

Author's response to reviews

Title: Improving Benchmarking by Using an Explicit Framework for the Development of Composite Indicators: An Example Using Pediatric Quality of Care

Authors:

Jochen Profit (profit@bcm.edu)
Katri Typpo (ktyppo@bcm.edu)
Sylvia J Hysong (hysong@bcm.edu)
LeChauncy D Woodard (lwoodard@bcm.edu)
Michael A Kallen (makallen@mdanderson.org)
Laura A Petersen (laurap@bcm.edu)

Version: 2 **Date:** 24 November 2009

Author's response to reviews:

24/11/2009

Brian Mittman, Ph.D.

Editor-in-Chief

Implementation Science

Re: Manuscript #: 5627251682829895

J Profit, K Typpo, S Hysong, L Woodard, M Kallen, L Petersen.

Improving Benchmarking by Using an Explicit Framework for the Development of Composite Indicators of Pediatric Care Quality

Dear Dr. Mittman,

We are very pleased with and appreciate the thoughtful suggestions by the reviewers. Below is a brief summary of the reviewers' comments along with our responses.

Below, is a point-by-point discussion of the changes made to the manuscript.

Reviewer 1

Major compulsory revisions

1) The authors need to acknowledge that the audience for this journal is international, and amend some of their introductory comments in this light. What about the experience outside the US?

Response: The background section has been altered to reflect the reviewer's comment:

"In pediatrics, interest in provider health care performance is rising. Various

countries, such as the United Kingdom, Canada, and Australia, are developing scorecards which include measures of pediatric health care quality. Resources for health care are finite, and high-income countries are facing rising pressures to maximize the value of health care expenditures. Information on provider performance can reduce the information deficit between purchasers and providers of health care, providing incentives for purchasers and consumers of services to use the best providers, and for providers to improve performance.”

2) In general, the authors could make further use of illustrations from pediatric care than they do, in order to ensure that this article has a unique contribution within the indicator development canon. What challenges to the international indicator development literature does application within pediatric care bring? (There are some mentioned, e.g. on proxy HRQL measurement, but perhaps worthy of a paragraph or emphasis in the conclusions? Do any of these issues have implications for composite indicator development – more or less likely to be affected by the advantages/disadvantages presented?) And the examples they do use could be explored further. E.g. they make the crucial point that preventive activities and the importance of long-term outcomes are particularly important in pediatrics – this also raises general questions about the usual definitions (or focus within those definitions) of quality, and could be emphasised more. (Also, on a more specific note, the reality of Table 3 is that it is relatively poorly populated – can the authors suggest more indicators that could be useful if further developed?)

Response: We have tried to highlight several aspects of quality measurement that feature prominently in pediatrics. In general while most of these aspects are not unique to pediatrics, we believe they are of different import when compared to measuring quality of care provided to adults.

We have moved the pediatric specific section into the background section for emphasis and have added the following sections:

§ “Death is fortunately a relatively uncommon outcome in children, even in acute care settings. As such, mortality in isolation is a poor discriminator of care quality. Moreover, mortality does not always represent poor care quality but may reflect appropriate decisions by providers and parents to provide comfort care for children with irreversible and debilitating conditions. Attitudes towards comfort care are likely to vary among providers, regions, and parental caregivers, which further undermines the ability of mortality to discriminate hospital quality of care. Nevertheless, mortality is an important balancing measure, which ensures that hospitals do not receive undue credit for measures that are sensitive to mortality (e.g. length of stay). We therefore recommend including mortality into composite indicators measuring the quality of acute care settings. However, its effect on provider performance should be subject to sensitivity analysis, as should be its weighting.”

§ “In addition, studies by Saigal and colleagues suggest that patient utilities may not be stable over a patient’s life, even in light of stable chronic disease. This suggests that the effect of patient preferences on provider performance on a composite indicator of quality should be assessed by allowing preferences to

vary over a reasonable range in sensitivity analyses.”

§ “In addition, developers may choose to apply differential weights among preventive care measures based on their value to public health in a given society (e.g., the prevention of obesity may be of greater value than administration of polio vaccine).”

§ The following sentence has been added to the conclusion. “Pediatric quality of care measurement presents unique challenges to researchers in this field and much empirical work remains to create best practice in composite indicator development.”

