BMJ Open

# Systematic approach to evaluating and confirming the utility of a suite of national health system performance (HSP) indicators in Canada: a modified Delphi study

Omid Fekri, Kira Leeb, Yana Gurevich

**To cite:** Fekri O, Leeb K, Gurevich Y. Systematic approach to evaluating and confirming the utility of a suite of national health system performance (HSP) indicators in Canada: a modified Delphi study. *BMJ Open* 2017;**7**:e014772. doi:10.1136/bmjopen-2016-014772

▸ Prepublication history for this paper is available online. To view these files please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2016-014772).

Received 17 October 2016
Revised 24 December 2016
Accepted 16 February 2017

Canadian Institute for Health Information, Toronto, Ontario, Canada

**Correspondence to**
Kira Leeb;
kleeb@cihi.ca

## ABSTRACT

**Objectives:** Evaluating an existing suite of health system performance (HSP) indicators for continued reporting using a systematic criteria-based assessment and national consensus conference.

**Design:** Modified Delphi approach with technical and leadership groups, an online survey of stakeholders and convening a national consensus conference.

**Setting:** A national health information steward, the Canadian Institute for Health Information (CIHI).

**Participants:** A total of 73 participants, comprised 61 conference attendants/stakeholders from across Canada and 12 national health information steward staff.

**Primary and secondary outcome measures:** Indicator dispositions of retention, additional stakeholder consultation, further redevelopment or retirement.

**Results:** 4 dimensions (usability, importance, scientific soundness and feasibility) typically used to select measures for reporting were expanded to 18 criteria grouped under the 4 dimensions through a process of research and testing. Definitions for each criterion were developed and piloted. Once the definitions were established, 56 of CIHI's publicly reported HSP indicators were evaluated against the criteria using modified Delphi approaches. Of the 56 HSP indicators evaluated, 9 measures were ratified for retirement, 7 were identified for additional consultation and 3 for further research and development. A pre-Consensus Conference survey soliciting feedback from stakeholders on indicator recommendations received 48 responses (response rate of 79%).

**Conclusions:** A systematic evaluation of HSP indicators informed the development of objective recommendations for continued reporting. The evaluation was a fruitful exercise to identify technical considerations for calculating indicators, furthering our understanding of how measures are used by stakeholders, as well as harmonising actions that could be taken to ensure relevancy, reduce indicator chaos and build consensus with stakeholders.

## Strengths and limitations of this study

- This exercise utilised an extensive suite of criteria to evaluate health system performance indicators.
- Multiple evaluation modalities were used to solicit feedback from evaluators.
- A large number of stakeholders participated in an inperson consensus conference.
- Assessment criteria and processes may not apply in other evaluation contexts.

## INTRODUCTION

Health indicators offer valuable insight into the performance of health systems and the health of populations. As the discipline of health system performance (HSP) measurement has grown over the decades, so too have the number of available health measures. In Europe alone, journal publications related to performance indicators increased at a rate of ~20% annually between 2000 and 2009.[1] However, continuing to increase the number of indicators reported runs contrary to, and inhibits, the provision of concise findings on the performance of health systems.[2] Health measure producers and users are constrained with finite resources, and must make important decisions on which indicators they deem important, have high utility, are valid and are feasible. Periodic reviews of indicators and conceptual frameworks can ensure their continued relevance and efficacy.[3]

Two national agencies, the Canadian Institute for Health Information (CIHI) and Statistics Canada, have collaborated for more than 15 years on developing and publicly reporting health measures for health regions, provinces and territories as part of the Health Information Roadmap.[4] Over the years, the number of indicators has increased from 13 in 2000 to more than 80 in 2014. This in part reflects the growing information needs of healthcare systems in general. For example, new indicators measuring outcomes, wait times and patient safety were the areas of focus for development in recent years. CIHI also expanded its indicator reporting over the years by refining the granularity of public reporting, and in 2007 began public reporting of health indicators for acute care hospitals in Canada. The indicators were developed and reported on according to the CIHI–Statistics Canada Health Indicator Framework.[5] In 2012, the suite of publicly available indicators at the hospital level was expanded substantially and in 2015, was expanded again to include indicators for long-term care homes.

After a period of rapid growth in public reporting of indicators likely due to the rising demand for accountability and quality improvement data as well as increases in capacity-building activities across the country, health system managers identified that having too many indicators to monitor and respond to was not achieving the goal of helping understand how well the healthcare system was performing. In 2010, this phenomenon was coined 'indicator chaos',[6] and initiated a new focus on streamlining indicator reporting and development activities.

Partly in response to this notion of indicator chaos, but also in efforts to ensure relevancy and efficiency, CIHI initiated a programme of work aimed at streamlining health system reporting in Canada. As part of this work, CIHI developed a new HSP framework to better reflect the relationship between indicator measurement and health system goals.[5] CIHI also recognised the need to ensure that the indicators being produced and reported reflected these goals and contributed to a broader understanding of HSP rather than continuing to add to the reporting and monitoring burden across the country. This required a systematic indicator evaluation process that could be repeated periodically to inform indicator reporting initiatives across the organisation and possibly beyond.

