Variation in quality of care between hospitals: how to identify learning opportunities

Alex Bottle ,[1] Pia Kjær Kristensen [2,3]

[1]School of Public Health, Imperial College London Faculty of Medicine, London, UK
[2]Clinical Epidemiology, Aarhus Universitetshospital, Aarhus, Denmark
[3]Orthopedic, Region Hospital Horsens, Horsens, Denmark

Correspondence to
Professor Alex Bottle, School of Public Health, Imperial College London Faculty of Medicine, London, W12 7TA, UK; robert.bottle@imperial.ac.uk

Accepted 1 March 2024
Published Online First 8 March 2024

Linked

► http://dx.doi.org/10.1136/bmjqs-2023-016726

Check for updates

To cite: Bottle A, Kristensen PK. *BMJ Qual Saf* 2024;**33**:413–415.

In healthcare, as in life, the adage 'variety is the spice of life' often holds true. Variation can represent individual patient preferences, but when it comes to the quality of healthcare, variation can also be unwanted and harmful. Analysis of variation in a quality-of-care indicator assumes that finding only limited variation is a good thing, suggesting consistently high compliance with evidence-based guidelines and providing evidence of equity. In this editorial, we consider how variation is and should be quantified, comment on the findings of a review[1] in this issue of *BMJ Quality and Safety*, and explore whether measurement at the hospital level is best for learning. We conclude by reflecting on the assumption that only limited variation is good.

How is variation analysed? Take CT scanning for suspected stroke as an example. This should be done soon after the patient arrives in the emergency department. The scan rate within 24 hours can be easily calculated at each hospital as the proportion of patients who get a scan within that time frame, but how should we summarise the spread of 24-hour scan rates across hospitals? There are two ways: absolute (eg, an interquartile range of 65–95% between hospitals) and relative (eg, a median odds ratio of 1.25). Unfortunately, regarding the former, news outlets will often focus on the rates' range, which is the most misleading summary; the hospital with the lowest rate may even be named and shamed as the 'worst'. Identification of unwarranted variation between healthcare providers often captures media attention, and such variation needs to be quantified accurately. The range is the most sensitive to random variation (largely due to small numbers), data entry errors, inadequate case-mix adjustment and other distortions. The preferred method to deal with randomness and give relative measures of variation is multilevel modelling. This method allows us to break down the sources of variation by 'level', such as patient, doctor and hospital. A common alternative analytical method, which also gives relative measures of variation, is to use generalised estimating equations. However, this can only handle two levels in the data and cannot separate the sources of variation, making it less useful for learning. In the stroke scan example, some types of patient will get a scan sooner than others due to their symptoms or lack of other medical problems, some doctors may be better at recognising the need for an urgent scan and some hospitals will be set up better to facilitate rapid access to a CT machine. We consider later which of these levels is best for learning. In a multilevel model, the variation in the quality indicator is separated or 'partitioned' by level, giving a variation partition coefficient (VPC), also called the intraclass correlation coefficient. This quantifies the between-level variation as a proportion of the sum of the between-level and within-level variation, typically expressed as a percentage. The higher the value, the more the indicator varies by that level, for example, hospital. Multilevel modelling can also be formulated in different ways for risk-adjusted outcomes, as reviewed elsewhere.[2]

How much non-random variation in quality indicators exists between hospitals? This matters because a lot of effort is targeted at assessing hospitals in initiatives such as accreditation, public reporting of performance and regulator inspections, and many national clinical audits and collaboratives such as the USA's *Get With The Guidelines* series are set up to analyse and feed back performance data at this

