults and notes, medication administration records, flow sheets and other interdisciplinary notes and consults) for resident care and status documentation. Assessors looked for documentation on comprehensive assessments, problems/issues, change in status (within certain time frames), referrals, treatments and nursing care plans. A MDS Supplement was used to conduct assessments on all patients in the sample, including assessment areas from the MDS in selected areas (cognitive patterns, communication/hearing patterns, mood and behavior patterns, physical functioning and structural problems, continence, disease diagnoses, health conditions, oral/nutritional status, skin conditions, activity pursuit patterns, medications, special treatment procedures, and discharge potential and overall status). The Administrative Questionnaire included questions regarding staff responsibilities, staff/resident/family involvement in care, resident status, access to specialists/consultants, clinical communication channels, staff turnover, staffing ratios, planning processes, information on the organization; and training and orientation of staff. The Environmental Walk Through/ Resident Observation was used to gain an overall understanding regarding whether the facility is “resident-centered”, what the “feel” of the facility, and what the nature of staff-resident interactions. A series of general environmental measures were employed to describe the responsiveness of the milieu to resident strengths, needs, and
problems that include general care environment measures (e.g., nature of physical environment, communication strategies, environmental manipulation and resident interactions with staff). These measures were collected through assessment, surveillance, and observation of staff technique. The nurse researchers recorded observations three times per day.

Nurse Researcher Qualifications, Training and Reliability
Peer Review Organizations (PROs) in participating states were contracted to hire field data collectors, with priority for registered nurses (ultimately only 1 was not an RN) with chart review experience and experience in a long-term care setting and/or in completing the MDS Version 2.0. Nurse researchers attended a five-day training and certification program led by the Mega-QI CMS Project Officer, Steering Committee members (including two RNs), five experienced RN researchers with experience in similar data collection activities. A training manual was developed and each assessor was provided a copy. Half of the training program was devoted to training in how to conduct resident assessments using a subset of items from MDS Version 2.0. To certify competency, each trainee completed a case and met individually with the lead trainer for review. To enhance and maintain consistency in coding, project staff held weekly one-hour conference calls with the assessors during the course of data collection. Minutes of the calls. Reliability among nurse researchers was assessed by having nurses complete two paired assessments and medical reviews with their partner per facility. Agreement statistics for the MDS inter-rater reliability of nurse researchers were very good (average kappa coefficient 0.78)[5].

Data Quality
Data quality was ensured by using computerized assisted interviewing by trained nurse assessors, frequent teleconferences among the research nurses and project staff, and by fax-back and call-back of facility administrators. For the variables used in this analysis, data were missing data about for about 8% of data elements. Missing items were handled with mean-based single imputation. Single imputation is known to produce biased standard errors, but our inferences are based on relative differences in means and covariances among a set of QIs relative to validation scales between second and third generation QI adjustment procedures, and any bias introduced through this handling of missing data will affect both adjustment procedures equally.

8. Assessment of reliability of the new QIs: stability of the condition between the measurements is a requirement before calculating the reliability. How was this controlled? What about the reliability of the other QIs?

We agree that stability of the condition in large part determines the magnitude of the reliability of a QI. However, in this manuscript, what is important is the relative reliability of QIs scored under a second and third generation risk adjustment framework. We now
detail the reliability of each QI scored in second and third generation framework in Appendix D.

9. Assessment of validity: Please describe the validation elements and the expert clinical panel review.

The included validation elements are described in an on-line appendix to the Mega-QI final report (Appendix F). We include a pointer in the text to that document (page 14):

The included validation elements are described in an appendix to the Mega-QI final report (Mega QI Report Appendix F, accessible at http://tinyurl.com/MegaQIValidationScales-PD).

10. The multiple correlation coefficient determine the level of validity. Please name references for building the three levels. Why are there different coefficients depending on the model?

The rationale for the multiple coefficient method for determining level of validity was developed and described in the Mega-QI study, which has been cited. Our goal is not to establish that procedure as an appropriate method for validating QIs, but rather to show how conclusions regarding QI validity would differ depending upon whether a second or third generation QI risk adjustment framework is adopted (page 15). The rationale for different cut-points depending on the model is that the coefficient of determination will be greater when there are more predictors (i.e., preventive plus responsive elements).

