Should indicators be correlated? Formative indicators for healthcare quality measurement

Iris Blotenberg, Laura Schang, Dennis Boywitt

Department of Methodology, Institut für Qualitätssicherung und Transparenz im Gesundheitswesen, Berlin, Germany

Correspondence to
Dr Iris Blotenberg;
blotenberg.iris@gmail.com

INTRODUCTION

Unwarranted variation in the quality of care remains a pressing issue for health systems globally.[1] In this context, there is a growing need for well-grounded quality indicators that help inform decisions by patients (eg, when choosing a provider), policy makers (eg, for accountability purposes and quality-based payment) and providers (eg, to stimulate benchmarking and improvement).[2] However, studies repeatedly report low or no correlations across healthcare quality indicators[3–5] leading to seemingly paradoxical results when different indicators are used for constructing league tables.[6 7] Is this result surprising or is it to be expected?

Quality indicators can either be conceptualised as reflective (also called the 'psychometric' approach to measurement) or as formative (also called the 'clinimetric' approach to measurement).[8 9] With reflective indicators, the indicator values are tacitly assumed to be reflections of quality as a single underlying property of a provider—in this case, correlations between indicators are expected. In contrast, with formative indicators, indicator values are understood as representations of different aspects of quality which, when taken together, form a provider's quality of care profile.

While the distinction between formative and reflective indicators has attracted much debate in psychometrics,[8 10] in health services research, assumptions about measurement have not been discussed as extensively as necessary. So far, discussions have been limited to specific fields, for example, composite indicators of healthcare quality[11] or patient-reported experience measures.[12] The underlying issue and the broader implications for quality measurement have, however, never been brought together.

In this paper, we argue that the distinction between reflective and formative indicators has important implications for the selection and interpretation of quality measures. If quality indicators are wrongly conceptualised as reflective although they are actually formative, valid indicators might be discarded[8 13] or conclusions about a provider's quality of care may be misguided.[14] Next, we highlight key arguments why quality indicators should typically be conceptualised as formative. We then bring together practical implications of formative quality indicators for indicator selection, rankings based on composite indicators, different methods for quality measurement and the use of tracer conditions.

THE CASE FOR FORMATIVE QUALITY INDICATORS

The reflective approach to measurement assumes that indicator values reflect a common, underlying property (eg, of an individual or an organisation) and should therefore be highly correlated.[8 9] This is often the case in psychometrics, for example, when a person's general intelligence determines how well she or he does in various cognitive tests (eg, reasoning, knowledge, working memory).[15] In contrast, with the formative approach to measurement, the goal is to measure a combination of features and thus the indicators need not be correlated.[8 9] A typical example is the Apgar score, an assessment of the state of a newborn infant based on five components (respiratory effort, heart rate, muscle tone, reflexes, colour).[16] Whether or not these components are correlated has no bearing on the usefulness of the Apgar score. Figure 1 illustrates the two different approaches to measurement.

In many cases, quality indicators are formative: they do not reflect a single underlying property but rather different aspects of quality, which do not necessarily have a common cause and thus are not expected to be correlated.[3 11 12] For example, quality of care frameworks are typically multidimensional, distinguishing between, for example,
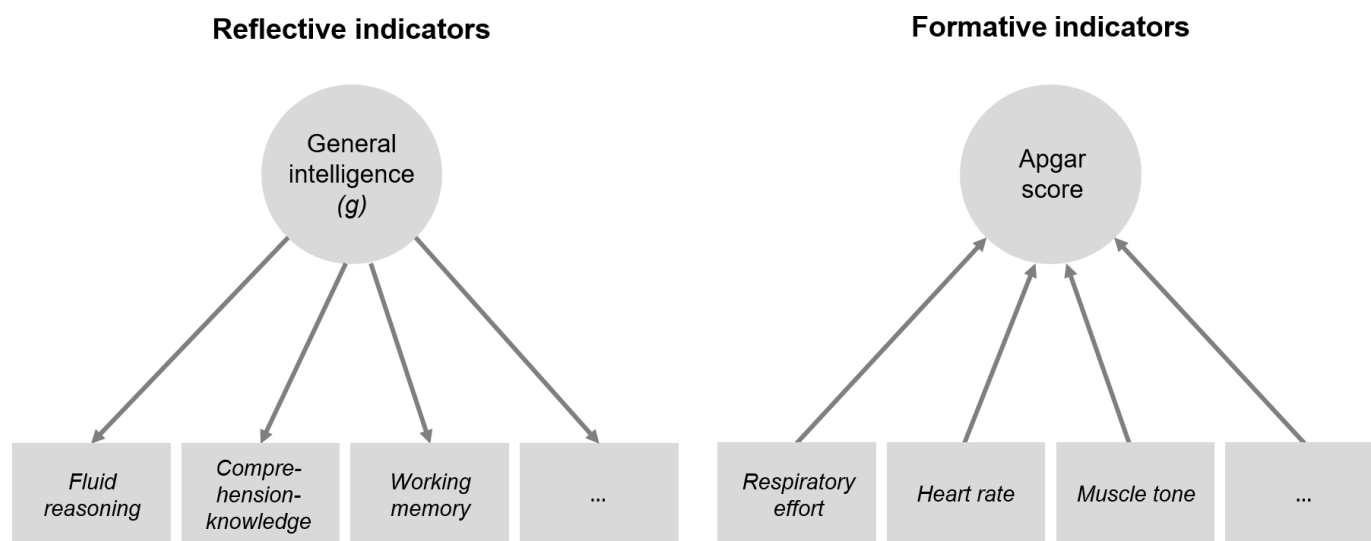
Figure 1 Two fundamental concepts for measurements. Reflective indicators (left) reflect an underlying property of an individual or an organisation. For example, different indicators of cognitive ability (reasoning, knowledge, working memory) reflect an individual's general intelligence. Formative indicators (right) measure a combination of features of a phenomenon of interest. For example, the Apgar score summarises five components of a newborn infant's health (respiratory effort, heart rate, muscle tone, reflexes, colour).

