

Author's response to reviews

Title: Using a summary measure for multiple quality indicators in primary care: the Summary Quality Index (SQUID)

Authors:

Paul J Nietert (nieterpi@musc.edu)
Andrea M Wessell (wessell@musc.edu)
Ruth G Jenkins (jenkinrg@musc.edu)
Chris Feifer (feifer@usc.edu)
Lynne S Nemeth (nemethl@musc.edu)
Steven M Ornstein (ornstesm@musc.edu)

Version: 2 **Date:** 5 February 2007

Author's response to reviews: see over

February 5, 2007

Implementation Science
BioMed Central Ltd
Middlesex House
34-42 Cleveland Street
London W1T 4LB, UK

Dear Editors,

I am pleased to resubmit the manuscript entitled “Using a summary measure for multiple quality indicators in primary care: the Summary Quality Index (SQUID)” for publication as a research article in *Implementation Science*. We appreciated the reviewers’ comments (*in italics*) and have responded to their concerns (in normal text) below.

Please let me know if I can provide you with any further information.

Sincerely,

Paul J. Nietert, PhD
Medical University of South Carolina
135 Cannon Street, Suite 303
PO Box 250835
Charleston, SC 29425
Phone: 843-867-1204
Fax: 843-876-1126
E-mail: nieterpj@musc.edu

Reviewer: Dr Martin Lee

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

The fundamental concept here is sound. It is reasonable to establish a single pooled quality estimate from multiple quality indicators and to use this to evaluate the performance of a healthcare unit or system. However, the attempt here is incomplete and, in my opinion, overly simplistic, or, at least, the evidence provided is not convincing that this approach is sufficient. We have responded to the comments below and have made substantial changes to the paper. We believe that the paper is more convincing that our approach is sufficient in certain circumstances.

The authors use a basic proportion of indicators met to define their index (and spend far too much space in the paper describing this extremely basic statistical concept).

We have reduced the description of the manner in which the SQUID is calculated while maintaining a clear description of the methodology. We wanted to make sure, however, that the reader understands that whether or not a given patient has met his/her recommended target depends on his/her comorbidity profile. Thus we were reluctant to reduce the algorithm description any further.

As a basic premise for this paper, I find this not properly developed. Much more needs to be done in order to provide support to publish this simplistic concept. For instance, could this index be used in a prospective evaluation of quality improvement?

Although we did report that the SQUID was used in a prospective evaluation of quality improvement (i.e. the ATRIP project), we recognize that we didn't emphasize its utility within QI projects in general. We added some clarification to this point in the background and discussion. Although the SQUID was not introduced in the beginning of ATRIP, participating practices were provided their SQUID score as a part of quarterly performance reports. We also recognize that we were a little unclear about the dates of ATRIP and have now included these.

Why is an unweighted index as good as one that weights according to difficulty of achievement or the severity of the disease involved? (Although the authors do mention this in their discussion, they need to pay much more attention to this as most composite health indices use some sort of weighting in their development.)

We understand this concern and gave serious consideration towards incorporating weights. As we now point out in the discussion, deriving them typically requires either building some type of group consensus or deriving them empirically from statistical methodology such as item response theory. We would argue that any such weighting scheme would result in a loss in the ease of interpretability of the SQUID, a factor we feel is key in communicating with an extremely varied audience that includes both providers (doctors, physician assistants, nurses), office staff, and potentially, according to the physician quoted as part of the qualitative analysis, even patients.

Their attempt to provide validation information is commendable, but again I think this is underdeveloped. One key statistic that is missing is the distribution of E among the patients. Clearly, if E tends to be very small for a reasonable percentage of the subjects, then the possible

values for SQUID are, of course, very limited. Under these circumstances, the responsiveness of this index is going to be quite different than when E is closer to the maximum.

Although E has a distribution that is skewed to the right, the way our indicators are defined, each adult has an E value that is 6 or greater. Thus we don't feel the potential problem suggested by the reviewer is that substantial within the context of this paper. The median of E is 9, and the mean is 10.6 (s.d.=4.9). We have now incorporated these data into the paper and added a figure describing this distribution.

They report the responsiveness in terms of the change in percentage over a 15 month period. The difference seems to be small in absolute terms and is strongly statistically significant only because of the huge number of subjects involved.

Although there were a lot of patients involved, the results were also significant across practices, with significance being measured via paired t-tests among the n=85 practices.

It would be useful to also know what this numerical change is for all subjects and not for just those who were active during the entire period (i.e. overall value for the 3rd quarter, 2004 versus the overall value for the 3rd quarter, 2005) as I would suspect that those who may not be doing as well might not continue to appear in the system for evaluation.

We have now incorporated these findings into the results and Table 3 – although the mean values for the SQUIDs at the 2 time points is lower, the absolute change (3.2%) is comparable.

