Peer review of quality of care: methods and metrics

Julian Bion , Joseph Edward Alderman

Intensive Care Medicine, University of Birmingham College of Medical and Dental Sciences, Birmingham, UK

Correspondence to
Professor Julian Bion, Intensive Care Medicine, University of Birmingham College of Medical and Dental Sciences, Birmingham B15 2TT, UK; J.F.Bion@bham.ac.uk

Accepted 12 July 2022
Published Online First 21 July 2022

► http://dx.doi.org/10.1136/bmjqs-2021-013864

To cite: Bion J, Alderman JE. *BMJ Qual Saf* 2023;**32**:1–5.

The privilege of professional self-regulation rests on clinical peer review, a long-established method for assuring quality of care, training, management and research. In clinical peer review, healthcare professionals evaluate each other's clinical performance. Based originally on the personal experience and expertise (and prejudices and biases) of one's peers, the process has gradually been formalised by the development of externally verifiable standards of practice, audit of care processes and outcomes and benchmarking of individual, group and organisational performance and patient outcomes. The spectrum of clinical peer review ranges from local quality improvement activities such as morbidity and mortality reviews, to medical opinion offered in courts of law. Peer review can therefore have different purposes ranging from collaborative reflective learning to identification of malpractice.

Given the ubiquity and importance of clinical peer review, it would be reasonable to expect some evidence of reliability of judgements made by different reviewers. And yet the literature tells a rather different story. A systematic review[1] of the inter-rater reliability of audited case records reported mean kappa values ranging from 0.32 to 0.7, with higher reliability when reviewers employed explicit criteria. Reviewers may give inconsistent judgements, change their opinions over time[2] and be susceptible to a variety of biases including implicit,[3] cognitive[4] and outcome or hindsight bias.[5] To some extent, this may be mitigated and reliability improved by using a combination of both criterion-based and implicit (global) assessment[6] combined with structured judgement templates,[7 8] or when a smaller group of reviewers is employed to detect well-characterised signals such as adverse events.[9] In a comparison of weekend and weekday quality of care across two epochs of time, using a combination of structured judgement and global (implicit) reviews of case records,[10] we found modest levels of agreement between reviewers examining the same case, but a high level of agreement when cases were aggregated at organisational level: the big picture was more informative than the individual case. In legal settings in the UK, the Woolf recommendations encourage consensus between expert witnesses by requiring a single, joint assessment from experts appointed by the courts.[11] While this approach may have improved matters for the courts,[12] the evidence that consensus-based reviews produce more accurate judgements is elusive.[13] This is problematic when the stakes are so high for patient care and for individual and organisational reputations. These methodological challenges to peer review have the potential to undermine the edifice of self-regulation: if the instrument of investigation (the reviewer) is so flawed, how can we have confidence in the outcome—the judgement of quality?

In an attempt to determine the utility of peer review, Schmitt *et al*[14] in this edition of the journal performed a cluster randomised trial of 60 hospitals, nested within the German 'Initiative Qualitätsmedizin' (IQM), a voluntary national multiprofessional quality improvement collaboration established in 2009 involving 385 hospitals. The population they chose for review was intensive care unit patients receiving mechanical ventilation for >24 hours. The 60 hospitals selected were those with the highest hospital mortality rates in 2016, the rationale being that these hospitals would have the greatest headroom for improvement. The logic model therefore

contains the assumption that higher mortality rates are attributable, at least in part, to deficiencies in care processes which can be identified and corrected through peer review.

Ultimately, the analysis was based on 30 intervention hospitals and 29 control hospitals caring for 12 085 and 13 016 patients in the pre-intervention and post-intervention periods, respectively. Thirty-three of the 60 hospitals had previously participated in clinical peer review of mechanical ventilation. Data from the non-participating hospitals ('observation arm') were used to derive standardised hospital mortality ratios based on a range of characteristics which included hospital coding for emergency admission status and comorbidities, but not acute physiology or severity of illness.