Coincidentally, every 5 years (1999, 2004, 2009, 2014), CIHI and Statistics Canada invite stakeholders from across the country to a national Consensus Conference on Health Indicators to discuss priority setting of indicator development and reporting for the next half decade.[7–10] The latest such conference (held in 2014) provided an opportunity to present the results of the internal evaluation of publicly reported indicators and to validate the results with stakeholders.

This paper describes CIHI's approach to evaluating a set of HSP indicators using a systematic criteria-based assessment tool and process. The results of the pilot—including achieving reconfirmation through a national consensus process—and possible next steps for broader implementation of the strategy are also presented in the paper.

## METHODS

This project had four distinct components:

1. Process and criteria development for systematic evaluation of HSP indicators.
2. Internal CIHI modified Delphi sessions.
3. Preconference survey of stakeholders on indicator recommendations.
4. Presentation and ratification of results at the national Consensus Conference.

### Systematic evaluation of HSP indicators

The Institute of Medicine's (IOM's) *Recommendations for Measure Selection Criteria*[11]—usability, importance, scientific soundness and feasibility—are consistently used in the evaluation and selection of health measures.[12] While many examples in the literature employ these four domains of criteria, we saw the need to expand the dimensions to include other criteria within three of the four domains. Through a process of research and testing, we arrived at a total of 18 criteria points organised around the 4 IOM domains (see table 1) that were feasible to apply and that held meaning to our project objective regarding continued reporting of indicators. There is congruence between these criteria and CIHI's Data Quality Framework[13] dimensions of accuracy, timeliness, comparability, usability and relevance. Over a period of 2 months, 56 of CIHI's suite of HSP indicators were assessed against these 18 criteria. To aid evaluators in their subsequent reviews, we created a one-page summary for each indicator denoting results for each evaluation criterion.[14]

### Internal CIHI modified Delphi sessions

Two groups within CIHI participated in the evaluation. First, a technical group of experts (n=6) (comprised epidemiologists, methodologists and statisticians) independently reviewed each indicator and criterion point, and provided a Likert Scale score between 1 and 9. Likert scores were assessed as follows: 7–9 was considered as robust strength for the indicator and agreement for continued reporting; 4–6 denoted equivocal evidence and further discussion at inperson Delphi session is required and 1–3 was considered as weak support for the indicator suggesting it should be retired. Respondents were instructed to produce a Likert score and disposition recommendation based on their assessment of all 18 evaluation criteria as a whole. We therefore forewent weighting evaluation criteria. This allowed for flexibility and context in instances where some criteria proved more informative than others.

Likert scores were averaged and presented at an internal inperson Delphi session as a basis for discussions, but were not automatically tied to a final result of continued indicator reporting. The mean was used to average scores as there were no outlier values across responses. Furthermore, all individual respondent ratings were shown alongside the mean score, thereby illustrating the level of concordance. Beginning with the

**Table 1** Evaluation criteria

| Domain/criterion | Definition |
|---|---|
| **Usability** | |
| Granularity of reporting | Reporting at national, provincial/territorial, regional and facility levels |
| Pan-Canadian coverage | Extent of participation from all provinces and territories |
| Comprehensiveness | Proportion of providers submitting data for the indicator |
| Usage | Level and extent of usage |
| Dimensionality | Ability to break down results by age, sex, socioeconomic status and other dimensions |
| Timeliness | Latest year of available results |
| Reporting frequency | Whether indicator is reported quarterly, annually or other |
| Accessibility | Whether the indicator is publicly and/or privately reported |
| Trendability | Number of years of available results for trending |
| **Importance** | |
| Relevance | Environmental scan identified uses of indicator by stakeholders |
| Actionability | Extent to which providers can meaningfully influence the indicator |
| Stakeholder follow-up | Number of data and methodological requests within last fiscal year |
| Sufficient volumes | Percentage of results suppressed (due to low counts) |
| Significance of variation | Degree of variation across reported values |
| **Scientific soundness** | |
| Data quality | Strength of data quality, ability to validate results, based on standards |
| Validity review | Extent and frequency of reviewing indicator's validity/methodology |
| Participation bias | Mandatory or voluntary participation by providers |
| **Feasibility** | |
| Production cost | Extent of staff/resources to produce indicator |

lowest average scores, each indicator was discussed, pertinent commentary synthesised and a final consensus reached on a disposition recommendation. Disposition options for indicators were retain, recommend further research and development (R&D) or consultation, or retire.