level. In this issue of *BMJ Quality and Safety*, van der Linde *et al*[1] set out to systematically review the literature on how much non-random variation in quality indicators exists, updating an earlier 2010 review[3] and exploring differences by diagnosis area and type of indicator. They included 44 studies and 144 indicators; it is notable that 52 of the 65 excluded studies were excluded because they did not use multilevel modelling, or the variation was not expressed relative to the total variation in another way. They recommend that absolute measures of variation are needed but should be given alongside relative measures. Hospital-level and doctor-level variation was low overall, especially for outcome indicators, with VPCs for outcomes just 1.4%. No clear differences were found by diagnosis. As the authors acknowledge, one would expect more variation in process than outcome indicators because the former are much more under the hospital's control and more likely to be affected by patient preferences and contraindications. Only two studies assessed their measures' reliability, which is a function of sample size. With small samples, random variation will often dominate, and users of quality indicators need to know how likely this is. It is well-known that small numbers of events, which occur particularly when expressed per doctor but can also happen per hospital, make it hard to distinguish between low and high performers.[4 5] Other gaps in the literature that van der Linde *et al*'s review revealed are for the levels ward and nursing unit within the hospital (just two studies found) and for healthcare professionals (also just two VPC estimates).

How could this study's findings be used in practice? The authors conclude that targeting hospitals might not be the best approach for quality of improvement initiatives. This is because of the limited variation at that level and sometimes inadequate statistical reliability, although they acknowledge some advantages of doing so in practice, for example, when investigating process performance measures. The authors suggest that efforts to ensure high-quality care should be at the level of the healthcare system, which could be a national or hospital system, rather than targeted at particular hospitals. However, the appropriateness of this also depends on the range of absolute rates. For example, if a top-performing hospital has a process measure rate that is genuinely twice as high as that in other hospitals, it could indicate important learning opportunities. Therefore, analysing variations in quality indicators using both an absolute and a relative approach is a crucial step before implementing targeted quality improvement initiatives. Compared with outcome measures, process indicators showed more variance at the hospital level in van der Linde *et al*'s review and are more relevant when evaluating hospitals in initiatives such as accreditation, public reporting of performance, regulator inspections and national clinical audits. These measures offer quick feedback on the efficiency and effectiveness of specific healthcare processes within an organisation and, in theory, do not require case-mix adjustment. However, there are challenges associated with process-level quality indicators, including the practicalities of data collection and the potential for manipulating ('gaming') processes to align with specific standards or targets (outcome measures can also be gamed but less easily).[6] This manipulation distorts performance data, undermining the effectiveness of the measurement system and resulting in missed learning opportunities.

The review has some limitations, of course. One is the exclusion of models that included hospital-level variables, for example, teaching status. This is understandable, as such factors can be proxies for quality, but the authors may have gone too far with this; also, in three studies, they were unable to separate hospital characteristics from baseline models that did not include them, so the exclusion process was not a perfectly clean procedure. A second limitation is that they included only studies using the multilevel approach, which means that the proportion of papers reporting absolute variation and/or reliability may not be representative of the wider literature. Moreover, except in the minority of studies that measured variation by doctor or by unit such as ward or department, hospitals are assumed to be homogeneous organisations. Data for a few wards are assumed to represent the whole site. We know that surgical outcomes in the same hospital can differ by department or depending on whether the surgery was planned or unplanned.[7] Even for the same group of patients, outcomes may differ by ward, for example, patients with heart failure do better if admitted to the coronary care unit than to a general ward.[8] Patients in general do worse if not in the most appropriate ward for their condition, for example, medical patients on surgical wards.[9] A related point is that hospitals can be part of larger entities that may have some influence on performance and therefore on the estimate of between-unit variation: hospital sites can be part of trusts in the UK or of networks or consortia. Studies not done on national data, for example, those for a single Canadian province or Australian state, have the added complication of restricted generalisability to that region.