Clinical peer review in the intervention hospitals consisted of several linked steps: self-selection of 12–16 records of patients who had died; self-assessment of care quality; on-site assessment by the review team consisting of trained doctors and nurses from other IQM hospitals and a structured report and discussion between reviewers and staff to agree 'clear and precisely formulated potentials for improvement to derive an action plan', the implementation of which was the responsibility of the local clinicians. This therefore fulfils the Medical Research Council's criteria for a complex intervention.[15] The authors used a difference-in-difference analysis to mitigate the impact of case mix and organisational differences between hospitals. The primary outcome was the difference in the pre-intervention and post-intervention standardised mortality ratios 1 year before and 1 year after peer review. What did they find? There was no impact of the intervention on either crude or adjusted mortality ratios. Peer review had no perceptible impact on mortality.

How should we interpret this apparent lack of effect? The authors took care to ensure adequate power for their study. Could the context have been unfavourable? It seems unlikely that intensive care staff would 'lack receptors' for quality improvement interventions since, even when metrics are disputed,[16 17] clinical staff achieve improvements in care over time when given the tools.[18] Moreover, when the authors invited participation in the project, they received positive responses from 237 hospitals, which does not suggest lack of interest. Or should we accept the null hypothesis and conclude that clinical peer review can join the list of other ineffective interventions in critical illness,[19] with wide implications for the whole of medicine?

We suspect that in addition to the unreliability of peer review discussed above, its impact will have been diminished further by methodological issues, some of which are acknowledged by the authors. These need to be addressed in future research. We consider here case selection, the choice of process or outcome measures, and the content of the intervention.

In terms of case selection, restricting the investigation to mortality reviews limits the generalisability of the study and introduces the problem of simultaneity or endogenous selection bias[20] in which the selected population (in this case patients requiring mechanical ventilation for >24 hours) and the target of the intervention (reliability improvement or error prevention) lie on the causal pathway to the primary outcome (mortality), and the risk of that outcome is itself a potential ('simultaneous') contributor to the probability of requiring prolonged ventilation or of experiencing an error or omission in care,[21] a form of 'reverse causation'.[22 23] While mortality reviews may reveal valuable opportunities for improvements in care of individual patients, the rationale for examining *only* those episodes of care which ended in death may be erroneous when reviewing care quality aggregated at unit or organisational level. Deficiencies in care processes do not generally result in death, while patients who die may have received exemplary care, even though they may have had more complex pathways with greater opportunity for errors. When evaluating quality of care therefore, it may be better to study this in a fully representative patient population, not just in those who died.

Standardised mortality ratios as a measure of care quality suffer from several methodological deficiencies,[24] including sensitivity to unmeasured aspects of case mix,[25 26] and the low proportion of deaths classed as avoidable, with the majority of deaths being a consequence of the patients' acute or comorbid diseases.[27–29] Importantly, there is no clear relationship between organisation-level standardised mortality ratios and clinical judgements of care quality.[30] Institutional reporting rates of incidents involving severe harm or death similarly show no relationship with mortality or patient satisfaction.[31 31] The signal-to-noise ratio is therefore adverse, since the opportunity for identifying improvements in care processes leading to improvements in outcomes is dependent on the total number of cases and the proportion which are genuinely avoidable.[32] The question then is 'how much of the (adjusted) mortality risk can be attributed to deficiencies in processes of care which can be detected by peer review and controlled by the clinical team'? The answer to the first part may be 'not much', but the answer to the second element (detection and control) could be anything from 'variable' to 'substantial'. And incremental improvements in care processes over time may add up to important gains which emerge as gradual secular trends,[18 33 34] the 'rising tide' phenomenon.[35]

Should one use care processes or outcomes to assess quality improvement interventions? Outcomes matter to patients and to staff; but at what point