As an overall evaluation of this paper, I think that conceptually the authors have a interesting idea, but the approach taken here is convincing enough yet. Either they consider a more in depth quantitative evaluation of their index or they use this forum to describe the qualitative issues and difficulties encountered in the actual use and development of such a concept in a large healthcare system.

Based on reviewer comments and discussion with the journal's editor, we have chosen to provide a more detailed quantitative evaluation of the index.

Reviewer: Asch

The problem of aggregating quality indicators is a very important one, especially as indicators proliferate. Individual providers, consumers and even managers have difficulty interpreting multiple simultaneous data on various aspects of their performance. As the authors point out, there is a need for summative measures to screen for problems and guide efforts toward quality improvement. They have developed an aggregate measure of the 30 or so common primary care indicators of quality by simply adding them up and averaging either for an individual patient or a practice.

The description of the simple methods for calculating these scores is unnecessarily long. We have reduced the description of the manner in which the SQUID is calculated while maintaining a clear description of the methodology.

This approach has been used before both by CMS and RAND, as the authors note, as well as by the VA, as they do not.

We now mention the use of the VA indices.

This approach has problems, some of which the authors acknowledge. The first is that not all indicators are created equal with regard to clinical outcomes. It is more important to measure LDL in a diabetic than measure total cholesterol in all other adults. Worse yet, because there are many more adults without diabetes than with, the latter measure will have a much higher weight in the overall practice score. Various approaches have been tried to address this, and the authors could discuss them more. Some research teams have developed weights based on either expert opinion or truncated decision trees predicting health related quality of life. Others have partitioned overall scores into conditions e.g. diabetes, coronary artery disease, and weighted care for each disease equally. The authors might rightly say this was beyond their scope, but they should at least point the reader in that direction, or provide some basic simulation studies. While we agree that much of what the reviewer states is beyond the scope of our paper, we have tried in the discussion section to do a better job discussing these alternatives to our methods.

Another problem is that of interaction between indicators, which the authors fail to address. A third problem is that even if one were to assume equal weights, some indicators are inherently more difficult to pass. There are statistical methods for adjusting for difficulty of passing derived from item response theory. The authors could have addressed both of these last two problems more thoroughly.

We have now addressed these issues more thoroughly in the discussion. As mentioned in response to Dr Lee's review, we argue that the use of a complicated weighting scheme would result in a loss in the ease of interpretability of the SQUID, a factor we feel is key in communicating with an extremely varied audience that includes both providers (doctors, physician assistants, nurses), office staff, and potentially, according to the physician quoted as part of the qualitative analysis, even patients.

The component indicators are simple ones, purposively, so that they might be abstracted from an electronic medical record without a lot of clinical data on contraindications. For implementation purposes, this was a good idea. Still it would be good to know to what extent the unmeasured

contraindications are common or non-randomly distributed, which would make them more worrisome, an analysis that could have been accomplished with a limited chart review.

To maintain patient privacy under our IRB approval, we are unable to do chart reviews other than using the electronic data already provided to us. Because physicians do not record in a standard fashion whether a patient has a contraindication for a given measure, we choose instead to ignore this in our reporting as the reviewer points out. We do educate PPRNet members that we don't expect them to be able to reach 100% on an individual indicator or on the SQUID for this exact reason. We have now tried to clarify this limitation in the discussion. In future work, we may be able to quantify allergies or age or disease-based contraindications from structured fields in upgraded versions of the EMR.

The analyses of the statistical properties have some strengths and weaknesses. The idea of reliability testing by evaluating sequential measures during a quality improvement initiative seems misguided. Ideally, two different measures would have taken place at the same time, though if these are just abstracted from a database, how they could change from measure to measure is unclear. I would suggest removing the reliability section of the manuscript.

We inserted a sentence in the methods discussion the reviewer's point above and removed the reliability section from the results, including Table 4.

Measuring the responsiveness of SQUID to quality improvement is more clearheaded, yet the data here are unconvincing. Only a very small improvement was noted.

Although the absolute change might seem small, the change was consistent over the 3.5 years of the ATRIP project. We do think it is also of importance that 88% of practices participating in a demonstration project exhibited a positive increase in their SQUID during from 10/04 to 10/05, a result that would be highly unlikely given chance alone. We have now mentioned this in the results, along with more thoroughly describing the statistical analysis in the methods.

The internal consistency of SQUID is high, but this reviewer is unclear as to what that means. Do we really expect that simple quality measures should be highly intercorrelated? If they were not, would that mean that we should not aggregate them? There has been a lot of debate about whether there is an underlying construct called quality or whether there subdomains (see Katherine Kahn's work on this), and the authors don't touch upon it. More encouraging is the variation in SQUID scores across practices and patients.

