Original Article

SAGE Open Medicine

Who is watching the watchmen: Is quality reporting ever harmful?

SAGE Open Medicine
2: 2050312114523425

DOI: 10.1177/2050312114523425
smo.sagepub.com

R Scott Braithwaite[1] and Arthur Caplan[2]

Abstract

Background: Quality reporting is increasingly used as a tool to encourage health systems, hospitals, and their practitioners to deliver the greatest health benefit. However, quality reporting systems may have unintended negative consequences, such as inadvertently encouraging "cherry-picking" by inadequately adjusting for patients who are challenging to take care of, or underpowering to reliably detect meaningful differences in care. There have been no reports seeking to identify a minimum level of accuracy that ought to be viewed as a prerequisite for quality reporting.

Method: Using a decision analytic model, we seek to delineate minimal standards for quality measures to meet, using the simplest assumptions to illustrate what those standards may be.

Results: We find that even under assumptions regarding optimal performance of the quality reporting system (sensitivity and specificity of 1), we can identify a minimal level of accuracy required for the quality reporting system to "do no harm": the increase in health-related quality of life from a higher rather than lower quality practitioner must be greater than the number of practitioners per patient divided by the proportion of patients willing to switch from a lower to a higher quality provider.

Conclusion: Quality measurement systems that have not been demonstrated to improve health outcomes should be held to a specific standard of measurement accuracy.

Keywords

Quality reporting, decision analysis, quality, pay for performance, physician reporting

Date received: 17 December 2013; accepted: 16 January 2014

Quality reporting is increasingly used as a tool to encourage health systems, hospitals, and their practitioners to deliver the consistent care.[1–4] Indeed, there has been no shortage of published reports describing the possible advantages of quality reporting,[5–9] and a "report card" displaying adherence with quality measures can theoretically help consumers, payers, and employers make informed health plan choices and identify the highest quality providers.[10] Report cards could even be part of real-time feedback mechanisms incentivizing health plans to be active participants in improvement of cycle planning, doing, studying, and acting.[11–13] Quality reporting is given great importance in health reform and state- and federal-sponsored exchanges of health information.[4]

However, no measurement system is perfect. While there have been many reports describing the obvious limitations and unintended consequences of quality reporting systems (e.g. inadvertently encouraging "cherry-picking"[14] by inadequately adjusting for patients that are challenging to take care of, or underpowering to reliably detect meaningful differences in care),[15–17] when quality metrics have not been demonstrated to improve health benefits, the dangers of harms from unintended consequences loom particularly large. We seek to delineate minimal standards for quality measures to meet, using the simplest assumptions to illustrate what those standards may be.

Development of a criterion

We believe a minimal moral requirement for any quality measurement ought to be based on the normative requirement exhibited in the Hippocratic Oath and many other codes of ethics of "first do no harm." Accordingly, we lay out minimal metrics for quality measures to meet and using the simplest assumptions to illustrate what those points are.

[1]Division of Comparative Effectiveness and Decision Science, Department of Population Health, School of Medicine, New York University, New York, NY, USA
[2]Division of Bioethics, Department of Population Health, School of Medicine, New York University, New York, NY, USA

Corresponding author:
R Scott Braithwaite, Division of Comparative Effectiveness and Decision Science, Department of Population Health, School of Medicine, New York University, 550 First Avenue, VZ30 6th Floor, 615, New York, NY 10016, USA.
Email: Scott.Braithwaite@nyumc.org

One simple way to operationalize the principle of "doing no harm" is to ensure that expected utility is not decreased. Accordingly, a criterion for quality measurement ought not to decrease expected utility and, ideally, should increase it. Other simplifying assumptions can be identified that may facilitate our objective, regarding characteristics of a particular health system, its patients, and practitioners:

