What is a performance outlier?

David M Shahian,[1,2] Sharon-Lise T Normand[3,4]

[1]Department of Surgery, Center for Quality and Safety, Massachusetts General Hospital, Boston, Massachusetts, USA
[2]Harvard Medical School, Boston, Massachusetts, USA
[3]Department of Health Care Policy, Harvard Medical School, Boston, Massachusetts, USA
[4]Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts, USA

Correspondence to
Dr David M Shahian, Department of Surgery, Center for Quality and Safety, Massachusetts General Hospital, 55 Fruit St, Boston, MA 02114, USA;
dshahian@partners.org

Accepted 5 January 2015

Linked

▶ http://dx.doi.org/10.1136/bmjqs-2014-003405

To cite: Shahian DM, Normand S-LT. *BMJ Qual Saf* 2015;**24**:95–99.

Healthcare performance measurement is a complex undertaking, often presenting a number of potential alternative approaches and methodological nuances. Important considerations include richness and quality of data sources; data completeness; choice of metrics and target population; sample size; patient- and provider-level data collection periods; risk adjustment; statistical methodology (eg, logistic regression vs hierarchical models); model performance, reliability and validity; and classification of outliers. Given these many considerations, as well as the absence of nationally accepted standards for provider profiling, it is not surprising that different rating organisations and methodologies may produce divergent results for the same hospitals.[1–6]

Outlier classification, the last step in the measurement process, has particularly important ramifications. For patients, it may lead them to choose or avoid a particular provider. For providers, outlier status may positively or negatively impact referrals and reimbursement, and may influence how scarce hospital resources are deployed to address putative areas of concern. Misclassification is probably more common than generally appreciated. For example, partitioning of hospitals (eg, terciles, quartiles, quintiles, deciles) to determine outliers may lead to excessive false positives—hospitals labelled as having above or below average performance when, in fact, their results do not differ significantly from the mean based on appropriate statistical tests.[7 8]

THE CURRENT STUDY

In this issue, Paddock *et al*[9] address a seemingly straightforward question—what precisely does it mean to be a performance outlier? Using *Hospital Compare* data, the authors demonstrate an apparently contradictory finding. When *directly* compared one to another, some individual hospitals in a given performance tier may not be statistically significantly different than individual hospitals in adjacent tiers, even when those tier assignments were made using appropriate tests of statistical significance. For instance, Paddock *et al*[9] show that for each bottom-tier hospital, there was at least one mid-tier hospital with statistically indistinguishable performance. Among mid-tier ('average') hospitals, 60–75% had performance that was not statistically significantly different than that of some bottom-tier hospitals.

How can this be? On the one hand, hospitals appear to have been appropriately divided into three discrete groups based on their performance rankings—bottom, mid and top tiers. On the other hand, direct comparisons between specific pairs of hospitals in adjacent tiers often showed no statistically significant difference, which seems inconsistent with their original rankings The answer to this apparent paradox illustrates several statistical concepts, some unfamiliar to non-statisticians but of fundamental importance to the correct interpretation of risk-adjusted outcomes and outlier status.

First and most fundamentally, Paddock *et al*[9] use a completely different statistical methodology for their direct hospital to hospital comparisons than the approach used in the original *Hospital Compare* tier assignments.[10–12] The latter employed Bayesian hierarchical regression models with 95% credible intervals (similar to CIs) to determine outliers. From the perspective of causal inference theory,[13–17] the *Hospital Compare* approach considers the following unobservable counterfactual: 'What would the results have been if this hospital's patients had been cared for by an "average" hospital in the reference population?' This is often referred to as the 'expected' outcome. A level of statistical certainty for the hospital-level estimates is chosen (eg, 95% credible interval), the actual results of a given hospital are compared to the expected or counterfactual outcomes, and any hospital whose 95% credible interval for their risk-adjusted mortality rate excludes the expected mortality rate is designated an outlier.

Because Paddock *et al*[9] did not have access to the patient-level data on which

the *Hospital Compare* analyses were based, they first converted CIs to SEs, then re-estimated performance tiers (presumably, though not stated, using one-sample z-tests), which were similar to the original *Hospital Compare* ratings. Finally, they performed two-sample z-tests using the results from various hospital combinations in adjacent performance tiers. Their counterfactual is not the expected outcome if a hospital's patients were cared for by an average hospital, but rather by one specific alternative hospital. Their corresponding null hypothesis is that the difference in mean mortality rates between the two hospitals being compared is zero (or, alternatively, that the ratio of their mean mortality rates is unity).