Recommendations of the technical group's Delphi sessions were then presented to the CIHI HSP leadership group (n=6) (comprised senior managers and researchers) who repeated the preceding exercise. First, they were asked to independently review all results to date (including indicator assessments and Likert scores, commentary and disposition recommendations). Results of their individual assessments were collated and presented at an inperson session. Disposition recommendations for each of the 56 indicators were consolidated and finalised based on group consensus. The RAND/UCLA Appropriateness Method[15] guided our internal iterative modified Delphi sessions.

### Preconference survey of stakeholders on indicator recommendations

A pre-Consensus Conference survey solicited initial feedback on recommendations. The online survey was available for a period of 6 weeks prior to the conference. Consensus Conference participants were chosen from an existing list of CIHI partners, stakeholders and clients; participants were largely hospital/health region CEOs, academics and researchers, representatives from ministries of health, clinicians and national collaboration partners involved in measuring and monitoring the performance of the healthcare system. An electronic survey was emailed to conference participants along with background documentation on the evaluation process, methodology and recommendations. The survey asked respondents whether they Agreed, Disagreed or had No opinion on the recommendation to retire select HSP indicators as per recommendations from CIHI's internal review.

### National consensus conference presentation

There were 61 participants at the invitational inperson Consensus Conference held in Toronto on 16 and 17 October 2014. Results of the preconference survey were presented. A threshold of 70% agreement by respondents was used to automatically pass recommendations or to otherwise hold further group discussion at the conference. An external moderator facilitated discussion and voting on final indicator dispositions.

## RESULTS

### Systematic evaluation of HSP indicators

The systematic evaluation of HSP indicators was a summative process considering 18 criteria points. Some criteria differentiated indicators more than others. For example, a small number of criteria resulted in mostly uniform findings for the suite of HSP indicators. However, when assessed alongside remaining criteria, important contextual considerations can be gleaned. Notable findings are summarised below by criterion.

### Usability

The *granularity of reporting* criterion identified nuances inherent within public reporting purposes. There are ~100 administrative health regions in Canada, and ~600 acute care hospitals. Twenty-nine indicators are reported at the regional level, and 27 are reported at the hospital/facility level. All indicators are reported at an aggregate provincial/territorial and national level.

With respect to *pan-Canadian coverage*, 44 of 56 indicators provided complete pan-Canadian coverage (all provinces and territories). The province of Quebec does not have available or comparable data for a dozen indicators. Similar to the criterion of pan-Canadian coverage, the *comprehensiveness* criterion assessed the inclusiveness of health services providers that submit data towards the indicator. For example, the mental illness hospitalisation indicator includes data on mental health patients treated in general hospitals only, while hospitalisations at free-standing psychiatric institutions are not included due to the differences in data collection.

For the *usage* criterion, we polled CIHI HSP staff responsible for interacting with clients on indicators and data requests. This provided a proxy for the level and extent of the indicator's usage by clients. The 56 indicators under evaluation were rated as high (n=33), medium (n=15) or low usage (n=8).

With regard to *dimensionality*, breakdowns of indicator results by dimensions of sex and socioeconomic status (SES) are available where applicable. Thus, 15 indicators are reportable by SES and 14 are reportable by sex.

In terms of *Timeliness, Reporting frequency and Accessibility*, all 56 indicators were publicly reported annually within 10 months of the relevant data being available for analysis. At the time of the evaluation, all HSP indicators were accessible publicly through online publications such as the Health Indicators e-Publication. Additionally, a majority of facility-level indicators are available to providers through private online tools to allow for more granular breakdowns of results and peer comparative reports.

For the *trendability* criterion, it was found that time trends vary by indicator. For example, the set of facility-level indicators was largely first reported beginning with 2007 data. Results for select regional indicators dated back to 1997. Overall, regional indicators possessed almost twice as many available years of results compared with facility indicators, a nature of the timing of reporting programmes.

## Importance

As a proxy measure for *relevance*, an environmental scan was conducted to understand stakeholder utilisation of indicators. A total of 232 instances online were recorded. The top five indicators were hospital standardised mortality ratio (HSMR) (n=23), 30-day overall readmission (n=18), wait times for hip fracture repair (n=17), ambulatory care sensitive conditions (n=14) and caesarean section rate (n=13).

Detailed statements on the *actionability* of each indicator were provided to evaluation participants. Specifically, summations on the purpose of indicator, strengths, caveats and scientific evidence in support were considered.

To measure the degree of *stakeholder follow-up*, we reviewed all instances of patient-level data requests from providers. In 2013–2014, there were 298 requests, with 11 facility-level readmission indicators accounting for 58% of all requests (n=173).

The criterion *sufficient volumes* quantifies the proportion of indicator results that are suppressed per CIHI's data privacy protocols. In general, indicator results with cell counts <5 are suppressed, and results based on <50 denominator cases per hospital are flagged as low volume and unstable rates. Facility-level indicators are particularly affected by low volumes and suppressed results: 23 of 27 facility-level indicators had at least one-fifth of all results flagged as low volume. A further seven of these indicators had at least one-fifth of all results suppressed due to small cell counts. At the extreme, we note the 28-day readmission after stroke and acute myocardial infarction (AMI) indicators with ~75% low-volume rates and one-third of all results suppressed.