1. The relevant population consists of patients served by the health system and the practitioners who are being subjected to the quality standard.
2. The quality measurement system will partition all practitioners into a higher performing subgroup and a lower-performing subgroup. While it is common and desirable for quality rating systems to have many more stratifications, we choose a simple partition for the sake of illustration, and with the explicit understanding that this model can be generalized to more complicated systems.
3. Any measurement system can be characterized in terms of its operating characteristics (e.g. sensitivity and specificity, true positives, and false positives). A perfect classification system would have a sensitivity and specificity of 1, with all true positives and no false positives. Inadequate risk adjustment and/or insufficient statistical power will be reflected in subpar sensitivity and specificity (e.g. a practitioner who is reported falsely as exhibiting negative quality because of insufficient consideration of her willingness to take care of poor adherers). It is incumbent upon the creator of the measurement system to seek to optimize its performance characteristics.
4. In the long-term, there will be realignment of the supply of practitioners and the demand by patients for those practitioners as a result of the quality measurement system. The realignment will not be perfect. Some people really like their doctor or have insurmountable barriers to choosing alternative doctors and will stick with them regardless of what a quality measurement system recommends. However, in the long-term, some realignment will occur and will extend over a sufficiently long duration so that its impacts will dwarf short-term effects.[18] Accordingly, consequences of true positives from the quality measurement system will be that some additional patients will seek out and intentionally receive above-average care, whereas consequences of false positives will be that some additional patients will seek out and unintentionally receive below-average care.
5. The quality metric under consideration has not yet been demonstrated to improve health outcomes.
6. Additional assumptions include the following: (a) truly labeling a practitioner as low quality may transiently lower practitioner utility, but this utility will revert to baseline with rapidity that is insignificant for our analysis; (b) falsely labeling a practitioner as high quality will transiently raise practitioner utility, but this utility will revert to baseline with a rapidity that is insignificant for our analysis; and (c) truly labeling a practitioner as high quality may raise practitioner utility (e.g. due to enhanced self-regard and pride), but this elevation will be sufficiently small to be ignored. Falsely labeling a practitioner as low quality will lower utility by a substantial amount. We tested decrements ranging from 0 (base case analysis) to 0.5 (consistent with other negative transformative events).[19]

In order for a quality reporting system to do no harm, the overall utility after adoption of the quality rating system needs to be at least as great as the overall utility before adoption of the quality rating system. In accord with the above assumptions, standard decision analytic methods can be used to create a decision tree comparing the expected value of "use of a quality measurement" system to the expected value of "no use of a quality measurement system" (Figure 1). This decision tree can then be used to identify the circumstances under which expected utility would be expected to increase by using a quality measurement system.

Results

Assuming that a quality measurement system has perfect performance characteristics (sensitivity of 1 and specificity of 1) and assuming a large loss of utility for a provider who is labeled as "low quality," we can identify a particular criterion in order for this measure to "do no harm": the increase in health-related quality of life that would occur as a consequence of correctly identifying a higher rather than lower quality practitioner must be greater than the number of practitioners per patient divided by the proportion of patients willing to switch from a lower to a higher quality provider. For example, if there is 0.001 practitioners for every patient (corresponding to a practice panel of 1000 for a primary care physician), and the proportion of patients who would switch from a low-quality provider to a high-quality provider based on information from the quality measurement system is 0.1 (meaning that 10% would switch), this would imply that higher quality practitioners would need to confer a health-related quality-of-life improvement of at least 0.01 utility units compared to the lower quality practitioners (0.001 divided by 0.1).

In sensitivity analyses in which we assume that practitioners sustain more mild decrements in utility from being labeled as "low quality," results are still notable, with higher quality practitioners needing to confer a health-related quality-of-life improvement of at least 0.005 utility units compared to the lower-quality practitioners.