Thus, the direct hospital–hospital comparisons performed by Paddock *et al*[9] ask a different question than the original *Hospital Compare* analyses, with a different counterfactual statement and statistical approach. Viewed from this perspective, it is no longer paradoxical but completely logical that they found different results. In this particular study, failure to reject the null hypothesis of no difference in performance among pairs of hospitals from adjacent tiers was also driven by the large SEs (resulting from small hospital sample sizes—see below). Indistinguishable performance would be particularly likely for pairs of hospitals whose performance was close to the boundary between two adjacent performance categories. That the authors only required at least one hospital from an adjacent tier to be statistically indistinguishable is a relatively low bar.

DIRECT AND INDIRECT STANDARDISATION

Notwithstanding the results from this specific study, which are largely a function of small sample sizes, the authors do not address the more fundamental error of using indirectly standardised results to directly compare pairs of hospitals. The differences between direct and indirect standardisation[13 17 18] remain unappreciated by most non-methodologists, resulting in their frequent misapplication and misinterpretation. In direct standardisation, rates from each stratum of the study population are applied to a reference population. This type of standardisation is common in epidemiological studies where there may be only a few strata of interest (eg, age–sex strata). Directly standardised results estimate what the outcomes would have been in the reference population if these patients had been cared for by a particular study hospital. In causal inference terminology, this is the unobservable counterfactual. The results from many different hospitals can be applied to the reference population in exactly the same fashion, and it is therefore permissible to directly compare their directly standardised results.

The conditions that make direct standardisation possible are not found in most profiling applications because of the large number of risk factors and the fact that any given hospital may have no observations for patients having certain types of risk factors. Consequently, in virtually all healthcare profiling applications, risk adjustment is performed using indirect rather than direct standardisation. The incremental risks associated with each predictor variable (eg, a risk factor such as insulin-dependent diabetes) are derived from the reference population using regression. As in the original *Hospital Compare* approach discussed above, the expected outcomes in the study population reflect the anticipated results if those patients had been cared for by an average hospital in the reference population, a quite different counterfactual than in direct standardisation.[17] Expected results for each patient of a given hospital are summed and compared with their observed results to estimate an O/E ratio (eg, standardised mortality ratio), which can be multiplied by the average mortality to yield a risk-adjusted or risk-standardised rate.

COVARIATE OVERLAP

Direct hospital–hospital comparisons using indirectly standardised observational data are inappropriate in virtually all profiling scenarios. The only exception is a very specific circumstance—when all regions of the covariate space defined by patient risk factors contain observations from all hospitals being compared—which would be an uncommon and chance occurrence in most profiling applications.[17 19] In the absence of covariate overlap, there may be patients from one hospital for whom there are no comparable patients in the other hospital (in causal inference parlance, there is no empirical counterfactual[19]), and thus no way to fairly compare performance in all patients cared for by the two hospitals. For example, it is unlikely that each hospital would have octogenarians with renal failure and chronic liver disease who underwent emergency aortic valve replacement (AVR), but one of them might. No adjustment (eg, model-based extrapolation) can reliably remedy the lack of data in the area of non-overlap, and statistical inferences should generally be limited to regions where there is overlap.

Thus, 'risk-adjusted' results derived using indirect standardisation cannot be used to directly compare two hospitals unless their patient mix has been demonstrated to be similar (eg, overlapping propensity score distributions).[17] Indirectly standardised rates for each hospital are estimated only for the patients they actually treated, and their results only apply to their particular case mix. It cannot be assumed that a hospital achieving better than average results in a generally low risk population could do the same in a population of very high risk patients that it has never treated. Because their indirectly standardised rates were obtained by applying reference population rates to their low risk patients, assuming that they would have similar performance if confronted with a high-risk, tertiary patient population is optimistic and unwarranted.