the effectiveness, safety and patient-centredness of care.[2] A provider that achieved desirable patient outcomes (effectiveness) may not necessarily have provided care in a responsive and respectful manner (patient-centredness). Moreover, even within the same quality dimension, a provider may do well in one aspect of care, but less so in another. For instance, a patient may have been provided with information about different treatment options (one aspect of patient-centredness), but the staff may not have answered his or her questions clearly (another aspect of patient-centredness).

Evidence about (the strength of) correlations between different quality measures is inconsistent. While some studies report significant correlations between indicators,[17] others report low or no correlations across pairs of measures.[3–5] This suggests that quality of care should not be seen as a single underlying property of a provider but rather as a combination of various quality aspects that should consequently be measured using formative indicators.

FORMATIVE INDICATORS FOR QUALITY MEASUREMENT: PRACTICAL IMPLICATIONS
Indicator selection
Depending on whether indicators are conceptualised as formative or reflective, different criteria are appropriate for indicator selection. Quantitative psychometric methods (eg, item–total correlations, internal consistency, factor analysis) have repeatedly been used to evaluate quality indicator sets, especially for patient-reported indicators.[12 13 18] However, quantitative psychometric methods would only be appropriate with reflective indicators,[8 9 12] thus implicitly assuming that quality of care

is a single underlying property of a provider. As we have argued above, this assumption seems questionable.

If quantitative psychometric methods of indicator selection are applied although the indicators are formative, important aspects of quality of care could be omitted.[8 13] In a development of an asthma quality of life questionnaire,[19] for instance, the authors contrasted two different selection methods: first, item selection based on patients' opinions and, second, item selection based on quantitative psychometric properties. While 20 items were common to both selection methods, several items that mattered highly to patients would have been excluded if item selection had solely been based on quantitative psychometric criteria.

With formative indicators, indicator selection should be guided by the criterion of content validity, which is defined as the degree to which indicators represent all relevant aspects of the targeted quality topic properly.[13 20] A content-valid indicator set enables users of measurement results such as patients, providers and policy makers to draw appropriate conclusions about a provider's quality with regard to the measured quality topic. To assure content validity of an indicator set, it is important to clearly specify the relevant content domains and to select indicators commensurate with these domains by involving stakeholders.[20]

Rankings based on composite measures of quality
The issue of formative versus reflective indicators also has implications for the construction of composite measures that aggregate multiple indicators into a summary measure. In many countries, they are widely used to provide an overview of performance and thus facilitate public reporting and comparisons between providers

(eg, in the form of rankings or league tables).[21 22] To ensure fair comparisons, it is essential to compare 'like with like'. However, until 2021, the Centers for Medicare & Medicaid Services (CMS) star rating used to compare hospitals on very different numbers and different types of indicators.[6] The star quality rating system introduced by the CMS provides a single public measure of hospital quality. For this measure, various quality indicators are aggregated into one measure per hospital, which is displayed in terms of one to five stars. However, such comparisons of providers based on different sets of indicators would only be fair if each indicator measured the same underlying property overall—the core assumption of reflective indicators.