We performed *significance of variation* analysis to determine the variability within indicator results. For example, the hip fracture surgical procedures performed within 48 hours indicators (both within one and across facilities) had the lowest relative SD values of 16% and 17%, respectively, indicating minimal differences across indicator results.

## Scientific soundness

The criterion *data quality* garnered the greatest discussion during Delphi reviews. Limitations of using administrative data were considered. Examples of concern include the inability to assess indications for angiography for AMI patients for the indicator use of coronary angiography following AMI, and the ability to properly identify denominator cases for the hysterectomy indicator.

The evaluation revealed that *validity reviews* were performed for each indicator on an annual basis. These included significance testing of risk factors, monitoring of diagnosis and procedure coding updates, and outlier and significant change analyses. Indicators recommended for further consultation and R&D were identified as such mainly for the purpose of seeking feedback from stakeholders on the validity and clinical relevance of current calculation methodologies.

The criterion *participation bias* assessed whether data submission and participation in the calculation of indicator results were a nature of voluntary participation. All but two indicators—physician specialists and general/family physicians per 100 000 population—required mandatory participation. In other instances, such as indicators produced for long-term care facilities, participation is not yet mandatory across the country, and therefore, results published may contain a participation bias.

### Feasibility

*Production cost* was considered based on the extent of staff resources required to produce each indicator. Indicators with complex linkages across multiple databases and those requiring building of episodes of care necessitate a larger degree of resources.

### Modified Delphi sessions of CIHI technical and leadership groups

The mean Likert scores, recommendations and rationale are noted in tables 2–4. Nine indicators were recommended as candidates for retirement (table 3), seven were identified as requiring additional consultation and three were recommended to undergo further redevelopment (table 2). Thirty-five indicators were recommended for retention (table 4). The rationale to retain these HSP indicators was based on the assessment of all 18 evaluation criteria as a whole. Although retained indicators correlate strongly with high mean Likert scores, this was only one contributor to the recommendation. Ultimately, the discussion during the inperson Delphi sessions allowed for the most pertinent and informative of the 18 evaluation criteria to be considered above others.

CIHI leadership and technical groups identified indicators for additional consultation and redevelopment. These indicator recommendations were not forwarded to Consensus Conference participants, but were instead identified for internal R&D efforts in the interim.

### Pre-Consensus Conference survey

Forty-eight Consensus Conference participants completed the online survey (response rate of 79%). Eighty-five per cent of conference participants had more than 10 years of healthcare experience. Geographic distribution of respondents correlated well with Canada's population across provinces/territories. Stakeholders from federal and provincial government agencies accounted for three-quarters of survey respondents, followed by regional health authority executives, hospital administrators and academic/research funding organisations. The mean survey agreement score (as a percentage of responses) for all nine indicators proposed for retirement was 70%, and was used as a benchmark for automatic ratification. The option to select No opinion for each indicator under survey accounted for an average of 20% of responses (ranging between 12% and 30% across indicators); such an option was made available in the event that respondents held insufficient knowledge on the indicator or did not utilise the indicator within their setting; a response of Agreed, Disagreed or No opinion was mandatory in the survey.

### National Consensus Conference

Of the nine indicators recommended for retirement, six received more than 70% agreement as a proportion of responses in the preconference survey, and therefore were automatically accepted for retirement (table 3). The remaining three indicators were discussed as a group, and subsequently also ratified for retirement by conference participants. The majority of indicators recommended for retirement were condition-specific readmission indicators. Ultimately, the decision to retire these indicators was based on appropriateness for continued public reporting. While these indicators were

**Table 2** Indicators identified for additional consultation and further redevelopment

| Type | Indicator | Mean Likert score | Rationale |
|---|---|---|---|
| **Additional consultation** | | | |
| Region | Hip replacement | 5.0 | There are concerns of utility and actionability for these indicators as they represent procedure counts per population. |
| | Knee replacement | 4.8 | |
| | Coronary artery bypass graft (CABG) | 6.6 | |
| | Percutaneous coronary intervention (PCI) | 6.6 | |
| | Cardiac revascularisation | 6.6 | |
| Facility | Vaginal birth after caesarean section | 4.4 | There are concerns of validity and utility for these indicators. |
| | Birth trauma | 5.4 | |
| **Further redevelopment** | | | |
| Region | Hysterectomy | 4.4 | R&D is required to improve identification of appropriate denominator cases. |
| Facility | Nursing sensitive adverse events for medical patients | 6.8 | There is an opportunity for incorporation within newly developed hospital harm indicator. |
| | Nursing sensitive adverse events for surgical patients | 6.8 | |

Mean Likert Scale Score: 7–9, robust indicator, recommending continued reporting; 4–6, equivocal indicator, further discussion at inperson Delphi session required; 1–3, weak indicator, recommending indicator retirement.