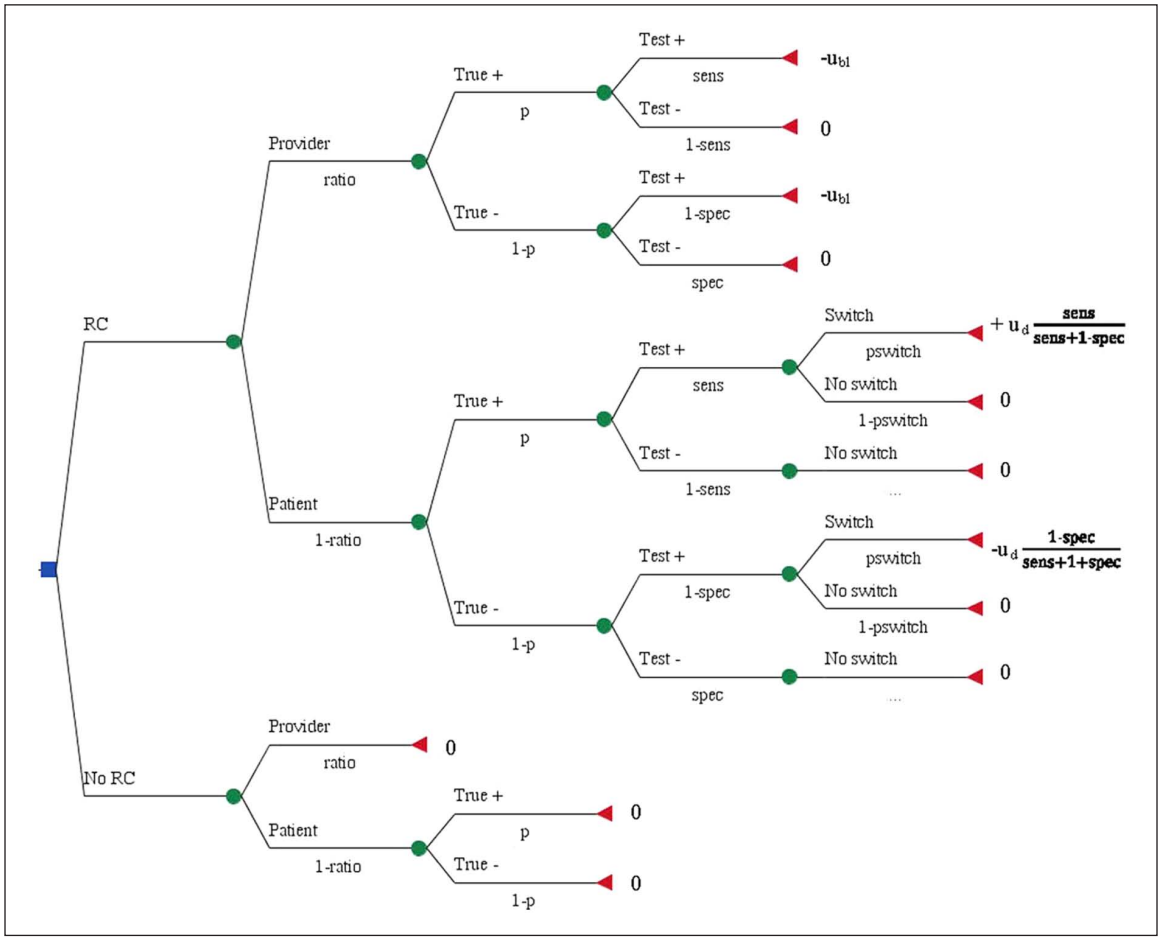

Figure 1. Decision tree and relevant calculations. The decision tree reads from left to right and represents possible pathways through the model. The square node at the left of the diagram is a "choose" node, representing the choice of using a quality reporting system (Report Card (RC)) or not using a quality reporting system (No RC). The circles at the origin of each branch are chance nodes, representing events that may or may not occur with a specified probability, depending on the use of RCs. The relevant population consists of providers and patients. Under the RC scenarios, providers are either in a high performing group (True+) or a lower-performing group (True−). However, the RC can either categorize that provider as high performing (Test+) or low performing (Test−).

Ratio: ratio of practitioners to providers; *p*: probability that a provider is high quality; *sens*: sensitivity of the RC; *spec*: specificity of the RC; *pswitch*: proportion of patients who would switch from a low-quality provider to a high-quality provider based on information from the RC; μ_{bl}: baseline utility; μ_d: change from baseline utility (magnitude of improvement in health-related quality of life) that would be expected to result from changing to a higher-quality provider. True positive is $TP = [sens / (sens + 1 - spec)]$. False positive is $FP = [1 - spec / (sens + 1 - spec)]$.