should the measurements be censored—28-day, hospital survival, 3 months postdischarge, 12 months? And what about quality of survival? Duration alone is insufficient for those living with multimorbidity. Process measures may be more laborious to collect, but they offer a more rapidly available quality signal than outcome, and are more 'empowering' as they give a clearer indication of what staff need to do to improve care by providing an explicit link between the metrics, the content of the intervention and consequential actions. Process measures and the criteria used for the reviews may have both technical and behavioural-social components. The technical components will include evidence from randomised trials of interventions which influence outcome and for which there is a performance gap. The obvious example in this case is lung-protective ventilation which is still not used reliably in around one-third of eligible patients with adult respiratory distress syndrome[36] and even fewer receiving intraoperative ventilation for elective or emergency surgery[37] even though standardising best practice reduces mortality.[38] Adherence may be higher in patients ventilated for COVID-19 pneumonia[39] suggesting that the trend to standardisation of treatment was accelerated by the pandemic. Other interventions could include venous thromboembolism prophylaxis, sedation minimisation, use of neuromuscular blockade and selective digestive decontamination depending on the patient population. Behavioural and social components of quality interventions may include communication, teamworking, use of checklists[40] and ability to challenge or raise concerns.[41] Behavioural barriers—which may be subtle and difficult to detect—include disputes about evidence, loss of autonomy and divergent views on clinical responsibilities,[42] disagreement about the validity of performance metrics[43] and difficulty in sustaining improvement over time.[44]

The contents of the intervention—the process 'targets'—should lie on the causal pathway to the desired outcome, and as far as possible should be supported by evidence both for impact and a gap analysis indicating headroom for improvement. Schmitt *et al* found 132 discrete recommendations for improvement in the intervention hospitals, 81 of which had been, or were being, implemented, but 53 (66%) of these were regarded as unlikely to affect mortality. This captures succinctly the problem of peer review: there are many aspects of the care pathway which might be done differently, or better, but which of these is really important? To answer that question needs preliminary diagnostic work to understand the problem before deciding that peer review is the right vehicle, and which treatments it needs to bring to bear.

Quality improvement research is still at an early stage in the development of rigorous methodologies.[43] Schmitt *et al* are to be applauded for having employed a cluster randomised trial to evaluate a quality improvement intervention and for having documented the contents of the intervention with clarity. Future work needs to address the issue of what constitutes a representative patient population; to consider incorporating contextual factors; to determine which processes of care really influence outcomes and to identify gaps in current practice and whether there is sufficient headroom for improvement. These elements should be brought together in the form of a logic model[44] offering a theory of change which may then be tested using methods such as realist evaluation.[45 46] Increasing sophistication of the electronic patient record (EPR) may reduce dependence on peer review, since if the correct processes of care are known, then error correction can be incorporated in real time in the form of prompts, reminders and automated control limits, with performance benchmarked against one's peers. This will work for the technical aspects of care, but the non-technical, behavioural aspects such as effective compassionate communication and teamworking cannot really be determined from the EPR. The future of peer review may lie not in the retrospective examination of case records, but in the contemporaneous observations of practice by peers and patients, within a model of workplace-based reflective learning.[47]

Twitter Joseph Edward Alderman @jaldmn

Contributors Both authors contributed to the review and writing the editorial.

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
Julian Bion http://orcid.org/0000-0003-0344-5403
Joseph Edward Alderman http://orcid.org/0000-0001-8273-9009

REFERENCES

1 Lilford R, Edwards A, Girling A, *et al*. Inter-Rater reliability of case-note audit: a systematic review. *J Health Serv Res Policy* 2007;12:173–80.
2 Belur J, Tompson L, Thornton A, *et al*. Interrater reliability in systematic review methodology: exploring variation in Coder decision-making. *Sociol Methods Res* 2021;50:837–65.
3 FitzGerald C, Hurst S. Implicit bias in healthcare professionals: a systematic review. *BMC Med Ethics* 2017;18:19.
4 Gopal DP, Chetty U, O'Donnell P, *et al*. Implicit bias in healthcare: clinical practice, research and decision making. *Future Healthc J* 2021;8:40–8.