**Table 3** Indicators recommended for retirement

| Type | Indicator | Mean Likert score | Rationale | Pre-Consensus Conference Survey Agreement for retirement (as a % of responses) |
|---|---|---|---|---|
| Facility | 28-day readmission after prostatectomy | 5.2 | These indicators have low volumes of cases leading to unstable rates as well as to the suppression of a large number of results for public reporting. Furthermore, these cases are included in the surgical/medical readmission indicators, and can still be derived through private reporting tools. | 82%* |
|  | 28-day readmission after hysterectomy | 5.6 |  | 80%* |
|  | 90-day readmission after knee replacement | 6.4 |  | 73%* |
|  | 90-day readmission after hip replacement | 6.4 |  | 72%* |
|  | 28-day readmission after stroke | 6.2 |  | 58% |
|  | Use of coronary angiography following AMI | 6.4 | Angiography may not be indicated for every AMI patient, depending on his or her clinical history, and the clinical appropriateness of angiography is difficult to ascertain from the administrative hospitalisation data. Therefore, it is challenging to interpret and compare the results for this indicator. | 78%* |
|  | Hip fracture surgical procedures performed within one facility (48 hours) | 6.4 | This indicator does not measure the true proportion of surgeries performed within 48 hours of admission to an acute care hospital, since it does not account for transfers across hospitals. Many patients are transferred from their initial admitting acute care facility to another facility for surgery. The indicator hip fracture surgical procedures performed within 48 hours, which measures total time across all acute care facilities, will continue to be produced and reported on. | 72%* |
|  | 28-day readmission after AMI | 6.4 | Concerns have been raised regarding hospitals' ability to take action on this indicator. It is felt that with the regionalisation of cardiac care, it is more appropriate to measure readmission after AMI at the regional level (by patient residence) than at the hospital level. In addition, having a low volume of cases leads to unstable rates and to the suppression of a large number of results for public reporting. Therefore, it was proposed to keep the Readmission after AMI indicator at the regional level and to retire the facility-level indicator. Furthermore, readmissions after AMI are included in the 30-day overall readmission indicator at the facility level. | 59% |
|  | Primary caesarean section rate | 4.6 | A new indicator (low-risk caesarean section) measures the rate of deliveries via caesarean section among singleton term cephalic pregnancies for women without placenta previa or previous C-section. Since this new indicator is limited to women who have not had a previous C-section, it can take the place of primary caesarean section rate and be a better indicator of appropriateness. | 57% |

Mean Likert Scale Score: 7–9, robust indicator, recommending continued reporting; 4–6, equivocal indicator, further discussion at inperson Delphi session required; 1–3, weak indicator, recommending indicator retirement.

*Passing the threshold (of 70% agreement among responses) for automatic ratification.

Table 4 Indicators retained

| Type | Indicator | Mean Likert score |
|---|---|---|
| Region | 30-day AMI inhospital mortality | 8.8 |
| | 30-day stroke inhospital mortality | 8.8 |
| | Hospital standardised mortality ratio (HSMR) | 8.8 |
| | Ambulatory care sensitive conditions | 8.6 |
| | Wait times for hip fracture repair | 8.4 |
| | 30-day readmission for mental illness | 7.8 |
| | Repeat hospital stays for mental illness | 7.8 |
| | Self-injury hospitalisation | 7.6 |
| | 30-day AMI readmission | 7.4 |
| | Hospitalised hip fracture event | 7.2 |
| | Hospitalised strokes | 7.2 |
| | Hospitalised AMI event | 7.0 |
| | Inflow/outflow ratio | 7.0 |
| | 30-day readmission: patients age 19 and younger | 6.8 |
| | 30-day obstetric readmission | 6.8 |
| | 30-day medical readmission | 6.8 |
| | 30-day surgical readmission | 6.4 |
| | Mental illness patient days | 6.2 |
| | Mental illness hospitalisation | 6.0 |
| | Injury hospitalisation | 5.4 |
| | Caesarean section rate | 4.8 |
| Facility | 30-day AMI inhospital mortality | 8.8 |
| | 30-day stroke inhospital mortality | 8.6 |
| | Hip fracture surgery within 48 hours | 8.4 |
| | 30-day overall readmission | 8.0 |
| | 30-day inhospital mortality following major surgery | 8.0 |
| | 30-day readmission: patients age 19 and younger | 7.8 |
| | 30-day obstetric readmission | 7.8 |
| | 30-day medical readmission | 7.6 |
| | 30-day surgical readmission | 7.4 |
| | Inhospital hip fracture in elderly (age 65+) patients | 7.4 |
| | Obstetric trauma—vaginal delivery with instrument | 7.4 |
| | Obstetric trauma—vaginal delivery without instrument | 7.4 |
| | Caesarean section rate | 6.8 |
| | Low-risk caesarean section | 6.8 |

Mean Likert Scale Score: 7–9, robust indicator, recommending continued reporting; 4–6, equivocal indicator, further discussion at inperson Delphi session required; 1–3, weak indicator, recommending indicator retirement.

ratified for retirement over concerns of rate stability and small numbers, facilities can continue to calculate and monitor these indicators through CIHI private reporting tools. Consensus on retiring these indicators was achieved with greater ease, given that a provider's capacity to continue to privately monitor performance would be maintained.

