Editorial – *BMJ Quality and Safety*

Composite measures of health care quality: sensible in theory, problematic in practice

Rocco Friebel [1,2] and Adam Steventon [3]

[1] Department of Health Policy, London School of Economics and Political Science, London, United Kingdom

[2] Center for Global Development, Washington DC, United States of America

[3] Data Analytics, The Health Foundation, London, United Kingdom

All health care systems show variation in the quality of care provided, whether that means access to primary care services,[1] ambulance response times,[2] Accident & Emergency waiting times,[3] or treatment processes and outcomes.[4–6] Monitoring this variation in quality can serve multiple purposes: informing patients about where best to seek care; [7] allowing clinicians to compare their performance with that of their peers and thus identify targets for local-level quality improvement efforts, and supporting the development of national policy. Though, what all these have in common is a trust in the reliability of the data to adequately reflect health care quality – sometimes a questionable assumption.

In this issue of BMJ Quality and Safety, Hofstede et al. (2018) have addressed a common situation whereby providers (such as hospitals, general practices or community teams) are ranked according to their performance on a quality indicator. Rankings are used often to make direct performance comparisons between providers and used to identify positive or negative outliers. Yet, one of the downsides of this approach is that the ranks of providers can be susceptible to chance fluctuations in the indicators. The precision of rankings, *i.e.* their reliability, therefore has to be carefully assessed when developing these kinds of approaches to reporting the quality of care. This is particularly the case when payment is linked to performance,[9] or when on-going quality improvement efforts might be undermined by errors in the measurement.

Performance measures are driven by patient case-mix, differences in care provided and chance variation, with their accuracy to reflect on real quality variation determined by two components.[10] The first is the reliability of the indicator for each health care provider. This component (the 'within-provider uncertainty') is highly dependent on the number of patients receiving the type of care in question at each provider and is likely to be affected by random variation especially in smaller population groups. The second component is the variance in the indicators between providers. This 'between-provider uncertainty' relates to the true variation in the indicators between providers, setting aside chance variation within the individual providers. These distinctions are relevant because the reliability of the ranking system will depend on both the within and between provider uncertainty. One way to combine the two is to measure the 'rankability', which is defined as the ratio of between-hospital variation and the sum of between-hospital variation and within-hospital variation multiplied by 100.[10] This calculated percentage describes the level of variation due to true hospital differences, as opposed to random noise. Low values for this

percentage imply that variation in performance across hospitals largely reflects chance, not true differences in performance. Referring to this situation as having low rankability conveys the idea that hospital rankings are unstable: chance variation could just as easily have reduced quite different rankings. By contrast, high values for rankability mean that most observed variation in performance reflect real differences between hospitals – any given ranking is thus quite stable.

In their article, Hofstede and colleagues (2018) examine whether it is possible to improve the reliability of rankings based on quality measures. Two strategies are assessed: combining indicator data across several years to increase the number of events (for example, reporting readmission rates based on the number of admissions occurring over a multi-year period rather than a single year) or generating a composite measure by combining information from two or more quality indicators. Both approaches might improve the rankability – yet with some downsides in terms of the usefulness of the quality indicators, as we will come on to discuss.

The article makes use of Dutch National Medical Registration data for over half a million patients treated in 95 hospitals, containing indicators for in-hospital mortality, length of stay and 30-day readmission rates across twelve years. The authors considered a rankability ratio below 50% as low, considered it moderate between 50% and 75%, and high above 75%. Findings from the analysis show that both strategies can lead to a significant improvement in rankability compared with the use of any single outcome measure. Yet, composite measures showed greatest reliability of rankings in this study, and the authors conclude that composite measures provide more information and more reliable rankings than combining multiple years of individual indicators. But of course, there are other considerations that we now address.

What are the benefits of using composite measures?

The focus on composite quality measures is timely because they are being used in many health systems: Center for Medicare and Medicaid Services (CMS) for example, has introduced star ratings to measure the performance of Medicare Advantage Plans and Part D plans. Star ratings are available for 5 categories, covering aspects such as patient experience and access, while overall star ratings for drug plans are assigned across 4

categories, covering aspects such as drug safety.[11] In Germany, overall ratings are made publicly available for residential and domiciliary care homes, covering 59 and 34 single criteria across multiple quality dimensions, respectively.[12, 13]

The rationale for the adoption of composite measures is simple. Over the years, administrative data collected have become available and complemented by disease specific data from audits. The result has been a proliferation of outcome measures, which can result in information overload. Composite measures can help condense this vast amount of information into a single indicator, which is easy to use and promises an overview of performance.[14] Composite measures provide information that summarises a range of quality dimensions. This might be particularly helpful for the patients who tend to place great importance on several different aspects of quality, namely they want care that is effective, safe, patient-centred and delivered compassionately.