5 Banham-Hall E, Stevens S. Hindsight bias critically impacts on clinicians' assessment of care quality in retrospective case note review. *Clin Med* 2019;19:16–21.
6 Hutchinson A, Coster JE, Cooper KL, *et al*. Comparison of case note review methods for evaluating quality and safety in health care. *Health Technol Assess* 2010;14:iii-iv, ix-x, 1-144.
7 Hutchinson A, Coster JE, Cooper KL, *et al*. A structured judgement method to enhance mortality case note review: development and evaluation. *BMJ Qual Saf* 2013;22:1032–40.
8 Vollam S, Gustafson O, Young JD, *et al*. Problems in care and avoidability of death after discharge from intensive care: a multi-centre retrospective case record review study. *Crit Care* 2021;25:10.
9 Hanskamp-Sebregts M, Zegers M, Vincent C, *et al*. Measurement of patient safety: a systematic review of the reliability and validity of adverse event detection with record review. *BMJ Open* 2016;6:e011078.
10 Bion J, Aldridge C, Girling AJ, *et al*. Changes in weekend and weekday care quality of emergency medical admissions to 20 hospitals in England during implementation of the 7-day services National health policy. *BMJ Qual Saf* 2021;30:536–46.
11 Woolf H. *Access to justice: interim report to the Lord Chancellor on the civil justice system in England and Wales*. Lord Chancellor's Department, 1995.
12 Jacob R. Experts and Woolf: Have Things Got Better? In: Dwyer D, ed. *The civil procedure rules ten years on*. Oxford Scholarship Online, 2012.
13 Hofer TP, Bernstein SJ, DeMonner S, *et al*. Discussion between reviewers does not improve reliability of peer review of hospital quality. *Med Care* 2000;38:152–61.
14 Schmitt J, Roessler M, Scriba P. Effect of clinical peer review on mortality in patients ventilated for more than 24 hours: a cluster randomised controlled trial. *BMJ Qual Saf* 2023;32:17–25.
15 Skivington K, Matthews L, Simpson SA, *et al*. A new framework for developing and evaluating complex interventions: update of medical Research Council guidance. *BMJ* 2021;374:n2061.
16 Dixon-Woods M, Leslie M, Bion J, *et al*. What counts? an ethnographic study of infection data reported to a patient safety program. *Milbank Q* 2012;90:548–91.
17 Dixon-Woods M, Leslie M, Tarrant C, *et al*. Explaining matching Michigan: an ethnographic study of a patient safety program. *Implement Sci* 2013;8:70.
18 Bion J, Richardson A, Hibbert P, *et al*. 'Matching Michigan': a 2-year stepped interventional programme to minimise central venous catheter-blood stream infections in intensive care units in England. *BMJ Qual Saf* 2013;22:110–23.
19 Landoni G, Comis M, Conte M, *et al*. Mortality in multicenter critical care trials: an analysis of interventions with a significant effect. *Crit Care Med* 2015;43:1559–68.
20 Elwert F, Winship C. Endogenous selection bias: the problem of conditioning on a Collider variable. *Annu Rev Sociol* 2014;40:31–53.
21 Moran JL, Santamaria JD, Duke GJ, *et al*. Modelling Hospital outcome: problems with endogeneity. *BMC Med Res Methodol* 2021;21:124.
22 de Grooth H-J, Girbes ARJ, van der Ven F, *et al*. Observational research for therapies titrated to effect and associated with severity of illness: misleading results from commonly used statistical methods. *Crit Care Med* 2020;48:1720–8.
23 Leisman DE. The Goldilocks effect in the ICU-When the data speak, but not the truth. *Crit Care Med* 2020;48:1887–9.
24 Lilford R, Pronovost P. Using hospital mortality rates to judge hospital performance: a bad idea that just won't go away. *BMJ* 2010;340:c2016.
25 Roessler M, Schmitt J, Schoffer O. Can we trust the standardized mortality ratio? A formal analysis and evaluation based on axiomatic requirements. *PLoS One* 2021;16:e0257003.
26 Girbes ARJ, de Grooth H-J. Time to stop randomized and large pragmatic trials for intensive care medicine syndromes: the case of sepsis and acute respiratory distress syndrome. *J Thorac Dis* 2020;12:S101–9.
27 Hogan H, Healey F, Neale G, *et al*. Preventable deaths due to problems in care in English acute hospitals: a retrospective case record review study. *BMJ Qual Saf* 2012;21:737–45.
28 Manaseki-Holland S, Lilford RJ, Te AP, *et al*. Ranking hospitals based on preventable Hospital death rates: a systematic review with implications for both direct measurement and indirect measurement through standardized mortality rates. *Milbank Q* 2019;97:228–84.
29 Rodwin BA, Bilan VP, Merchant NB, *et al*. Rate of preventable mortality in hospitalized patients: a systematic review and meta-analysis. *J Gen Intern Med* 2020;35:2099–106.
30 Hogan H, Zipfel R, Neuburger J, *et al*. Avoidability of hospital deaths and association with hospital-wide mortality ratios: retrospective case record review and regression analysis. *BMJ* 2015;351:h3239.
31 Howell A-M, Burns EM, Bouras G, *et al*. Can patient safety incident reports be used to compare Hospital safety? results from a quantitative analysis of the English national reporting and learning system data. *PLoS One* 2015;10:e0144107.
32 Girling AJ, Hofer TP, Wu J, *et al*. Case-Mix adjusted hospital mortality is a poor proxy for preventable mortality: a modelling study. *BMJ Qual Saf* 2012;21:1052–6.
33 Benning A, Dixon-Woods M, Nwulu U, *et al*. Multiple component patient safety intervention in English hospitals: controlled evaluation of second phase. *BMJ* 2011;342:d199.
34 Hutchings A, Durand MA, Grieve R, *et al*. Evaluation of modernisation of adult critical care services in England: time series and cost effectiveness analysis. *BMJ* 2009;339:b4353.
35 Chen Y-F, Hemming K, Stevens AJ, *et al*. Secular trends and evaluation of complex interventions: the rising tide phenomenon. *BMJ Qual Saf* 2016;25:303–10.
36 Bellani G, Laffey JG, Pham T, *et al*. Epidemiology, patterns of care, and mortality for patients with acute respiratory distress syndrome in intensive care units in 50 countries. *JAMA* 2016;315:788–800.
37 Patel JM, Baker R, Yeung J, *et al*. Intra-Operative adherence to lung-protective ventilation: a prospective observational study. *Perioper Med* 2016;5:8.
38 Parhar KKS, Stelfox HT, Fiest KM, *et al*. Standardized management for hypoxemic respiratory failure and ARDS: systematic review and meta-analysis. *Chest* 2020;158:2358–69.
39 Levy E, Scott S, Tran T, *et al*. Adherence to lung protective ventilation in patients with coronavirus disease 2019. *Crit Care Explor* 2021;3:e0512.
40 NICE Guideline Committee. Emergency and acute medical care in over 16s: service delivery and organisation. National Institute of Health and Care Excellence guideline [NG94], 2018. Available: https://www.nice.org.uk/guidance/ng94/chapter/Recommendations#emergency-and-acute-medical-care-in-hospital

