PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	A systematic approach to evaluating and confirming the utility of a suite of national health system performance (HSP) indicators in Canada: a modified Delphi study
AUTHORS	Fekri, Omid; Leeb, Kira; Gurevich, Yana

VERSION 1 - REVIEW

REVIEWER	Fabrizio Carinci University of Surrey, United Kingdom
REVIEW RETURNED	09-Nov-2016

GENERAL COMMENTS	The paper reports on the experience of a systematic evaluation of performance indicators at system level in Canada. Consistently with recent work conducted by the OECD, this procedure reflects the needs of making indicators relevant for national policies, involving experts and representatives from different disciplines that should judge on different criteria in a transparent procedure. Such a dialogue between policy and academia is essential to move the analytical level closer to everyday practice. However, it does not happen in many countries and for this reason the paper is useful to help overcoming the existing gaps and strengthening the existing systems of indicators. A revision of the paper should be submitted, taking into account each of the following points: 1. explain why average likert scores have been preferred over the median. Demonstrate how excess variability eg range or distance from median has been dealt with to resolve discordances and/or unclear situations 2. explain better if/how results from 18 dimensions were summarized into a composite measure used to flag each indicator for candidate inclusion/retirement. If this is done only using the mean, this could hide very variable results. Some criteria may be more important than others, eg actionability (as reflected in part of the paper). This is not clear from the current text. Explain better how such variability is considered (if at all) by the consensus survey. 3. the consensus panel does not seem to include representatives of citizens and patients. Please clarify. That seems important to ensure transparency of the system. Explain why eventually this was not needed in this case. Otherwise, state as a clear limitation that should be overcome in other experiences. 4. CIHI uses indicators to populate the national portal on health systems performance "Your health". How will this experience

	contribute to that? Apparently should: if so, it could be argued that the inclusion of citizens in the procedure was certainly needed. 5. Discuss to what extent this experience is applicable elsewhere. What would be the barriers and which systems (eg NHS vs private, centralized vs decentralised, etc) would be best placed to adopt this approach? Make practical examples including references. Please consider issues related to the information infrastructure (some of the literature from the OECD could help). This would really strengthen the usability of the paper. 6. Throughout the paper, authors refer to indicators in a very abstract manner. It would be much better to refer directly to the actual indicators/diseases/areas kept or dropped in the scheme. This would make the paper more directly usable (and less methodological oriented)

REVIEWER	Richard H Osborne NHMRC Senior Research Fellow Head, Health Systems Improvement Unit School of Health and Social Development Centre For Population Health Research, Faculty of Health, Deakin University, Australia. Affiliate Professor of Health Literacy, University of Copenhagen, Denmark.
REVIEW RETURNED	13-Nov-2016

GENERAL COMMENTS	This is an important report of Canada's process to reconfirm national health system performance indicators. Many countries undertake such reviews but few document the process. The process described has the potential to be taken up by other countries to improve local selection of performance indicators. Running title: "HSP" will be unknown to readers. consider revising. Introduction I am not convinced 'more is better' is a common adage, it is more common to hear 'less is more'. Consider revising. Consider that the growth in indicators has also come from the desire of lead agencies to operationalize their accountability to the public, and consumer demand for transparency. Methods Present the options within the Likert scale – this will help the reader understand the scoring, and robustness of the assessment approach. Was the same scale used for all surveys? Currently in P5L22 it is described as a score between 1 and 9, this sounds like a rating scale, not a Likert scale. Results It is unclear what is meant by a holistic process. If this is method, explain in Methods. On P5L28 is stated "…based on a holistic assessment of all 18 criteria." This is not clear. P5L40 also seems to add new/additional procedures over what is

