https://ijhpm.com
Int J Health Policy Manag 2026;15:10217

OPEN ACCESS

doi 10.34172/ijhpm.10217

IJHPM
International Journal of Health Policy and Management

Commentary

Measuring Quality or Measuring Noise? Lessons From the Dutch Breast Cancer Audit

Comment on "Comprehensive Evaluation of Quality Indicators: Analyzing the Dutch Breast Cancer Audit"

Olaf Johan Hartmann*

Abstract

Cancer registries and clinical audits are increasingly used to compare healthcare performance across institutions. However, observed differences between hospitals are often interpreted as true differences in quality without considering whether the indicators can reliably distinguish between providers. This commentary discusses the framework applied by Verheul and colleagues to the Dutch Breast Cancer Audit. It distinguishes registry data quality from indicator performance and describes four complementary properties in sequence: feasibility, discriminative ability, validity, and reliability, with rankability as a measure of the latter. Reliability deserves particular attention when indicators guide public reporting, reimbursement, policy, or patient choice, because low rankability indicates that apparent differences may largely reflect random variation. International standards and the broader reliability literature support pre-use testing, transparent reporting of uncertainty and sample-size requirements, periodic re-evaluation, and restriction or retirement of unreliable indicators. Without adequate reliability, institutional differences should not be interpreted as true differences in quality of care.

Keywords: Quality Indicators, Clinical Registries, Reliability, Rankability, Benchmarking, Netherlands

Citation: Hartmann OJ. Measuring quality or measuring noise? Lessons from the Dutch Breast Cancer Audit: Comment on "Comprehensive evaluation of quality indicators: analyzing the Dutch Breast Cancer Audit." *Int J Health Policy Manag.* 2026;15:10217. doi:10.34172/ijhpm.10217

Article History:
Received: 29 July 2026
Accepted: 10 September 2026
ePublished: 29 September 2026

***Correspondence to:**
Olaf Johan Hartmann
Email:
olaf.johan.hartmann@ahus.no

Imagine two hospitals treating patients with breast cancer. One appears to perform better than the other according to publicly reported quality indicators. Patients may choose one hospital over the other. Politicians may use the data to guide healthcare policy. Hospital managers may reward or penalise departments based on the results.

But what if the observed difference is largely due to chance?

This simple question is central to the study by Verheul and colleagues. In their evaluation of quality indicators used in the Dutch Breast Cancer Audit, they ask an important but often overlooked question: How reliable are the measures we use to judge healthcare quality?[1]

Quality indicators have become central tools for benchmarking healthcare performance, informing policy decisions, supporting accreditation processes, and guiding quality improvement initiatives. Most clinicians support measuring variation in healthcare delivery. However, measurement only improves care if the indicators themselves accurately reflect the quality of care being delivered.[2-4]

Substantial resources are devoted to collecting and reporting quality indicators. Many of these indicators influence hospital benchmarking, public reporting, reimbursement, and strategic decision-making. Yet relatively little attention is paid to whether the indicators themselves are sufficiently robust.

Two distinct levels of quality assessment are recognised. The first concerns the quality of the registry data, including completeness, accuracy, comparability, and timeliness. The second concerns the measurement quality of the indicator itself: whether it reflects the quality of care being delivered, distinguishes providers reliably, and is sufficiently protected from differences in case mix and random variation. A registry may therefore contain data of excellent quality while producing an indicator that provides little meaningful information about the quality of care received by patients. High-quality data are essential, but they cannot compensate for a poorly designed indicator: one may measure the wrong thing with great precision.

Four related but distinct properties should be considered when evaluating a quality indicator (Table). Feasibility concerns whether the necessary data can be collected completely and with a reasonable burden. Discriminative ability concerns whether scores vary sufficiently between hospitals to identify potential differences or room for improvement. Validity concerns whether the indicator reflects the aspect of healthcare quality it is intended to measure; in Verheul et al, this was operationalised as the influence of measured case-mix variables on hospital scores. Reliability concerns whether observed differences are reproducible and

Department of Breast and Endocrine Surgery, Akershus University Hospital, Lørenskog, Norway.