§ With regard to table 3, we had already included metrics not deemed of high enough quality by the expert panellists. The empty cells of the matrix highlight opportunities for measure development. We prefer not to add metrics at this point as these would be unlikely to find acceptance among providers without further study.

3) Page 7 – another disadvantage of composite indicators is that performance in one area might cancel out performance in another. This is not identified at this point (though is implied later in the article where the authors talking about the need for ‘unidirectionality’ (page 11) and the importance of this consideration when determining the appropriate aggregation methods (page 12-13)).

Response: The section on page 7 was altered to include the reviewers suggestion:

“In addition, a summary score may inaccurately suggest that providers are average if good scores on one metric compensate for poor performance on some metrics. In fact, “average” providers may be “poor” providers for patients whose needs are within the low scoring performance areas. Some of these dangers can be countered by using dissemination formats that convey results accurately while avoiding oversimplification (such as the ability to “drill down” into individual components of the composite), and by making the process of indicator development explicit and transparent to all stakeholders. In addition, statistical techniques, such as multi-criterion analysis mitigate the problem of performance averaging.”

4) Page 7 to 8 – use of composite indicators for quality improvement – this exploration needs to be unpicked further. While a comprehensive/composite indicator approach is ideal for identifying the need for improvement (so as not to overfocus on individual areas that might not be the most important), it might not be the most appropriate approach to track improvement. E.g. the initial comprehensive assessment is likely to identify specific areas for attention, and specific evidence-based interventions should then be developed. The most useful tracking measures are likely to be focused within this area and linked to the relevant evidence-based interventions. Of course, later checking with the comprehensive indicator set to ensure that this focus has not distorted quality of care is important.

Response: We have attempted to incorporate the reviewers suggestion and

added the following paragraph:

“A multi-dimensional approach to quality measurement via composite indicators may support such a multi-dimensional approach to quality improvement. Composite indicators and their individual components may identify specific areas for attention, for which specific evidence-based interventions are then developed. The success of improvement can then be cross-checked with the comprehensive measure set to ensure that this focus has not worsened quality of care in another area. However, targeting individual quality metrics may lead to piecemeal rather than system-based efforts in quality improvement. Potentially, larger leaps in improvement may result from systems-based interventions that affect multiple areas of care simultaneously and have the potential to spread throughout the care service and the institution at large. Improving safety attitudes among staff is an example of a system-based intervention which may improve outcomes and propagate throughout an institution. Whether composites are used to track improvement targeting individual or multiple metrics will depend on local resources, support systems, expertise and institutional capacity. In either application, composites would allow tracking of overall improvement and alert users to potential effects on other measures of quality, while their individual components retain the ability to monitor specific improvement activities.”

5) It emerges during the article that the framework used is a combination of the OECD approach and Profit et al's quality matrix for NICU. It would be helpful if this were signposted in the abstract and earlier in the article.

Response: The abstract already contains a reference to the combination of frameworks applied for this methods paper. On the reviewer's suggestion, we have also added the following sentence to the background section:

“The final approach to composite indicator development is the result of a combination of approaches described by Profit and colleagues with methods developed by the European Commission Joint Research Center (EC-JRC) and the Organization for Economic Cooperation and Development (OECD), henceforth simplified as JRC. (12;15) In the discussion section we will spotlight pediatric-specific aspects in composite indicator development which require empirical research. These include paucity of interactions with the health care system, paucity of critical health outcomes, and availability of quality of life and prevention metrics. We will focus on aspects important to pediatrics because aggregate performance measurement is comparatively new to this field.”

6) Pages 11 to 12 – many of the (excellent) technical points made on the framework (including on aggregation methods) can presumably be referenced from the OECD papers and others?

Response: We have added an explicit statement that refers readers to the handbook for additional detail.

7) Page 14 – is there a further step? To assess the validity of the resulting composite indicators? (e.g. face validity, criterion validity against other global indicators of quality in this field?)

Response: The suggested validation step is described by the OECD/JRC in Step 8. We have added the following to this step.

“In addition, a PICU indicator can be correlated with indirect measures of quality, for example, measures of patient safety culture, for purposes of criterion validation of an inherently immeasurable construct.”