	stated in the Methods – namely the CIHI leadership "performed a similar exercise" - is unclear if results for this are presented or integrated. While I appreciate that these types of consultations can evolve over time, and it is difficult to describe what happens in traditional Methods/Results sections without narrative, it isn't clear what was done at what time. The authors should consider a flow diagram of the processes, inputs and outputs (decisions). The term 'solidified' is ambiguous – do you mean ratified, consolidated, or other? P8L54 please provide more information on the demographics of the pre-conference survey participants and whether they were reasonably/appropriately representative and authoritative to derive robust decisions. Where there any gaps in representation? Consider this for all stages. Discussion The process employed appears to be a confirmatory exercise limited to the indicators currently in place. Since the last review, have healthcare, community attitudes, community needs, or health inequalities etc changed such that new indicators are needed? This is a weakness of the Delphi approach which doesn't easily allow new issues/insights to be included when a predefined set is the starting point. Delphi may also force the removal of infrequent but critical concepts from the pool of items/concepts. The authors should consider whether this is a potential weakness. The paper would be strengthened by some critical reflection of the processes used, including the breadth and depth of those consulted, who was missed or not sufficiently represented, and how they might do it better next time. Tables These need to be standalone. The 'Mean Likert score' (if it is a true Likert scale) should be defined. In table 3 the last column is "Pre–Consensus Conference Survey Agreement (as a % of responses)". It is important to state the number of non-respondents and reasons for these. Also, there were 48 respondents (representing a 79% response rate). Of the 48, some answered Agreed, Disagreed or had No Opinion. Also consider providing the number of respondents offering 'no-opinion' and why such response might come from 'experts'. A flow diagram would help.

VERSION 1 – AUTHOR RESPONSE

Response to Reviewer 1 (Fabrizio Carinici)

We thank Dr. Carinci for these helpful comments.

1. explain why average likert scores have been preferred over the median. Demonstrate how excess variability eg range or distance from median has been dealt with to resolve discordances and/or unclear situations

• We now note our use of mean scores, and review of all respondent ratings (2nd paragraph of Internal CIHI modified-Delphi sessions within Methods section).

2. explain better if/how results from 18 dimensions were summarized into a composite measure used to flag each indicator for candidate inclusion/retirement. If this is done only using the mean, this could hide very variable results. Some criteria may be more important than others, eg actionability (as reflected in part of the paper). This is not clear from the current text. Explain better how such variability is considered (if at all) by the consensus survey.
• We have added further text in Methods section, second paragraph, last two sentences.

3. the consensus panel does not seem to include representatives of citizens and patients. Please clarify. That seems important to ensure transparency of the system. Explain why eventually this was not needed in this case. Otherwise, state as a clear limitation that should be overcome in other experiences.
• We now note the preceding public engagement exercises undertaken by CIHI to incorporate the patient/public voice (paragraph 2 of strengths and weaknesses of the study).

4. CIHI uses indicators to populate the national portal on health systems performance "Your health". How will this experience contribute to that? Apparently should: if so, it could be argued that the inclusion of citizens in the procedure was certainly needed.
• This item has been addressed with previous comment (paragraph 2 of strengths and weaknesses of the study).

5. Discuss to what extent this experience is applicable elsewhere. What would be the barriers and which systems (eg NHS vs private, centralized vs decentralised, etc) would be best placed to adopt this approach? Make practical examples including references. Please consider issues related to the information infrastructure (some of the literature from the OECD could help). This would really strengthen the usability of the paper.
• We have elaborated on the applicability of this experience to other environments, and cited an appropriate source on health information infrastructure.

6. Throughout the paper, authors refer to indicators in a very abstract manner. It would be much better to refer directly to the actual indicators/diseases/areas kept or dropped in the scheme. This would make the paper more directly usable (and less methodological oriented)
- We thank the reviewer for the suggestion. We have reviewed the entirety of the manuscript to address this. Within the Results section, we have already referred, when appropriate, to specific indicators to provide further details for the reader. The explanatory notes within the Tables refer to diseases/indicators that were retired due to their limitations. We feel these references currently suffice.

Reviewer 2 (Richard H. Osborne)

We thank Dr. Osborne for these helpful comments.

Running title: "HSP" will be unknown to readers. consider revising.
• Health System Performance is now spelled in full in the running title.

Introduction
I am not convinced ‘more is better’ is a common adage, it is more common to hear ‘less is more’. Consider revising.
Consider that the growth in indicators has also come from the desire of lead agencies to operationalize their accountability to the public, and consumer demand for transparency.