Two contextual health human resources indicators at the regional level—physician specialists and general/family physicians per 100 000 population—were also included in the modified Delphi review process, and rated low in Likert Scale scoring (both received a mean score of 3.2). While these indicators provide some context on HSP characteristics, they are already reported elsewhere within CIHI. It was agreed to continue reporting on these indicators but outside of the HSP framework.

Table 4 lists 35 HSP indicators retained for continued public reporting. Although retained indicators correlate strongly with high mean Likert scores, this was only one contributor to the recommendation. For example, the regional level caesarean section rate indicator received a mean Likert score of 4.8 from the technical group, but was retained for public reporting after discussion by the leadership group due to continued concerns over high rates in Canada and therefore, a need for continued monitoring.

## DISCUSSION

This exercise proved to be an informative, objective, systematic, transparent, inclusive and likely repeatable process for evaluating and reconfirming a national set of HSP indicators. Overall, the approach of using 18 subcriteria was manageable and informative, with feedback from participants that the added information and context made it easier to make a final disposition recommendation for each indicator. The overall timeline of the evaluation process from inception to completion was

18 months. Three distinct phases stand out, each requiring ~6 months to complete: initial R&D of the evaluation plan, executing the evaluation internally at CIHI and achieving consensus across stakeholders.

An initial Likert score of indicators provides a baseline to proceed with group Delphi reviews. We found it beneficial to begin with the lowest scores and work our way to the highest rated indicators. We also found it operational to have our technical group first review indicators and to pass on recommendations to a leadership group that would consider these in addition to their knowledge and understanding of the use of HSP information in the field. The iterative process of having participants first review indicators independent of other Delphi members and to then convene as a group to discuss findings allowed for a balanced and participatory discussion among participants. These iterative methods ensured a summative process whereby findings were transparent and confirmed at each stage.

The national Consensus Conference provided an opportunity to pilot-test the results of a rigorous, mostly internal methodology for evaluating indicators produced by CIHI. Most recently, CIHI has been incorporating the learnings from this exercise into a broader 'lifecycle' approach to indicator development, evaluation and retirement recognising that all too often there is a tendency to add new indicators to the suite of those reported paying little attention to the utility of those reported in some instances for years. The internally developed evaluation process including the 18 criteria used for assessing previously reported indicators will also lend itself to midcycle reviews of suites of indicators that could be modified for such a process. The ability to affirm our internal process with external stakeholders at a national conference provided further confidence in the process. And, while stakeholders appreciated the opportunity to review and ratify our findings, going forward, they expressed comfort with CIHI implementing a systematic evaluation of the indicators and making decisions about reporting. There was congruence in opinion on the suitability of HSP indicators for public reporting throughout the evaluation process, beginning with Likert scores and assessments from CIHI technical staff, to CIHI leadership, and finally with stakeholders.

## Strengths and weaknesses of the study

We recognise that the overall evaluation process required considerable time and resources, there are important benefits to such a comprehensive approach. For example, we ensured a transparent and sequential evaluation, whereby discourse and findings were accumulated and presented in a summarised manner at each phase. We solicited feedback from a wide array of expertise including those responsible for monitoring the results of these indicators on a regular basis. An external moderator facilitating the discussion ensured independence during the consensus process. These processes have been described as favourable conceptual approaches to aid exercises of indicator development, maintenance and evaluation.[14]

One main weakness of this process was the lack of involvement of the 'patient/public' voice in evaluating the utility of CIHI's current suite of publicly reported HSP indicators. Traditionally, the approach to HSP reporting has largely been targeted to system decision makers. With the growing recognition that HSP includes measuring things that are important and relevant to the patient/public, it is clear that the patient/public perspective needs to be embedded in future aspects of this work. In 2013, CIHI solicited input from 3000 Canadians (randomised, representative sample) through small group dialogues and online questionnaires about which types of indicators and domains of HSP they would like to see publicly reported. In an attempt to obtain broader input to the evaluation process discussed in this paper, the same survey sent to Consensus Conference participants was made available on CIHI's website for public participation. The survey responses from the general public were highlighted and considered at the Consensus Conference. However, a more systematic approach to including the patient/public perspective within the 'lifecycle' approach to development, evaluation and retirement is needed to going forward.