Table. Four Key Properties for Evaluating Quality Indicators, as Applied by Verheul and Colleagues, and Their Practical Meaning

Property	Question It Answers	Practical Meaning
Feasibility	Can the required data be collected?	Data are sufficiently complete and collection is practical.
Discriminative ability	Do scores vary between hospitals?	There is enough between-hospital variation to identify possible differences or room for improvement; variation alone does not prove differences in quality.
Validity	Does the indicator reflect the aspect of healthcare quality it is intended to measure?	Scores should be clinically meaningful. In Verheul et al, validity was operationalised as the influence of measured case-mix variables on hospital scores.
Reliability (assessed by rankability)	Can observed differences between hospitals be distinguished from random variation?	Rankability estimates the proportion of observed between-hospital variation not attributable to chance. High rankability is necessary, but not sufficient, for meaningful ranking.

can be distinguished from random variation.

These properties are consistent with broader frameworks for indicator evaluation, although terminology and operational definitions remain far from standardised. Reviews and assessment instruments have identified a wider range of relevant properties, including clinical importance, scientific soundness, usability, actionability, and feasibility.[5,6]

Although all four properties are necessary, reliability deserves particular attention when indicators are used to compare or rank hospitals. This is not because reliability supersedes the other properties, but because it determines whether observed between-hospital variation can support meaningful comparison. An indicator may be feasible, demonstrate between-hospital variation and show little influence of measured case mix, yet still be unsuitable for benchmarking if the observed differences are largely attributable to chance. In the framework applied by Verheul and colleagues, reliability was assessed using rankability, which estimates the proportion of observed between-hospital variation not attributable to random variation.[1]

When applied to 18 indicators from the Dutch Breast Cancer Audit, feasibility was generally high and measured case mix had little influence on most indicators; however, none achieved high rankability.[1] These findings show that an indicator may perform well according to the other evaluated properties while remaining unsuitable for ranking hospitals. Expert consensus remains important for identifying clinically meaningful measures, but it cannot establish reliability. Indicators that lack reliability may create misleading conclusions, encourage inappropriate benchmarking, and shift attention towards improving metrics rather than patient care.

The United States provides one model for systematic oversight. Battelle, a Centers for Medicare & Medicaid Services-certified consensus-based entity, conducts measure endorsement and maintenance reviews through the Partnership for Quality Measurement. Measures undergo a transparent, consensus-based assessment that includes evidence, feasibility, reliability, validity, and usability.[7] Guidance developed by the National Quality Forum Scientific Methods Panel for risk-adjusted outcome measures similarly emphasises reliability at the accountable-entity level, appropriate risk adjustment, and statistical uncertainty.[8] This lifecycle approach contrasts with treating expert consensus as sufficient for permanent adoption.

Evidence from other settings reinforces the practical consequences. In cardiac surgery, only 4.4% of hospitals achieved reliability above 0.70 for 30-day readmission rates, and 599 cases over three years were required to reach that threshold.[9] More recently, a comparison of 14 reliability methods showed substantially different estimates, with discrepancies widening as sample size decreased.[10] A RAND Corporation technical report on physician profiling further demonstrates how low reliability increases the risk of provider misclassification.[11] Reliability is therefore not merely a statistical refinement; it determines whether labels, rankings, and their associated consequences can be defended.

These findings have implications extending far beyond breast cancer care. Increasingly, healthcare systems use quality indicators to inform patients, regulators, payers, policy-makers, and healthcare providers. When indicators are used for public reporting or reimbursement decisions, confidence in their validity and reliability becomes essential. Indicators that inadequately distinguish between providers may create unintended consequences, including reputational harm and misallocation of quality improvement efforts.

The outcomes that matter most to patients—survival, recurrence, complications, functioning, and quality of life—may take years to observe and are influenced by disease severity and other patient characteristics. Healthcare systems therefore rely heavily on process indicators, such as whether recommended diagnostic or treatment steps were completed. These measures are easier to collect and often actionable, but adherence to a process does not by itself prove better outcomes or better care for an individual patient. Their value depends on a strong evidence-based link to meaningful patient benefit and on the indicator itself being valid and reliable, particularly when it is used to compare hospitals or guide policy.[4]

Moreover, valid individual indicators do not necessarily constitute a valid representation of healthcare quality. An indicator set may overrepresent easily measured processes while omitting outcomes and experiences that matter most to patients. The objective should therefore not be to produce more indicators, but to identify a parsimonious set of indicators that measures what matters.[12,13]

The study also highlights an important distinction between internal quality improvement and external benchmarking. Indicators with limited reliability may still provide value as tools for local learning and multidisciplinary discussion. However, the standards required for public reporting should

arguably be higher. Before indicators are used to compare hospitals publicly, they should demonstrate that observed differences represent meaningful differences in care rather than statistical noise.