Limitations of composite measures

The potential benefits of composite measures might be outweighed by their substantial limitations (see table 1).[15] An independent review by the Health Foundation about approaches to measuring the quality of general practice in England discouraged the development and dissemination of composite scores.[15] One problem is that composite measures can lack the ability to signal changes in care quality that are specific enough to be the target of improvement projects. Quality improvement efforts are often directed towards a specific problem with care delivery and measured through a precisely defined set of indicators. Improvements against these indicators might not translate into changes in composite measures that also include information in other quality domains.

Another problem is that composite measures might pick-up potential spill-over effects. For example, a reduction in mortality can lead to a subsequent increase in hospital readmissions, since a greater proportion of patients now survive the initial hospital stay who would otherwise have died. If a composite measure was formed by combining data on mortality and readmission rates, then the two effects might be cancelled out. Another possible type of spill-over effect occurs when improvements to one area of care come at the cost of deteriorations elsewhere, for example due to limited resources. Although composite measures purport to offer a comprehensive and balanced view of quality across

several domains, this is only possible if the requisite data are available, yet if data on some domains are missing, then those domains will not be reflected as well as they should be on the composite score, which is potentially misleading.

Of course, individuals and stakeholder groups might differ in their assessment about the relative importance of the constituent measures. For example, patients place a great value on receiving care that is delivered compassionately and in a timely manner,[16] while clinicians might sometimes place greater emphasis on the delivery of effective treatments. A key challenge in the use of composite measures is therefore the weighing of selected single outcome measures to reflect individual preferences,[17] with different weighing methods being used such as equal-, numerator-, and opportunity-based weighting, or weighting on expert judgement.[18] Importantly, to ensure the usability of composite measures, their construction and selection of outcome measures has to be guided by the overall purpose of their use and tailored towards the end-user. Composite measures can be misleading when data on certain domains relevant to the end user are not available. Also, it might be challenging to adjust composite measures for confounders that can differ from one quality indicators to the next.

<<< Insert Table 1 here >>>

Conclusion

Clinicians, health care managers and policy-makers depend on most reliable information to make judgements about the impact of past initiatives on quality, and to guide future improvements. Composite measures are a good idea in theory as they can provide a way to make sense of the growing number of measures on various aspects of care quality. The companion paper also found that a composite measure of in-hospital mortality, 30-day readmission and prolonged length of stay showed better rankability than did individual indicators for some important medical and surgical examples commonly subjected to performance measurement. Indeed, rankability which describes the proportion of performance variation due to true differences rather than chance represents an important technical consideration for any performance measure. However, in practice composite measures suffer from significant limitations because of missing data, complex causalities, and difficulties setting the right weights to reflect individual preferences. Unless these

limitations are addressed, for instance through improving the transparency around the composites inherent aims and limitations, or by allowing users to adapt composites to reflect individual preference which could be aided through data visualisation tools,[19] their main applications are likely to be about helping patients to decide where to go for care, rather than quality improvement. People who are producing rankings of providers might be better advised to combine data across multiple years to make impact assessments, but ultimately as with any evaluation, the purpose of the quality measurement should determine the selection of the method.

References

1   Irving G, Neves AL, Dambha-Miller H, *et al.* International variations in primary care physician consultation time: a systematic review of 67 countries. *BMJ Open* 2017;**7**:e017902. doi:10.1136/bmjopen-2017-017902

2   National Audit Office. NHS Ambulance Services. 2017. https://www.nao.org.uk/wp-content/uploads/2017/01/NHS-Ambulance-Services.pdf

3   Giuntella O, Nicodemo C, Vargas-Silva C. The effects of immigration on NHS waiting times. *J Health Econ* 2018;**58**:123–43. doi:10.1016/j.jhealeco.2018.02.001

4   Friebel R, Hauck K, Aylin P, *et al.* National trends in emergency readmission rates: A longitudinal analysis of administrative data for England between 2006 and 2016. *BMJ Open* 2018;**8**:1–10. doi:10.1136/bmjopen-2017-020325

5   Friebel R, Hauck K, Aylin P. Centralisation of acute stroke services in London: Impact evaluation using two treatment groups. *Heal Econ (United Kingdom)* 2017;:1–11. doi:10.1002/hec.3630