Shekelle[16] notes that there is little agreement on methodologies for developing performance indicators, and this can also be said regarding their evaluation. Nonetheless, Stelfox and Straus[14] emphasise the importance of clearly establishing the chosen evaluation criteria in advance of launching a consensus process. In the majority of the studies we reviewed and cite, a smaller number of evaluation criteria were applied: most often, usability, importance, scientific soundness and feasibility (or a variation thereof that drew on similar domains). Conversely, we found it helpful to apply multiple subcriteria to comprehensively reflect the evaluation of indicators for their suitability of ongoing public reporting. Furthermore, providing a more granular evaluation schema for participants ensured more consistent definitions of domains and structured evidence/results for evaluators' consideration. Nonetheless, while these evaluation criteria were informative and applicable to this precise context, not all would apply for other evaluation purposes. Further efforts are necessary to determine the level of customisation required to ensure that the process and criteria are applicable to other sectors of care and types of indicators.

In addition to convening an inperson consensus conference (or expert panel) to evaluate indicators, Santana *et al*[17] forwarded their evaluation survey to 101 trauma centres across 4 countries involved in the use and assessment of injury care indicators. Moreover, a novel subsequent process has been described by Bobrovitz *et al*[18] whereby the discussion occurring throughout the consensus conference is transcribed and undergoes qualitative content analysis to identify key themes raised throughout the deliberations. These additional activities can provide complementary evidence to the evaluation process, such as qualitative findings to an otherwise objective and quantitative exercise, and reaching a broader group of stakeholders and users of health measures.

There are certain characteristics of the Canadian healthcare system that are favourable for such an evaluation exercise. As the national healthcare system information steward, CIHI receives data for virtually all hospitalisations across the country in a standardised manner. All but 2 of the 56 HSP indicators are calculated using this standardised data source. Therefore, the application of 18 evaluation criteria to these indicators can be performed so in a systematic process, so that objectivity is maintained. A centralised healthcare information system is more conducive for cross-country analysis and reporting.[19] This also extends to the convening strength of CIHI to bring together stakeholders from all provinces and territories to agree on a national agenda.

To balance the limiting aspects of a Delphi exercise on a set of existing indicators, the Consensus Conference also included working group sessions on identifying priority areas for future indicator development (organised by health system quadrants of Inputs and Characteristics, Outputs, Outcomes and Social Determinants of Health). From these discussions, along with a cross-country consultation process, CIHI has embarked on a path to develop new indicators for the domains of safety (eg, infections), mental health and addictions (alcohol attributable hospitalisations), and others relating to recently identified priority populations such as seniors and ageing (eg, palliative care), and children and youth.

## CONCLUSION

The proliferation of health measures required to fulfil reporting gaps occurred with minimal consideration to alignment and utility with pre-existing indicators. Not surprisingly, then, stakeholders were overwhelmingly in favour of implementing a process that would result in a leaner, more applicable suite of HSP indicators.

CIHI will gradually expand this evaluation methodology to applicable sectors of care. We will also continue to work with external partners to reduce indicator chaos and increase alignment with reporting requirements across the country.[6]

This exercise generated identified analytical alignment actions that can be taken at CIHI throughout indicator production and maintenance with a view to reduce indicator chaos. Furthermore, we gained new knowledge about how the HSP indicators we produce are used by stakeholders through an internet-based environmental scan and via discussions held at the Consensus Conference.[10]

In line with established practices of convening a Consensus Conference every 5 years, we feel that it is highly beneficial to inform those discussions with a wholesale and systematic criteria-based review of indicators just prior. A broad consultation process encompassing diverse public health stakeholders from across the country helps ensure the development and use of indicators most appropriately reflecting the health of populations and the performance of health systems.[20] Similarly, a retrospective exercise on national HSP practices can identify important lessons, of which the selection of indicators suitable for public reporting is an integral component.[21]

**Contributors** OF participated in the design of the evaluation, performed data analysis, facilitated internal Delphi sessions and drafted the manuscript. KL participated in the design of the evaluation, facilitated internal Delphi sessions, organised the national consensus conference and drafted the manuscript. YG participated in the design of the evaluation, facilitated internal Delphi sessions and drafted the manuscript.

**Funding** This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

**Competing interests** None declared.

**Ethics approval** Ethics approval was not required as the study did not utilise any sensitive data.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data sharing statement** No additional data are available.