COVARIATE IMBALANCE AND BIAS

Irrespective of whether there is *overlap* in their respective distributions of patient risk, these distributions may still vary across hospitals being compared (ie, the prevalence of relevant risk factors may be different) and this *covariate imbalance*[19] may bias the interpretation of results and the determination of outliers.[20] Covariate imbalance is a common problem in profiling using observational data because patients are not randomised (the method used to achieve covariate balance in clinical trials). Standard regression-based adjustment may not completely address bias when there is substantial lack of covariate balance. Covariate imbalance was the motivation for the development of propensity score approaches for matching, modelling or stratification in studies using observational data,[21 22] and propensity approaches to profiling have been investigated.[20]

CASE MIX BIAS

Despite excellent patient-level risk adjustment, substantial case mix bias (eg, due to marked differences in the distributions of high and low risk cases between hospitals) may be present and may impact performance estimates and outlier status. For example, the target population (condition or procedure) may be very broadly defined, which is usually done in an effort to increase sample size. Instead of focusing only on isolated aortic valve replacement (AVR), a relatively homogeneous cohort, measure developers may include all patients with an AVR, even when this procedure has been combined with other operations (such as simultaneous coronary artery bypass grafting surgery).[7] These combined procedures generally are associated with higher average mortality than their corresponding isolated procedures, so the resulting study population will have a heterogeneous range of expected mortality rates. Sometimes, completely dissimilar conditions or procedures with quite different inherent risk are aggregated into a heterogeneous composite measure to increase sample size or to give the appearance of being broadly representative. For example, the hospital standardised mortality ratio (HSMR) encompasses nearly all of the admissions at a given hospital.[4]

In all these examples, even with perfect patient-level risk adjustment, comparisons among providers may be biased and inaccurate unless differences in the relative distributions of higher and lower risk cases are properly accounted for,[4 23] a profiling analogue of Simpson's paradox.[24 25] The impact of this phenomenon is not uniform. Centres performing a greater proportion of more complex cases, with higher inherent risk of adverse outcomes, may falsely appear to have worse results.

THE LIMITATION OF SMALL SAMPLE SIZE

Small sample sizes are common in provider profiling, and this makes it difficult to reliably differentiate hospital performance and classify outliers. In a study of major surgical procedures, Dimick *et al*[26] found that only coronary artery bypass grafting surgery was performed with sufficient volume by most providers to reliably allow detection of a doubling of mortality rate. Krell *et al*[27] found that most surgical outcomes measures estimated from the American College of Surgeons' National Surgical Quality Improvement Program (NSQIP) registry data had low reliability to detect performance differences for common procedures. Similar findings have been observed with common medical diagnoses.[28–30]

At volumes typically encountered in practice, and even assuming perfect patient-level risk adjustment, much of the variation in healthcare performance measures is random; the extent of random variation and potential misclassification is greater at lower volumes and event rates.[31] As a consequence, there is substantial fluctuation from one sampling period to another in the rates of adverse events and performance rankings among providers.[32] Longitudinal assessment of provider performance over longer periods of time and investigation of trends are more prudent approaches than relying on results in one sampling period.[33]

Different approaches have been used to address the limitations of small sample size in provider profiling and outlier classification. These include establishing lower limits for sample size below which estimates are not calculated; collecting provider data over longer time periods to increase the number of observations; broadening the target population inclusion criteria (although this may lead to aggregation issues discussed previously, including ecological bias); attribution of results to larger units (eg, hospitals rather than individual physicians); and use of composite measures that effectively increase the number of endpoints.[34] Many statisticians also advocate the use empirical Bayes or fully Bayesian approaches which shrink sample estimates towards the population mean.[35–38] This yields more accurate estimates of true underlying performance, with less chance of false positive outliers, especially in small samples.

STATISTICAL CERTAINTY

Closely related to these sample size concerns is the degree of statistical certainty chosen to classify a hospital as an outlier (eg, 90%, 95%, 99% CI). The overall health policy 'costs' of higher specificity and fewer false outliers versus higher sensitivity and more false outliers must be considered, and there is no one correct answer.[39] Furthermore, the p values and CIs from traditional frequentist approaches may sometimes be misleading. With very small sample sizes, virtually no provider can be reliably identified as an outlier; conversely, with very large sample sizes, outlying results identified by statistical criteria may have little practical difference from the average. Bayesian approaches may provide more intuitive interpretation, such as estimating the probability that a hospital's performance exceeds some threshold.[36 40–42]

THE IMPACT OF OUTLIERS ON THE REFERENCE POPULATION ('EXPECTED' VALUES)