6   Friebel R, Dharmarajan K, Krumholz HM, *et al.* Reductions in readmission rates are associated with modest improvements in patient-reported health gains following hip and knee replacement in England. *Med Care* 2017;**55**:834–40. doi:10.1097/MLR.0000000000000779

7   Gutacker N, Siciliani L, Moscelli G, *et al.* Choice of hospital: Which type of quality matters? *J Health Econ* 2016;**50**:230–46. doi:10.1016/j.jhealeco.2016.08.001

8   Hofstede SN. Ranking hospitals: do we gain reliability by using composite rather than individual indicators? *BMJ Qual Saf* 2018;**22 May**.bmjqs-2017-007669

9   Friebel R, Steventon A. The multiple aims of pay-for-performance and the risk of unintended consequences. *BMJ Qual Saf* 2016;**25**:827–31. doi:10.1136/bmjqs-

2016-005392

10   Van Dishoeck AM, Lingsma HF, Mackenbach JP, *et al.* Random variation and rankability of hospitals using outcome indicators. *BMJ Qual Saf* 2011;**20**:869–74. doi:10.1136/bmjqs.2010.048058

11   Centers for Medicare & Medicaid Services. Medicare star rating. 2018.https://www.medicare.gov/find-a-plan/staticpages/rating/planrating-help.aspx (accessed 25 Jul2018).

12   Spitzenverband G. Pflegenoten. https://www.gkv-spitzenverband.de/service/versicherten_service/pflegenoten_1/pflegenotenhinweise.jsp (accessed 9 Aug2018).

13   Rechel B, McKee M, Haas M, *et al.* Public reporting on quality, waiting times and patient experience in 11 high-income countries. *Health Policy (New York)* 2016;**120**:377–83. doi:10.1016/j.healthpol.2016.02.008

14   National Quality Forum. Composite Measure Evaluation Framework and National Voluntary Consensus Standards for Mortality and Safety—Composite Measures: A Consensus Report. 2009.

15   Dixon J, Spencelayh E, Howells A, *et al.* Indicators of quality of care in general practices in England. London, England: 2015. http://www.health.org.uk/publication/indicators-quality-care-general-practices-england

16   Joffe AS, Manocchia M, Weeks JC, *et al.* What do Patients Value in Their Hospital Care ? An Empirical Perspective on Autonomy Centred Bioethics. *J Med Ethics* 2003;**29**:103–8.

17   Schang L, Hynninen Y, Morton A, *et al.* Developing robust composite measures of healthcare quality - Ranking intervals and dominance relations for Scottish Health Boards. *Soc Sci Med* 2016;**162**:59–67. doi:10.1016/j.socscimed.2016.06.026

18   Shwartz M, Restuccia JD, Rosen AK. Composite Measures of Health Care Provider Performance: A Description of Approaches. *Milbank Q* 2015;**93**:788–825. doi:10.1111/1468-0009.12165

19   Barclay M, Dixon-Woods M, Lyratzopoulos G. The problem with composite indicators. *BMJ Qual Saf* 2018;:10.1136/bmjqs-2018-007798. doi:10.1136/bmjqs-2018-007798

20   Montori VM, Permanyer-Miralda G, Ferreira-González I, *et al.* Validity of composite end points in clinical trials. *Bmj* 2005;**330**:594–6.

Table 1: Advantages and disadvantages of composite quality measures

Advantages	Disadvantages
• Simple, avoiding information overload in light of the growing number of performance indicators	• Might not provide sufficient detail to enable practitioners to identify areas of need for improvement, and it can also be challenging to attribute changes in composite measures to specific quality improvement efforts
• Able to summarise information across several quality dimensions	• Masks specific changes to care quality, for example if changes to morbidity and mortality do not correlate with other dimensions such as the patient experience
• Improved rankability compared with the use of a single indicator, even if that indicator is produced using data that span multiple years	• The weights attached to the individual indicators might not reflect their importance to the users of the data, particularly if users place different value on aspects of care quality
• Helps to overcome problems associated with small sample sizes for individual quality indicators (*e.g.*, if there are only few mortality cases in one hospital, but large numbers of hospital readmissions), analogous to their appeal in clinical trials[20]	• The constituent indicators might have varying levels of robustness (for example with regards to gaming or coding practices), making it hard to assess the validity of the composite measure
	• It can be challenging to adjust composite measures for differences in patient characteristics between providers, when the confounders can differ between one constituent indicator to the next