Minor essential revisions

1) Page 16 – as the authors correctly identify with ref 18, there is little evidence that benchmarking information affects patient choice. So their earlier comment ‘can be leveraged’ should perhaps be more circumspect? (e.g. ‘might’?)

Response: This has been changed.

2) Page 9 – did the authors find any other guidance on composite indicator development other than the OECD? (Or can they state this is the only available?)

Response: This has been changed.

3) Page 10 – can the authors reference the desirable characteristics for indicators?

Response: This has been referenced.

4) Page 10 – I wasn’t sure about the definition used for reliability – is ‘systematic bias’ the right phrase?

Response: We changed this to: “precision of point estimates”.

5) Page 10 – ‘step 2’ implies that the framework is only interested in already developed (validated, reliability –tested etc) indicators. Does this need to be made clear earlier, or is there an option of developing new indicators? (In which case, some of the methodological challenges in this can be referenced from elsewhere.)

Response: In general it may be wise for developers to choose metrics supported by solid evidence; else the composite may not achieve acceptability among users. However, developers could certainly go through their own measure selection process and include metrics based on statistical properties or develop new metrics. Metric development is a difficult process its description is beyond the scope of this manuscript. In practice, we think it more likely that composite developer would draw from a finite set of already developed and available quality metrics.

However, we have added a section to the manuscript highlighting the trade-offs between measure selection based on expert opinion and that based on statistical properties:

“Given the high stakes involved with regard to comparative performance measurement, we think that the metric selection process is of cardinal importance to the composite indicator’s acceptability among users. Selection should therefore rely on a rigorous and explicit process so that each metric is appropriately vetted with regard to its strengths and weaknesses.”

And later... “The selection of metrics may be informed by expert opinion or based on statistical methods. The use of expert opinion and a formal metric vetting process may enhance the composite index’ external validity and thus user acceptability. On the other hand, a statistical approach to metric selection may be less time consuming and result in a more parsimonious measure set, but may lack external validity with users. Importantly, either approach should result in a measure set that clinically represents the underlying quality construct and balances external validity and parsimony. Future updates of the composite should incorporate user feedback and new scientific evidence, which may require changes to the existing measure set. As mentioned above, metric selection and attribution to domains of care informs the structure of the composite with regard to its sub-pillars. We recommend a minimum of three measures per pillar, meaning that given the dearth of available data, a PICU composite would currently lack at least the domains of equity and efficiency. Whether a metric, such as severity-adjusted length of stay, can nevertheless be incorporated into the composite can be investigated by examining whether it statistically maps on another domain.”

Reviewer 2

Major Areas

1) One of the strengths of this paper is the use of an example when discussing the 10 steps of composite development. However, for 5 of these steps (#4, 5, 6, 7, and 9), the example is not referred to. It would help clarify some of these complex steps, such as the treatment of missing data or weight and aggregation, if the authors could provide some examples from the development of the PICU composite measure.

a. Response: This manuscript was primarily conceived as a methods paper presenting to the research community an explicit and transparent framework for the development of composite indices of quality for pediatric care. Unfortunately, development of the entire composite index of PICU quality is beyond the scope of this methods paper. However, we have attempted to provide more clarity to some of the steps mentioned by the reviewer. We have added additional detail to most of the steps while attempting to limit the length of the manuscript.

2) There are 2 steps where further explanation is warranted. First, the issue of missing data is an important issue, and one that is relatively unappreciated in the pediatric quality measure literature. How would a developer determine whether missing data are randomly or non-randomly missing? Also, one or two clarifying sentences about the advantages and disadvantages of imputation of data would help a reader who is unfamiliar with this technique. Second, There should be more discussion of how a developer should select the metrics for a composite measure. What are the advantages and disadvantages to expert opinion versus statistical selection versus a broad selection of metrics that are reduced in future steps? Are the metrics reduced in future steps, and how does one do that?

Response:

i. The following has been added to the manuscript to address how a developer would go about determining whether missing data are randomly or non-randomly missing.

“Missingness status (random vs. non-random) can be investigated directly, with a missing data analysis (MDA) establishing whether missingness is associated with measured and available variables of interest. However, these investigations have limits: Variables potentially associated with identified missingness cannot be investigated if they have not been measured within the context of the study at hand and remain external to a MDA, constraining its conclusions.”