Additional problems with outlier classification can arise if the expected outcome for a particular provider is derived from a relatively small reference population (eg, the cardiac surgery programmes in a particular state). Every provider's outcomes impact not only their own observed value but also the 'expected' value (the E in O/E) for their programme, which is based on the reference population to which they belong.[13] A substantially aberrant result from one or two providers will expand the range of values that are considered average, and will reduce the likelihood of a truly abnormal outlying provider being correctly classified as such. Several approaches to this problem have been suggested, including replication with posterior predicted p values, and leave-one-out cross validation, in which the expected performance for each hospital is estimated from a model developed from all other hospitals.[37]

GRAPHICAL TOOLS FOR OUTLIER DETECTION

Finally, various graphical methods have also been used to monitor healthcare performance and to determine outliers. These include funnel plots,[43 44] in which unadjusted or adjusted point estimates of provider performance are plotted against sample size (volume), with superimposed CIs around the population average to indicate warning or outlier status. Other methods include real time graphical monitoring using cumulative sum (CUSUM) approaches, in which results are immediately updated with each patient or procedure.[45–47]

CONCLUSION

Outlier determination, the final step in the performance measurement process, is a more complicated undertaking than most non-experts appreciate, with many nuances in implementation and interpretation. Those involved in provider profiling have a responsibility to explicitly state the approaches they use for outlier classification, and to explain the proper interpretation of outlier status to end users of varying statistical sophistication. For example, as demonstrated in the study of Paddock *et al*,[9] it should be recognised that while the CMS website is named *Hospital Compare*, the statistically valid comparison is between each hospital and a hypothetical average hospital, not between pairs of hospitals.

Given the historical lack of comparative performance data in healthcare and the urgent need to foster informed consumer choice and performance improvement, it is understandable that various stakeholders (patients, payers, regulators) might be tempted to view the issue of outliers too simplistically, sometimes misinterpreting or unintentionally misusing outlier results. However, this may lead to consequences that are at least as undesirable as having no performance data at all. Misclassification of providers may misdirect consumers, unfairly discredit or commend certain providers, and lead to misallocation of scarce resources. Scientific rigour and sound judgment are required to accurately classify outliers and to constructively use this information to improve healthcare quality.

Competing interests None.

Provenance and peer review Not commissioned; internally peer reviewed.

REFERENCES

1 Healthcare Association of New York State. HANY's report on report cards: understanding publicly reported hospital quality measures. 2013. http://www.hanys.org/quality/data/report_cards/2013/docs/2013_hanys_report_card_book.pdf (accessed 5 Feb 2014).
2 Leonardi MJ, McGory ML, Ko CY. Publicly available hospital comparison web sites: determination of useful, valid, and appropriate information for comparing surgical quality. *Arch Surg* 2007;142:863–8.
3 Rothberg MB, Morsi E, Benjamin EM, *et al*. Choosing the best hospital: the limitations of public quality reporting. *Health Aff (Millwood)* 2008;27:1680–7.
4 Shahian DM, Wolf RE, Iezzoni LI, *et al*. Variability in the measurement of hospital-wide mortality rates. *N Engl J Med* 2010;363:2530–9.
5 Delong ER, Peterson ED, DeLong DM, *et al*. Comparing risk-adjustment methods for provider profiling. *Stat Med* 1997;16:2645–64.
6 Peterson ED, Delong ER, Muhlbaier LH, *et al*. Challenges in comparing risk-adjusted bypass surgery mortality results: results from the Cooperative Cardiovascular Project. *J Am Coll Cardiol* 2000;36:2174–84.
7 Shahian DM, He X, Jacobs JP, *et al*. Issues in quality measurement: target population, risk adjustment, and ratings. *Ann Thorac Surg* 2013;96:718–26.
8 Bilimoria KY, Cohen ME, Merkow RP, *et al*. Comparison of outlier identification methods in hospital surgical quality improvement programs. *J Gastrointest Surg* 2010;14:1600–7.
9 Paddock SM, Adams JL, Hoces de la Guardia F. Better-than-average and worsethan-average hospitals may not significantly differ from average hospitals: an analysis of Medicare Hospital Compare ratings. *BMJ Qual Saf* 2015;24:128–34.
10 Bratzler DW, Normand SL, Wang Y, *et al*. An administrative claims model for profiling hospital 30-day mortality rates for pneumonia patients. *PLoS ONE* 2011;6:e17401.
11 Krumholz HM, Wang Y, Mattera JA, *et al*. An administrative claims model suitable for profiling hospital performance based on 30-day mortality rates among patients with an acute myocardial infarction. *Circulation* 2006;113:1683–92.
12 Krumholz HM, Wang Y, Mattera JA, *et al*. An administrative claims model suitable for profiling hospital performance based on 30-day mortality rates among patients with heart failure. *Circulation* 2006;113:1693–701.
13 Draper D, Gittoes M. Statistical analysis of performance indicators in UK higher education. *J R Stat Soc Series A (Statistics in Society)* 2004;167:449–74.
14 Maldonado G, Greenland S. Estimating causal effects. *Int J Epidemiol* 2002;31:422–9.
15 Little RJ, Rubin DB. Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annu Rev Public Health* 2000;21:121–45.