41 Tarrant C, Leslie M, Bion J, *et al*. A qualitative study of speaking out about patient safety concerns in intensive care units. *Soc Sci Med* 2017;193:8–15.
42 Knighton AJ, Kean J, Wolfe D, *et al*. Multi-factorial barriers and facilitators to high adherence to lung-protective ventilation using a computerized protocol: a mixed methods study. *Implement Sci Commun* 2020;1:67.
43 Dixon-Woods M. How to improve healthcare improvement-an essay by Mary Dixon-Woods. *BMJ* 2019;367:l5514.
44 Damschroder LJ, Aron DC, Keith RE, *et al*. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. *Implement Sci* 2009;4:50.
45 Rycroft-Malone J, Seers K, Eldh AC, *et al*. A realist process evaluation within the facilitating implementation of research evidence (fire) cluster randomised controlled international trial: an exemplar. *Implement Sci* 2018;13:138.
46 Bucknall TK, Harvey G, Considine J, *et al*. Prioritising responses of nurses to deteriorating patient observations (PRONTO) protocol: testing the effectiveness of a facilitation intervention in a pragmatic, cluster-randomised trial with an embedded process evaluation and cost analysis. *Implement Sci* 2017;12:85.
47 Bion J, Brookes O, Brown C, *et al*. A framework and toolkit of interventions to enhance reflective learning among health-care professionals: the pearl mixed-methods study. *Health Serv Deliv Res* 2020;8:1–82.