In addition, the following section about data imputation has also been added to the manuscript.

“Because many benchmarking activities have reputational and/or financial implications, it may be prudent to assume data are not missing at random. The developer could give providers the benefit of the doubt and assign a probability of 0 to missing data, here implying a negative outcome did not occur. However, this may provide an incentive to game the system and not provide data on patients with poor outcomes. A similar incentive is provided if missing data are (a) excluded or (b) imputed using a hospital’s average performance. More sophisticated methods for imputing missing data, based on regression analysis or probabilistic modelling, attempt to impute a true value based on a hospital’s results with similar patients.(33;34) Yet even these methods may underestimate true data values if providers intentionally game the system. Conversely, assigning a value of 1 to a missing data point may punish providers unfairly for something beyond their control, e.g., data lost in the abstraction and transmission phase of the benchmarking activity. Despite other disadvantages, this approach may encourage complete record keeping. To be successful, missing value imputation must proceed via a carefully selected strategy appropriate for the dataset under analysis. An inappropriate imputation strategy may itself introduce bias into analytic results. Complete-case-analysis, which sidesteps imputation and missingness by use of missing case deletion (listwise or pairwise) will produce biased results when non-random missingness is present. Common imputation strategies, such as mean imputation, last observation carried forward, or mean difference imputation, will also introduce bias into results when missingness is non-random. A multiple imputation strategy, preserving the variance of a variable with missingness, will create multiple imputed values and weights to be combined in producing a consistent outcome estimator while accounting for errors in the imputation process itself.(35;36) Thus, a multiple imputation strategy carefully matched to the characteristics of the dataset containing missingness offers a “best practice” solution.”

ii. Measure selection: Please see response to Reviewer 1, Minor revisions, number 5.

3) This is a study specifically focused on pediatric composite measure development. The last section of the manuscript, beginning on page 14, discusses some differences between pediatric and adult measures; however,

these issues should be included in the background of the manuscript. Also, issues of long-term versus short-term outcomes of care and limited mortality of children should be raised, to emphasize how different children are from adults.

Response: We have restructured this section for greater clarity and provided the following subheadings which distinguish pediatric from adult indicators: Paucity of interactions with the health care system, paucity of critical health outcomes, quality of life metrics, and prevention metrics.

Regarding a section addressing the of limited mortality in pediatrics, please see response to Reviewer 1, major revisions, number 2.

Discretionary Revisions

1) The issue of improved power to detect an outlier or a difference between hospitals with a composite measure is not discussed in either the quality improvement or benchmarking sections of the paper (pages 6-8).

Response: Since this theory still has to be tested empirically we decided not to place undue emphasis on it.

2) The authors refer to the issue that some metrics may span multiple domains of care. The implications of this are not discussed, particularly if a metric has different and opposite effects on different domains of care.

Response: The following has been added to this section:

The resulting composite would combine metrics of structure, process, and outcomes, a combination suggested by others, and be based on sub-pillars based on the IOM domains of quality of care. Metrics within each pillar will correlate amongst each other and with those of other pillars. Ideally, one would expect moderately high correlations of metrics within pillars and low correlations between pillars. High correlation of metrics between pillars may indicate a metric's multi-dimensional nature.

3) The authors raise the new issue of preventive outcomes of care. How are pediatric patients different from adult patients? Have adult measures not focused on preventive outcomes in their composite development?

Response: The author's believe that many of the issues found in children such as those relating to preventive care, can also be found and are being applied in adult care, making the methods presented in this paper applicable to adult researchers. However, we believe that some of the aspects mentioned feature more prominently in children. For example, high quality preventive (and acute) care provided to children has the capacity to improve the health and social trajectory of a child for a lifetime, whereas a preventive medicine such as aspirin to prevent a heart attack, while important, may save only a few years of life. However, not all preventive care is equally important. Few children are likely to suffer from polio these days even if not immunized. Yet, a physician who is able to promote healthy lifestyles among his patients and reduces rates of obesity may have a huge impact on patients' long-term health outcomes. Therefore, compliance with high value preventive guideline recommended care potentially

should be weighted higher than other areas of preventive care. We have added a section in the manuscript highlighting this point (see also above).