There is much potential for further work in this area. There are still too few studies using multilevel modelling. One-third of the 44 included studies were from the USA, with very few outside of high-income countries. More work is needed to cover subhospital levels such as doctor or ward, where quality improvement efforts are much more actionable by local teams. There is scope for a review of studies using absolute estimates of variation that, unlike when simply reporting the range, do account for random variation, for example, through funnel plots. An example of that is an analysis of English administrative data on COVID-19 mortality by hospital that used both multilevel modelling and

reported the proportion of funnel plot outliers with and without case-mix adjustment.[10]

Variation in healthcare indicators is frequently denounced as a scandal, as care quality should not resemble a 'postcode lottery'. However, is variation in healthcare processes and outcomes inherently problematic? Or could it be the foundation for learning and development within the healthcare sector? Differences in outcomes that are clearly due to the use by some providers of practices proven to be inferior teach us little, but what about the many areas of medicine lacking strong evidence and/or consensus? For example, research using clinical vignettes shows substantial disagreement between surgeons on the role of surgical intervention, partly stemming from their perception of the risks and benefits of operative and non-operative management.[11] More generally, healthcare providers differ in their willingness to experiment with new ways of working, which, probably after some failures, may go on to redefine what 'good' quality of care is. The improvements (or otherwise) in patient outcomes that result from these differences in treatment choices need to be captured by continuous data monitoring. Data collection must be paired with robust statistical analysis capable of distinguishing between random variation and where the variation unfolds (particularly at patient, physician, department or hospital level). The review by van der Linde *et al* is a good summary of what we know about this, but, as with healthcare as a whole, there is still much to discover.

X Alex Bottle @DrAlexBottle and Pia Kjær Kristensen @pia_kjar

Contributors AB drafted the first version of the manuscript, and both authors edited it.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Commissioned; internally peer reviewed.

ORCID iDs
Alex Bottle http://orcid.org/0000-0001-9978-2011
Pia Kjær Kristensen http://orcid.org/0000-0001-5473-9386

REFERENCES

1 van der Linde M, Salet N, van Leeuwen N, *et al*. Between-hospital variation in indicators of quality of care: a systematic review. *BMJ Qual Saf* 2024;33:443–55.
2 Bottle A, Aylin P. Statistical methods for Healthcare performance monitoring. In: *Statistical Methods for Healthcare Performance Monitoring*. 1st ed. CRC Press, 2016.
3 Fung V, Schmittdiel JA, Fireman B, *et al*. Meaningful variation in performance: a systematic literature review. *Med Care* 2010;48:140–8.
4 Normand S-LT, Shahian DM. Statistical and clinical aspects of hospital outcomes profiling. *Statist Sci* 2007;22:206–26.
5 Austin PC, Reeves MJ. Effect of provider volume on the accuracy of hospital report cards: A Monte Carlo study. *Circ: Cardiovascular Quality and Outcomes* 2014;7:299–305.
6 Catlow J, Bhardwaj-Gosling R, Sharp L, *et al*. Using a dark logic model to explore adverse effects in audit and feedback: a qualitative study of gaming in colonoscopy. *BMJ Qual Saf* 2022;31:704–15.
7 Ingraham AM, Cohen ME, Raval MV, *et al*. Comparison of hospital performance in emergency versus elective general surgery operations at 198 hospitals. *J Am Coll Surg* 2011;212:20–28e1.
8 National Heart Failure Audit (NHFA) 2022 Summary Report. Healthcare Quality Improvement Partnership, . 2022Available: https://www.nicor.org.uk/wp-content/uploads/2022/06/NHFA-DOC-2022-FINAL.pdf
9 Goulding L, Adamson J, Watt I, *et al*. Patient safety in patients who occupy beds on clinically inappropriate wards: a qualitative interview study with NHS staff. *BMJ Qual Saf* 2012;21:218–24.
10 Bottle A, Faitna P, Aylin PP. Patient-level and hospital-level variation and related time trends in COVID-19 case fatality rates during the first pandemic wave in England: Multilevel Modelling analysis of routine data. *BMJ Qual Saf* 2022;31:211–20.
11 Sacks GD, Dawes AJ, Ettner SL, *et al*. Surgeon perception of risk and benefit in the decision to operate. *Ann Surg* 2016;264:896–903.