In fact, however, for CMS star ratings, analyses have shown that using different sets of indicators systematically disadvantages some groups of hospitals (eg, the more measures a hospital reported, the less likely it was to obtain five stars[14]). In Germany, empirical comparisons of different systems of hospital rankings showed that changing the indicator set can result in considerable shifts in relative rankings of hospitals, even from the top to the bottom half of the group or vice versa.[7] Moreover, conclusions based on single indicators need not match conclusions based on composite indicators. In fact, research shows low correlations between composite indicator value and component indicators of process of care, readmissions, mortality, efficiency and patient satisfaction.[3 11]

These results are again well in line with the idea that quality of care is multidimensional and that quality indicators should be understood as formative. With formative indicators, measurement results based on different (sets of) indicators are not expected to align. In other words, different indicator sets should not be expected to produce the same '(quality) signal'. Thus, with formative indicators, comparisons should be based on the same set of indicators so as to be fair.

Different methods, different results?

A formative perspective on quality measurement may also shed some light on why different methods for quality appraisal may produce different results. A good example is the Care Quality Commission's Intelligent Monitoring (IM) tool in the UK, which was devised to prioritise quality inspections in hospitals which were expected to perform poorly based on quantitative risk scores. Research found, however, that the risk scores produced by the IM tool failed to predict the inspection-based quality ratings.[23]

While the authors discuss several explanations for this gap—the IM tool might be too simplistic, too coarse or the inspections might be unreliable[23]—there may also be another explanation: quality was (implicitly) conceptualised as reflective although it is actually formative. The use of quantitative risk scores to prioritise hospitals for inspection assumes a single underlying 'hospital quality' that will be reflected in both the risk scores and the inspection-based quality ratings—in other words, assuming reflective indicators.

However, did the inspections and the IM tool actually measure the same things? The risk score was based on about 150 indicators that were intended to represent key risks to the quality of care, covering among others mortality rates, waiting times and whistleblower reports. The inspection teams, on the other hand, rated various individual hospital services against five 'key questions' (Is the service 'safe', 'effective', 'caring', 'responsive to people's needs' and 'well led'?). These service-level ratings were then aggregated to assign hospital-level ratings.[23] Thus, both the risk score and the inspection-based rates are in fact complex, heterogeneous composite indicators. Since the components of these composite indicators differed, it is not surprising—from a formative perspective on measurement—that the two methods 'formed' different pictures of quality. We suggest that the more similar the criteria underlying these two composite indicators are, the higher they will be correlated.

Tracer conditions: reflective indicators by another name?

In analysing healthcare delivery, Kessner *et al*'s 'tracer concept' has been influential: the premise is that a selected set of health problems could serve as 'tracers' of the general quality of care[24] and thus enable profiling the strengths and weakness of services delivery.[25] According to Kessner, tracers should be chosen based on specific criteria, such as that they are well defined and easy to diagnose, have high prevalence rates and represent a cross section of patient age and sex groups and a variety of medical care activities.[24] Although studies suggested caution in extrapolating from the management of tracer problems to the overall quality of care,[26] the 'tracer' concept is still widely used.[25 27]

So under which conditions is extrapolating from the results of a set of indicators to quality of care in other areas warranted? Only under the assumption of reflective indicators. Inferences from the measured to unmeasured health conditions would only be logically warranted, if quality of care were delivered equally across all tracer conditions (ie, has a single underlying cause).[28] Again, it seems doubtful that quality of care is a single property underlying the tracers and non-tracers. In fact, a provider might focus particularly on tracer diagnoses, neglecting other conditions.[28] Accordingly, the strong assumption of a common underlying cause that drives the management of all health conditions in the same fashion for all tracers appears logically and empirically questionable. Thus, we caution from extrapolating conclusions from a given set of indicators to other unmeasured aspects of quality of care.

Of course, prioritisation of conditions for which quality measures should be implemented is always necessary as not all aspects of healthcare delivery can and need be measured with limited resources. We suggest that the selection of conditions should be based on explicit criteria to ensure that the construct(s) of interest are

actually targeted. Importantly, quality of care in these prioritised health conditions cannot be taken as an indication of quality of the entire provider or of the entire health system.

CONCLUSIONS

Underlying assumptions of quality measurement and their consequences have not been discussed as extensively as necessary. Decisions on whether quality measures are conceptualised as formative or reflective have important implications for the selection and interpretation of quality indicators. Because there is little reason to believe in one underlying cause affecting the results of all indicators of a set and because different quality indicators are typically not highly correlated, we suggest that formative indicators as measures of quality of care are frequently appropriate. Developers of quality indicators should thus select indicators primarily based on how well they represent the targeted quality topic (ie, ensure high content validity of the indicator set) rather than with emphasis on psychometric properties. Also, providers, patients and policy makers and other users of quality measures should be aware that different indicator sets, different measurement methods and different assessment criteria are likely to show different pictures of quality rather than converging on a single underlying 'quality signal'— different criteria are bound to yield different results.