Several practical steps follow. The intended use of an indicator should be specified before testing, because a signal used for local learning need not meet the same reliability standard as a measure used for public ranking or payment. Developers should assess reliability at the level at which results will be reported, publish uncertainty, and determine minimum sample-size requirements. Where reliability is low, options include extending reporting periods, increasing sample size, combining related outcomes in clinically coherent composites, or restricting the indicator to internal use; if these approaches do not produce adequate reliability, the measure should be revised or retired. Statistical shrinkage may stabilise estimates but cannot create information absent from small samples. Indicators should also be re-evaluated periodically as evidence, practice patterns, case mix, and performance change. Future research should compare reliability estimators and define use-specific thresholds that minimise misclassification without sacrificing validity, feasibility, or actionability.

Most clinicians understand that laboratory tests require validation before entering routine practice. The same principle should apply to quality indicators, and the evidentiary standard should reflect the consequences of their use.

As healthcare increasingly embraces data-driven decision-making, the question is not whether quality should be measured, but whether measures justify the decisions based upon them. Verheul and colleagues show that without adequate rankability, apparent differences between institutions may represent statistical noise rather than genuine differences in quality. Quality indicators, like clinical instruments, require rigorous and repeated validation before they guide practice or policy.[1]

Disclosure of artificial intelligence (AI) use

OpenAI ChatGPT was used to assist with literature searching, language editing, and improving the clarity and organisation of the manuscript. The author independently verified the sources, critically reviewed and revised all output, and takes full responsibility for the final content.

Ethical issues

Not applicable.

Conflicts of interest

Author declares that he has no conflicts of interest.

Disclaimer

The views expressed in this commentary are those of the author and do not necessarily reflect those of Akershus University Hospital.

References

1. Verheul EM, van der Linde M, Lingsma HF, et al. Comprehensive evaluation of quality indicators: analyzing the Dutch Breast Cancer Audit. *Int J Health Policy Manag.* 2025;14:8943. doi:10.34172/ijhpm.8943
2. Vos EL, Lingsma HF, Jager A, et al. Effect of case-mix and random variation on breast cancer care quality indicators and their rankability. *Value Health*. 2020;23(9):1191-1199. doi:10.1016/j.jval.2019.12.014
3. Donabedian A. The quality of care. How can it be assessed? *JAMA*. 1988; 260(12):1743-1748.doi:10.1001/jama.260.12.1743
4. Maes-Carballo M, Gómez-Fandiño Y, Reinoso-Hermida A, et al. Quality indicators for breast cancer care: a systematic review. *Breast*. 2021; 59:221-231. doi:10.1016/j.breast.2021.06.013
5. Backes J, Vogel J, Geissler A, et al. Welcome to the jungle: collection and evaluation of quality indicators with the QUALICATOR instrument. *BMJ Qual Saf*. 2026;35:565-575. doi:10.1136/bmjqs-2026-020265
6. Jones P, Shepherd M, Wells S, Le Fevre J, Ameratunga S. Review article: what makes a good healthcare quality indicator? A systematic review and validation study. *Emerg Med Australas*. 2014;26(2):113-124. doi:10.1111/1742-6723.12195
7. Partnership for Quality Measurement. What endorsement means. https://p4qm.org/em/about. Accessed September 2, 2026.
8. Glance LG, Joynt Maddox K, Johnson K, et al. National Quality Forum guidelines for evaluating the scientific acceptability of risk-adjusted clinical outcome measures: a report from the National Quality Forum Scientific Methods Panel. *Ann Surg*. 2020;271(6):1048-1055. doi:10.1097/SLA.0000000000003592
9. Shih T, Dimick JB. Reliability of readmission rates as a hospital quality measure in cardiac surgery. *Ann Thorac Surg*. 2014;97(4):1214-1218. doi:10.1016/j.athoracsur.2013.11.048
10. Nieser KJ, Harris AHS. Comparing methods for assessing the reliability of health care quality measures. Stat Med. 2024;43(23):4575-4594. doi:10.1002/sim.10197
11. Adams JL, Mehrotra A, McGlynn EA. Estimating Reliability and Misclassification in Physician Profiling. RAND Corporation; 2010. TR-863-MMS. https://www.rand.org/pubs/technical_reports/TR863.html.
12. 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(3):mzab107. doi:10.1093/intqhc/mzab107
13. Meyer GS, Nelson EC, Pryor DB, et al. More quality measures versus measuring what matters: a call for balance and parsimony. *BMJ Qual Saf*. 2012;21(11):964-968. doi:10.1136/bmjqs-2012-001081