• We have revised this statement in the introductory paragraph.
• Additionally, in paragraph 3, we note the rise in accountability and demand for quality improvement data as contributors to “indicator chaos” (a key finding of citation #6).

Methods
Present the options within the Likert scale – this will help the reader understand the scoring, and robustness of the assessment approach. Was the same scale used for all surveys? Currently in P5L22 it is described as a score between 1 and 9, this sounds like a rating scale, not a Likert scale.
• We have elaborated on the Likert scale and indicator disposition options.

Results
1. It is unclear what is meant by a holistic process. If this is method, explain in Methods. On P5L28 is stated “…based on a holistic assessment of all 18 criteria.” This is not clear.
• We have replaced the three instances of “holistic” with improved text to describe the process of considering all 18 evaluation criteria.

2. P5L40 also seems to add new/additional procedures over what is stated in the Methods – namely the CIHI leadership “performed a similar exercise” - is unclear if results for this are presented or integrated. While I appreciate that these types of consultations can evolve over time, and it is difficult to describe what happens in traditional Methods/Results sections without narrative, it isn’t clear what was done at what time. The authors should consider a flow diagram of the processes, inputs and outputs (decisions).
• We clarify that the leadership group repeated the preceding exercise performed by the technical group (1st sentence, last paragraph, Internal CIHI modified-Delphi sessions (Methods)).

3. The term ‘solidified’ is ambiguous – do you mean ratified, consolidated, or other?
• We have replaced the term “solidified” to now read “consolidated and finalized based on group consensus” (2nd last sentence, last paragraph, Internal CIHI modified-Delphi sessions (Methods)).

4. P8L54 please provide more information on the demographics of the pre-conference survey participants and whether they were reasonably/appropriately representative and authoritative to derive robust decisions. Where there any gaps in representation? Consider this for all stages.
• We note in the Methods section the professions of survey respondents (2nd sentence, Pre-conference survey of stakeholders on indicator recommendations).
• We provide further demographic findings in the Results section, Pre-Consensus Conference survey, 3rd sentence.
• We also reflect in the Discussion section (2nd paragraph under Strengths and weaknesses of the study) on the exclusivity of stakeholders and clients in the survey.

Discussion
1. The process employed appears to be a confirmatory exercise limited to the indicators currently in place. Since the last review, have healthcare, community attitudes, community needs, or health inequalities etc changed such that new indicators are needed? This is a weakness of the Delphi approach which doesn't easily allow new issues/insights to be included when a predefined set is the starting point. Delphi may also force the removal of infrequent but critical concepts from the pool of items/concepts. The authors should consider whether this is a potential weakness.
• We reflect on this in a new paragraph at the end of the Discussion section, as well as in the Strengths and Weaknesses of the study.

2. The paper would be strengthened by some critical reflection of the processes used, including the breadth and depth of those consulted, who was missed or not sufficiently represented, and how they might do it better next time.

• We reflect on this in the third paragraph of the Discussion section.

Tables
1. These need to be standalone. The ‘Mean Likert score’ (if it is a true Likert scale) should be defined.
• We now define this in the three applicable tables.

2. In table 3 the last column is “Pre–Consensus Conference Survey Agreement (as a % of responses)”. It is important to state the number of non-respondents and reasons for these.
• We have noted in the Results section, subheading Pre-consensus Conference survey, that a response was mandatory for the questionnaire.

3. Also, there were 48 respondents (representing a 79% response rate). Of the 48, some answered Agreed, Disagreed or had No Opinion. Also consider providing the number of respondents offering ‘no-opinion’ and why such response might come from ‘experts’. A flow diagram would help.
• We have noted in the Results section, subheading Pre-consensus Conference survey, the potential reasons why respondents may select ‘No Opinion’ in the survey, the average and range for the indicators.

VERSION 2 – REVIEW

REVIEWER	Fabrizio Carinci University of Surrey, United Kingdom
REVIEW RETURNED	30-Jan-2017

GENERAL COMMENTS	The authors have duly addressed all points raised in the initial assessment of the manuscript. As far as my review is concerned, the paper may be accepted in its current revised format.