-Minor Essential Revisions

1. Please figure out the single strata of the 5 stratification factors. How were these strata built/ tested? (content and statistic!)

The thresholds for the five stratification factors were defined at the 20\textsuperscript{th} and 80\textsuperscript{th} percentiles of a convenience cross-national sample of nursing homes. We now alert the reader that details on the threshold values can be obtained by request as part of the technical QI specifications (page 9).

2. Figure 3: No labeling/ marking of x-nor y-axis.

We choose not to label and mark the x and y axis to maximally condense the image. They are intended to be as descriptive as sparklines [6], which are small, high resolution graphics. In our revised manuscript the plots in Figure 3 (now Appendix C) are recast as sparklines. Our goal was to provide a broad visceral impression of differences in the distribution of QIs in different jurisdictions, rather than precise numerical estimates. Axes are defined in the figure caption.
3. Some references seems quite old.

We have reviewed our citations and only included relevant references. We have added some more recent references in our revised text in new sections. Text has been moved from new page 12 to new page 19.

4. May be you can make a table that compares the characteristics of the validation sample with the reference sample.

The reference sample and the validation sample are used for very different purposes. It is very important that the validation sample be broadly representative of nursing facilities. In the absence of representativeness of the validation sample, validity inferences are meaningless. For the main goal of this manuscript (comparing second and third generation QI adjustment procedures) representativeness is merely a convenience as the same reference sample is used in the adjustment process for both frameworks. Regardless, we believe both samples are quite representative, but a comparison would add unnecessary detail to an already cumbersome manuscript.

5. Some phrases in the results section could better be placed in the discussion section. (pages 14-15)

We have moved some phrases from the results to the discussion.

6. Please explain “..the absence of validity evidence in this study does not indict a QI is invalid.” (page 15)

This phrase has been moved from the results to the discussion, where it is explained (page 19-20).

7. Please name clearer the limits of the study (see mayor comments 1-10) in the discussion section.

We have added three paragraphs on limitations (page 20-21):

> Finally, it is important to note that some limitations of QI scoring in first and second generation frameworks remain in our proposed third generation framework. Particularly, the third generation framework does not directly address the problem of low base rates (rare events). More work is needed to address this issue. One approach would be to accumulate observations over a long period of time. Typically nursing home QIs are scored quarterly, but longer periods may allow the accumulation of more events and lead to more stable estimates. The cost of such an approach is the greater time lag may not capture current quality or care practices at a given facility. A better approach may be to limit consideration of viable QIs to those that do not have a low base rate.
Beyond the issue of validity, there are limitations in the conceptualization and operationalization of nursing home quality indicators that require additional research. With few exceptions, the first, second and our new third generation quality indicators are ultimately measures of aggregate clinical state or course of a facility’s residents. As such, nursing home QIs reflect only one leg of the Donabedian structure-process-outcome model for quality of medical care [3]. Whether and how nursing home QIs are risk adjusted will not remedy this. A potential consequence of this framework is nihilism with respect to the prospects of developing and using measures of nursing home quality. Perhaps no better expression of the frustration of this state of affairs can be found in Charles Phillips et al.’s review, Where should Momma go? [4].

A limitation of our manuscript is that we do not present information comparing the relative proportion of facilities flagged as having poor quality under second and third generation adjustment procedures. Such an effort requires establishing a benchmark or threshold against which a facility may be judged as having poor quality of care. We are unaware of thresholds having been described for the first and second generation QIs, and they have not been described for third generation QIs. In the absence of clinically based thresholds, empirically based thresholds are typically used. For example basing judgments of poor quality on the basis of the mean, median or percentile of the adjusted QIs score. Use of empirically based thresholds would not be informative for comparing second and third generation QIs.

Please do not hesitate to contact me if I can in any way facilitate your role as editor (jones@hrca.harvard.edu, 617-363-8493).

Sincerely,

Richard N. Jones, Sc.D.
References cited in this letter