## REFERENCES

1. Klazinga N, Fischer C, ten Asbroek A. Health services research related to performance indicators and benchmarking in Europe. *J Health Serv Res Policy* 2011:16:38–47. http://hsr.sagepub.com/content/16/suppl_2/38.long (accessed 16 Oct 2016).
2. Kelley E, Arispe I, Homes J. Beyond the initial indicators: lessons from the OECD Health Care Quality Indicators Project and the US National Healthcare Quality Report. *Int J Qual Health Care* 2006;18 (Suppl 1):45–51. http://intqhc.oxfordjournals.org/content/18/suppl_1/45.long (accessed 16 Oct 2016).
3. Carinci F, Van Gool K, Mainz J, *et al*. Towards actionable international comparisons of health system performance: expert revision of the OECD framework and quality indicators. *Int J Qual Health Care* 2015;27:137–46. http://intqhc.oxfordjournals.org/content/27/2/137 (accessed 16 Oct 2016).
4. Canadian Institute for Health Information. *Roadmap initiative … launching the process*. Ottawa. 2000. https://www.cihi.ca/en/profile_roadmap_launch_pdf_en.pdf (accessed 16 Oct 2016).
5. Canadian Institute for Health Information. *A performance measurement framework for the Canadian health system*. Ottawa. 2013. https://secure.cihi.ca/free_products/HSP-Framework-ENweb.pdf (accessed 16 Oct 2016).
6. Saskatchewan Health Quality Council. *Think big, start small, act now: tackling indicator chaos: a report on a national summit*. Saskatoon, 30–31 May 2011. http://hqc.sk.ca/Portals/0/documents/tracking-indicator-choas.pdf (accessed 16 Oct 2016).
7. Canadian Institute for Health Information. *National consensus conference on population health indicators*. Ottawa. 1999. https://secure.cihi.ca/free_products/phi.pdf (accessed 16 Oct 2016).
8. Canadian Institute for Health Information. The health indicators project: the next 5 years. Report from the Second Consensus Conference on Population Health Indicators. Ottawa. 2005. https://secure.cihi.ca/free_products/Consensus_Conference_e.pdf (accessed 16 Oct 2016).

9. Canadian Institute for Health Information. *The Health Indicators Project: Report from the Third Consensus Conference on Health Indicators*. Ottawa. 2009. https://secure.cihi.ca/free_products/82-230-XWE_e.PDF (accessed 16 Oct 2016).
10. Canadian Institute for Health Information. *Rethink, Renew, Retire: Report from the Fourth Consensus Conference on Evaluating Priorities for Canada's Health Indicators*. Ottawa. 2015. https://secure.cihi.ca/free_products/Rethink_Renew_Retire.pdf (accessed 16 Oct 2016).
11. Hurtado M, Swift E, Corrigan J, eds. Committee on the National Quality Report on Health Care Delivery, Board on Health Care Services. Envisioning the National Health Care Quality Report. Washington, DC: National Academy Press, 2001. http://www.ncbi.nlm.nih.gov/books/NBK223318/pdf/Bookshelf_NBK223318.pdf (accessed 16 Oct 2016).
12. Mattke S. When should measures be updated? Development of a conceptual framework for maintenance of quality-of-care measures. *Qual Saf Health Care* 2008;17:182–6. (accessed 16 Oct 2016).
13. Canadian Institute for Health Information. *The CIHI Data Quality Framework, 2009*. Ottawa. 2009. https://www.cihi.ca/en/data_quality_framework_2009_en.pdf (accessed 16 Oct 2016).
14. Stelfox H, Straus S. Measuring quality of care: considering conceptual approaches to quality indicator development and evaluation. *J Clin Epidemiol* 2013;66:1328–37.
15. Fitch K, Bernstein S, Aguilar M, *et al*. *The RAND/UCLA appropriateness method user's manual*. Santa Monica, CA: RAND Corporation, 2001. http://www.rand.org/pubs/monograph_reports/MR1269 (accessed 16 Oct 2016).
16. Shekelle P. Quality indicators and performance measures: methods for development need more standardization. *J Clin Epidemiol* 2013;66:1338–9.
17. Santana MJ, Stelfox HT, Asbridge M, *et al*. Development and evaluation of evidence-informed quality indicators for adult injury care. *Ann Surg* 2014;259:186–92.
18. Bobrovitz N, Parrilla JS, Santana M, *et al.* A qualitative analysis of a consensus process to develop quality indicators of injury care. *Implement Sci* 2013;8:45.
19. OECD. Strengthening Health Information Infrastructure for Health Care Quality Governance: Good Practices, New Opportunities and Data Privacy Protection Challenges. Paris. 2013. (accessed 12 Dec 2016).
20. Klazinga N, Stronks K, Delnoij D, *et al*. Indicators without a cause. Reflections on the development and use of indicators in health care from a public health perspective. *Int J Qual Health Care* 2001;13:433–8. http://intqhc.oxfordjournals.org/content/intqhc/13/6/433.full.pdf (accessed 16 Oct 2016).
21. van den Berg MJ, Kringos DS, Marks LK, *et al*. The Dutch health care performance report: seven years of health care performance assessment in the Netherlands. *Health Res Policy Syst* 2014;12:1.