Solving the tree, we start with

$$-ratio * \mu_{bl} * [p * sens + (1-p)(1-spec)] + (1-ratio) * p * sens * pswitch * \mu_d * TP \times (1-ratio) * (1-p) * (1-spec) * pswitch * -\mu_d FP$$

This reduces to $-ratio * \mu_{bl} * [p * sens + (1-p)(1-spec)] + (1-ratio) * pswitch * \mu_d * [p * sens * TP - (1-p) * (1-spec) * FP]$

This reduces to $\mu_d > ratio / 1 - ratio * 1 / pswitch * \mu_{bl} * [p * sens + (1-p) * (1-spec)]^2 / (p * sens)^2 + [(1-p) * (1-spec)]^2$

Making the following assumptions: $sens = 1, spec = 1, \mu_{bl} = 1, p = 0.5$, the equation further simplifies to $\mu_d > ratio / [(1-ratio) * pswitch]$.

Discussion

This result has important implications for developers of quality measurement systems, for consumer groups, health systems, and payers: in situations in which implementation of the quality metric has not been demonstrated to improve health outcomes, the onus should be on proponents of using that metric to demonstrate that it "does no harm," and this evaluation can be accomplished based on few inputs (Table 1): the proportion of patients who would switch practitioners, the ratio of practitioners to patients, and the magnitude of improvement in health-related quality of life that would be expected to result from higher rather than lower quality practitioners (represented as μ_d in Figure 1). The proportion of patients who would switch

Table 1. Calculations of how much of an improvement in health-related quality of life for higher versus lower quality practitioners would be necessary in order for the quality rating to "do no harm." Note that these calculations assume the ideal scenario of a quality reporting system with a sensitivity and specificity of 1 of correctly identifying higher and lower quality practitioners. Health-related quality of life is expressed in terms of utility units, a preference-weighted quality-of-life metric on a scale of 0 (worst) to 1 (best).

Number of patients per practitioner	Proportion of patients willing to switch practitioners based on quality data (%)	Minimum increase in health-related quality of life between higher and lower quality physicians necessary to avoid doing harm
200	5	0.10
500	5	0.04
1000	5	0.02
2000	5	0.01
200	10	0.05
500	10	0.02
1000	10	0.01
2000	10	0.005
200	20	0.025
500	20	0.01
1000	20	0.005
2000	20	0.0025

practitioners can be estimated via surveys, the ratio of practitioners to patients is well known by any health system, and the magnitude of improvement in health-related quality of life can be estimated using validated models.[20]

While a 0.01 increment in utility may seem like a very small number intuitively, it is important to note that utility is an overall health-related quality-of-life measure rather than a disease-related quality-of-life measure. For this reason, very few improvements in medical care produce substantially large changes in utility once averaged across the entire population. For example, if higher quality practitioners improved pain control for 10% of their population with chronic pain, and if this lowering resulted in an improvement of 0.03 utility units for those patients who are affected (a typical minimum change reflecting clinical significance),[21] then the higher quality practitioners would be increasing overall utility across their panel by 0.003 utility units, which would be insufficient to meet the criterion for quality measurement.

Our base case calculations make the optimistic assumption that the quality measurement system has perfect sensitivity and specificity. In truth, sensitivity and specificity of many quality measurement systems will be far below 1 because of many factors, including inadequate risk adjustment[22,23] and insufficient statistical power. However, assumptions of perfect sensitivity and specificity often yield useful bounding analyses (i.e. if a quality reporting system causes harm even under the idealized assumption of perfect performance characteristics, it would also cause harm under actual performance characteristics). Additionally, the base case calculation is not grossly inaccurate even with more realistic sensitivity and specificity estimates. Across a wide range of sensitivity and specificity assumptions, the minimum difference in health for high- versus low-quality providers would vary between one and two times the number of practitioners per patient divided by the proportion of patients willing to switch from a lower to a higher quality provider (Table 1).