We have now done a more thorough job of describing why our high internal consistency is important. We have also now discussed the debate about whether quality should be measured as 1 or multiple domains and cited Dr. Kahn's relevant work in this area.

The strength of this manuscript is underdeveloped. Especially for this journal, the readers would benefit greatly from an expanded version of the subsection of Methods entitled "Use of SQUID in quality improvement" and the results section beginning at the bottom of page 11. Knowing more about how providers reacted to the idea of summative measures, the problems the authors had in implementing the measures would help others who plan similar efforts. This qualitative information is at the core of the real message of this paper and I would encourage the authors to develop it further.

We appreciate the reviewer's enthusiasm for this aspect of the paper. We have expanded this section and incorporated some additional qualitative feedback in the manuscript. However, after

discussions with one of the journal's editors, we feel it is preferable to keep the paper's main focus a quantitative one.

Reviewer: Robbie Foy

General

This review is mainly written from the perspective of a UK family physician and implementation researcher. Hence my comments will focus on clinical utility and implications for quality improvement evaluation. This manuscript describes an algorithm used to produce a summary measure for the quality of primary care. The algorithm was applied to an electronic medical record system used in 89 ambulatory care practices. The authors assessed and found acceptable the reliability, internal consistency, responsiveness to change and face validity of the summary measure.

This manuscript is well written. The criteria used in the summary measure are clinically relevant and evidence-based and the algorithm methods appear transparent. My feedback therefore mainly concerns discretionary issues.

Major Compulsory Revisions (that the author must respond to before a decision on publication can be reached): *None*

Minor Essential Revisions (such as missing labels on figures, or the wrong use of a term, which the author can be trusted to correct): *None*

Discretionary Revisions (which the author can choose to ignore)

1. A brief description of the active QI components of A-TRIP (e.g. whether it used audit and feedback) in a couple of sentences would be helpful to readers not familiar with this programme. The A-TRIP QI demonstration project was comprised of 3 specific types of interventions: practice performance reports (audit and feedback), optional semi-annual site visits to practices, and optional annual network meetings. This has now been clarified in the background.

2. I wasn't certain if there was a clear rationale for combining process and outcome measures – and whether this is routinely done in other quality measures. The achievement of clinical outcome measures may be partly determined by population characteristics. So, wouldn't process measures used alone represent a more sensitive indicator of performance? Outcome measures themselves are of interest but possibly are measuring something different unless adjusted for confounders. Please clarify in either manuscript text or accompanying correspondence.

The rationale for combining process and outcome measures was simply to derive a measure that incorporated all the quality indicators of interest. Given that the vast majority (31 of 36) of our measures are process measures, the SQUID is driven by the process measures. Combining process and outcomes in one composite score has been done within clinical areas such as diabetes, and we have now referenced articles that have done so.

3. How comprehensively do the summary measures capture the relevant patient populations? In other words, are there any risks of over- or under-playing performance based on patients being excluded deliberately or by default from the denominator? For example, were the estimated prevalences broadly in line with those predicted by literature? I'm not sure if this feature is

mainly determined by the comprehensiveness of the PPRNet electronic medical record system rather than the algorithm itself.

The way we have designed the SQUID, every adult is eligible for at least 6 individual quality measures – thus no adult is excluded from the denominator. Absent a thorough chart review of the text of progress notes (which we couldn't practically or ethically do), it would be hard to know for sure whether the estimated prevalences are over- or under-estimates. Also, because of varying definitions of how an active patient is defined, comparing prevalences to those in other published studies is problematic and beyond the scope of this paper.

4. How were the comments fed back to assess face validity selected – so as to avoid selection bias? Were there any negative results? How many of those asked for comments provided feedback?

The process for ascertaining qualitative feedback was informal, as it was solicited via an e-mail listserv for PPRNet members. There were no negative comments received. Because of the way the listserv works, it's difficult to know how many people read the e-mail requesting feedback. Also, since we were just trying to get an overall sense from the practitioners who were using the SQUID, we didn't press everyone to contribute feedback. We clarified that this was truly an informal request for feedback in the results.

5. One of the limitations of using summarised performance measures in assessing quality of care is that they might help detect general problems, but condition-specific indicators might be more useful to target quality improvement efforts at conditions where need is greatest.

We have added this limitation. We include the SQUID in practice performance reports that include condition-specific indicators that practices may use to identify priority areas.

6. That said, I thought that the most interesting feature of this algorithm is its use in measuring equity, i.e. levels of use by or access to recommended treatments by those in need of them. This issue can be marginalised in quality improvement efforts, where efficiency is usually implicitly more important than equity. In this case, were there any (probably predictable) characteristics of patients who received inappropriate care?

Although this is an interesting question, it is really beyond the scope of this particular manuscript.