REFERENCES

1 Corallo AN, Croxford R, Goodman DC, *et al*. A systematic review of medical practice variation in OECD countries. *Health Policy* 2014;114:5–14.
2 Arah OA, Westert GP, Hurst J, *et al*. A conceptual framework for the OECD health care quality indicators project. *Int J Qual Health Care* 2006;18 Suppl 1:5–13.
3 Shwartz M, Cohen AB, Restuccia JD, *et al*. How well can we identify the high-performing hospital? *Med Care Res Rev* 2011;68:290–310.
4 Gooiker GA, Kolfschoten NE, Bastiaannet E, *et al*. Evaluating the validity of quality indicators for colorectal cancer care. *J Surg Oncol* 2013;108:465–71.
5 Bradley EH, Herrin J, Elbel B, *et al*. Hospital quality for acute myocardial infarction: correlation among process measures and relationship with short-term mortality. *JAMA* 2006;296:72–8.
6 Bilimoria KY, Barnard C. An evolving Hospital quality StAR rating system from CMS: aligning the stars. *JAMA* 2021;325:2151–2.
7 Stausberg J, Halim A, Färber R. Concordance and robustness of quality indicator sets for hospitals: an analysis of routine data. *BMC Health Serv Res* 2011;11:106.
8 Bollen K, Lennox R. Conventional wisdom on measurement: a structural equation perspective. *Psychol Bull* 1991;110:305–14.
9 van Amelsvoort LGPM, Fleuren BPI, Kant IJmert. Improving measurement models in clinical epidemiology: time to move beyond the inherent assumption of an underlying reflective measurement model. *J Clin Epidemiol* 2020;118:119–23.
10 Bollen KA, Diamantopoulos A. In defense of causal-formative indicators: a minority report. *Psychol Methods* 2017;22:581–96.
11 Shwartz M, Restuccia JD, Rosen AK. Composite measures of health care provider performance: a description of approaches. *Milbank Q* 2015;93:788–825.
12 Sizmur S, Graham C, Bos N. Psychometric evaluation of patient-reported experience measures: is it valid? *Int J Qual Health Care* 2020;32:219–20.
13 Fayers PM, Hand DJ. Causal variables, indicator variables and measurement scales: an example from quality of life. *J R Stat Soc Ser A Stat Soc* 2002;165:233–53.
14 DeLancey JO, Softcheck J, Chung JW, *et al*. Associations Between Hospital Characteristics, Measure Reporting, and the Centers for Medicare & Medicaid Services Overall Hospital Quality Star Ratings. *JAMA* 2017;317:2015–7.
15 McGrew KS. CHC theory and the human cognitive abilities project: standing on the shoulders of the giants of psychometric intelligence research. *Intelligence* 2009;37:1–10.
16 Apgar V. A proposal for a new method of evaluation of the newborn infant. *Curr Res Anesth Analg* 1953;32:260–7.
17 Doyle C, Lennox L, Bell D. A systematic review of evidence on the links between patient experience and clinical safety and effectiveness. *BMJ Open* 2013;3:e001570.
18 Bull C, Byrnes J, Hettiarachchi R, *et al*. A systematic review of the validity and reliability of patient-reported experience measures. *Health Serv Res* 2019;54:1023–35.
19 Juniper EF, Guyatt GH, Streiner DL, *et al*. Clinical impact versus factor analysis for quality of life questionnaire construction. *J Clin Epidemiol* 1997;50:233–8.
20 Schang L, Blotenberg I, Boywitt D. What makes a good quality indicator set? A systematic review of criteria. *Int J Qual Health Care* 2021;33:mzab107.
21 Friebel R, Steventon A. Composite measures of healthcare quality: sensible in theory, problematic in practice. *BMJ Qual Saf* 2019;28:85–8.
22 Barclay M, Dixon-Woods M, Lyratzopoulos G. The problem with composite indicators. *BMJ Qual Saf* 2019;28:338–44.
23 Griffiths A, Beaussier A-L, Demeritt D, *et al*. Intelligent monitoring? assessing the ability of the care quality Commission's statistical surveillance tool to predict quality and prioritise NHS Hospital inspections. *BMJ Qual Saf* 2017;26:120–30.
24 Kessner DM, Kalk CE, Singer J. Assessing health quality--the case for tracers. *N Engl J Med* 1973;288:189–94.
25 Barbazza E, Kringos D, Kruse I, *et al*. Creating performance intelligence for primary health care strengthening in Europe. *BMC Health Serv Res* 2019;19:1006.
26 Kessner DM. Quality assessment and assurance: early signs of cognitive dissonance. *N Engl J Med* 1978;298:381–6.
27 Boerma T, AbouZahr C, Evans D, *et al*. Monitoring intervention coverage in the context of universal health coverage. *PLoS Med* 2014;11:e1001728.
28 Paschen U, Vitt K-D. Das Tracer-Konzept Der Qualitätssicherung Im Krankenhaus – eine kritische Überprüfung. *Gesundheitswesen* 1992;54:460–4.