Limitations

We seek to delineate minimal standards for quality measures to meet, using the simplest assumptions to illustrate what those standards are. Sensitivity and specificity in real life will be lesser than 1 and may be difficult to estimate because of ambiguity regarding the best gold standard;[24] it might not always be necessary to do so. If a quality reporting system would cause harm even under the idealized assumption of perfect performance characteristics, it would also cause harm under actual performance characteristics.

Quality metrics should be adjusted for those patient characteristics over which the practitioner and/or health system has locus of control, but not those characteristics over which the practitioner and/or health system does not have locus of control.[25] If this principle is disregarded, risk adjustment degenerates into a logistical rather than a scientific discussion, focused on the question of what data are routinely available for risk adjustment, rather than the question of the data's suitability, completeness for risk adjustment, or position in the causal pathway of quality of care.[26] Indeed, these and other principles in quality metric formulation have been well described, and disregarding them out of convenience (e.g. using what data are available even if other unavailable data are important) merely increases the likelihood of doing harm.

Consequently, it can be argued that practitioners who are going to be subject to a quality measurement themselves ought to make a list of patient characteristics that are likely to be associated with the quality outcome of interest and

that peers regard as being within their locus of control, and these characteristics should be used as the adjustors.[27] A fair, explicit, and transparent procedure such as this not only reduces the likelihood that a quality metric may cause harm but also may encourage "buy in" from practitioners themselves.

Other limitations of this approach to measuring quality involve an explicit consideration of the well-being of practitioners as well as the well-being of patients. It may be argued that health systems should only be concerned with optimizing the health of their subscribers. However, this is a shortsighted perspective. Practitioner noncompliance and burnout will ultimately have pernicious effects on the health system overall.

Finally, it can be argued that our approach is too simple, merely dividing practitioners into two strata, one of higher performers and one of lower performers. However, our approach can be applied to more sophisticated quality measurement systems and stratifications, albeit with a commensurate increase in mathematical complexity.

Conclusion

Quality measurement systems that have not been demonstrated to improve health outcomes should be held to a specific standard of measurement accuracy. The hypothesized benefit in quality of life resulting from the higher quality outcomes should exceed the number of practitioners per patient divided by the proportion of patients willing to switch from a lower to a higher quality provider. However, the most important reason to develop such a standard is to hold those who seek to measure the performance of health-care providers to the same standard demanded of the practitioners themselves—do no harm.

Declaration of conflicting interests

The authors declare no conflict of interest in preparing this article.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

1. Christianson JB, Volmar KM, Alexander J, et al. A report card on provider report cards: current status of the health care transparency movement. *J Gen Intern Med* 2010; 25: 1235–1241.
2. Sinaiko AD, Eastman D and Rosenthal MB. How report cards on physicians, physician groups, and hospitals can have greater impact on consumer choices. *Health Aff (Millwood)* 2012; 31: 602–611.
3. Krumholz HM, Anderson JL, Bachelder BL, et al. ACC/AHA 2008 performance measures for adults with ST-elevation and non-ST-elevation myocardial infarction: a report of the American College of Cardiology/American Heart Association Task Force on Performance Measures (Writing Committee to develop performance measures for ST-elevation and non-ST-elevation myocardial infarction): developed in collaboration with the American Academy of Family Physicians and the American College of Emergency Physicians: endorsed by the American Association of Cardiovascular and Pulmonary Rehabilitation, Society for Cardiovascular Angiography and Interventions, and Society of Hospital Medicine. *Circulation* 2008; 118: 2596–2648.
4. US Congress. Patient Protection and Affordable Care Act (2010), http://www.gpo.gov/fdsys/pkg/BILLS-111hr3590enr/pdf/BILLS-111hr3590enr.pdf (2012).
5. Shahian DM, Meyer GS, Mort E, et al. Association of National Hospital Quality Measure adherence with long-term mortality and readmissions. *BMJ Qual Saf* 2012; 21: 325–336.
6. Tu JV, Khalid L, Donovan LR, et al. Indicators of quality of care for patients with acute myocardial infarction. *CMAJ* 2008; 179: 909–915.
7. Kerr EA and Fleming B. Making performance indicators work: experiences of US Veterans Health Administration. *BMJ* 2007; 335: 971–973.
8. Pronovost PJ and Holzmueller CG. Partnering for quality. *J Crit Care* 2004; 19: 121–129.
9. Pronovost PJ, Nolan T, Zeger S, et al. How can clinicians measure safety and quality in acute care? *Lancet* 2004; 363: 1061–1067.
10. Henderson A. Surgical report cards: the myth and the reality. *Monash Bioeth Rev* 2009; 28: 20.1–20.20.
11. Curran ET and Bunyan D. Using a PDSA cycle of improvement to increase preparedness for, and management of, norovirus in NHS Scotland. *J Hosp Infect* 2012; 82: 108–113.
12. Koetsier A, Van der Veer SN, Jager KJ, et al. Control charts in healthcare quality improvement. A systematic review on adherence to methodological criteria. *Methods Inf Med* 2012; 51: 189–198.
13. White CM, Statile AM, Conway PH, et al. Utilizing improvement science methods to improve physician compliance with proper hand hygiene. *Pediatrics* 2012; 129: e1042–e1050.
14. Resnick DK, Rajpal S and Steinmetz MP. Common pitfalls in interpretation of medical evidence: a case demonstration of misleading interpretation in the analysis of cervical spine fusions. *Spine J* 2009; 9: 905–909.
15. Schold JD, Srinivas TR, Howard RJ, et al. The association of candidate mortality rates with kidney transplant outcomes and center performance evaluations. *Transplantation* 2008; 85: 1–6.
16. Schold JD. Evaluation criteria for report cards of healthcare providers. *Adv Health Econ Health Serv Res* 2008; 19: 173–189.
17. Gajewski BJ, Mahnken JD and Dunton N. Improving quality indicator report cards through Bayesian modeling. *BMC Med Res Methodol* 2008; 8: 77.
18. Wang J, Hockenberry J, Chou SY, et al. Do bad report cards have consequences? Impacts of publicly reported provider quality information on the CABG market in Pennsylvania. *J Health Econ* 2011; 30: 392–407.
19. Tengs TO and Wallace A. One thousand health-related quality-of-life estimates. *Med Care* 2000; 38: 583–637.

20. Eddy DM, Adler J and Morris M. The "Global Outcomes Score": a quality measure, based on health outcomes, that compares current care to a target level of care. *Health Aff (Millwood)* 2012; 31: 2441–2450.
21. Walters SJ and Brazier JE. Comparison of the minimally important difference for two health state utility measures: EQ-5D and SF-6D. *Qual Life Res* 2005; 14: 1523–1532.
22. Shahian DM, Normand SL, Torchiana DF, et al. Cardiac surgery report cards: comprehensive review and statistical critique. *Ann Thorac Surg* 2001; 72: 2155–2168.
23. Iezzoni LI. Risk adjusting rehabilitation outcomes: an overview of methodologic issues. *Am J Phys Med Rehabil* 2004; 83: 316–326.
24. Timmermans S and Berg M. *The gold standard: the challenge of evidence-based medicine*. Philadelphia, PA: Temple University Press, 2003.
25. Wallston KA, Wallston BS and DeVellis R. Development of the multidimensional health locus of control (MHLC) scales. *Health Educ Monogr* 1978; 6: 160–170.
26. Myers RP, Hubbard JN, Shaheen AA, et al. Hospital performance reports based on severity adjusted mortality rates in patients with cirrhosis depend on the method of risk adjustment. *Ann Hepatol* 2012; 11: 526–535.
27. Stanley R, Lillis KA, Zuspan SJ, et al. Development and implementation of a performance measure tool in an academic pediatric research network. *Contemp Clin Trials* 2010; 31: 429–437.