16 Pearl J. *Causality: models, reasoning, and inference*. Cambridge, UK; New York: Cambridge University Press, 2000.
17 Shahian DM, Normand SL. Comparison of "risk-adjusted" hospital outcomes. *Circulation* 2008;117:1955–63.
18 Fleiss JL, Levin BA, Paik MC. *Statistical methods for rates and proportions*. Hoboken, NJ: J. Wiley, 2003.
19 Gelman A, Hill J. *Data analysis using regression and multilevel/hierarchical models*. Cambridge; New York: Cambridge University Press, 2007.
20 Huang IC, Frangakis C, Dominici F, *et al*. Application of a propensity score approach for risk adjustment in profiling multiple physician groups on asthma care. *Health Serv Res* 2005;40:253–78.
21 Rosenbaum PR. *Observational studies*. New York: Springer, 2002.
22 D'Agostino RB Jr. Propensity scores in cardiovascular research. *Circulation* 2007;115:2340–3.
23 Glance LG, Osler TM. Comparing outcomes of coronary artery bypass surgery: is the New York Cardiac Surgery Reporting System model sensitive to changes in case mix? *Crit Care Med* 2001;29:2090–6.
24 Manktelow BN, Evans TA, Draper ES. Differences in case-mix can influence the comparison of standardised mortality ratios even with optimal risk adjustment: an analysis of data from paediatric intensive care. *BMJ Qual Saf* 2014;23:782–8.
25 Marang-van de Mheen PJ, Shojania KG. Simpson's paradox: how performance measurement can fail even with perfect risk adjustment. *BMJ Qual Saf* 2014;23:701–5.
26 Dimick JB, Welch HG, Birkmeyer JD. Surgical mortality as an indicator of hospital quality: the problem with small sample size. *JAMA* 2004;292:847–51.
27 Krell RW, Hozain A, Kao LS, *et al*. Reliability of risk-adjusted outcomes for profiling hospital surgical quality. *JAMA Surg* 2014;149:467–74.
28 Hofer TP, Hayward RA. Identifying poor-quality hospitals. Can hospital mortality rates detect quality problems for medical diagnoses? *Med Care* 1996;34:737–53.
29 Hofer TP, Hayward RA, Greenfield S, *et al*. The unreliability of individual physician "report cards" for assessing the costs and quality of care of a chronic disease. *JAMA* 1999;281:2098–105.
30 Thomas JW, Hofer TP. Accuracy of risk-adjusted mortality rate as a measure of hospital quality of care. *Med Care* 1999;37:83–92.
31 Austin PC, Reeves MJ. Effect of provider volume on the accuracy of hospital report cards: a Monte Carlo study. *Circ Cardiovasc Qual Outcomes* 2014;7:299–305.
32 Siregar S, Groenwold RH, Jansen EK, *et al*. Limitations of ranking lists based on cardiac surgery mortality rates. *Circ Cardiovasc Qual Outcomes* 2012;5:403–9.
33 Bronskill SE, Normand SL, Beth LM, *et al*. Longitudinal profiles of health care providers. *Stat Med* 2002;21:1067–88.
34 O'Brien SM, Shahian DM, Delong ER, *et al*. Quality measurement in adult cardiac surgery: part 2--Statistical considerations in composite measure scoring and provider rating. *Ann Thorac Surg* 2007;83(4 Suppl):S13–26.
35 Ash A, Fienberg SE, Louis TAP, *et al*. Statistical issues in assessing hospital performance. Commissioned by the Committee of Presidents of Statistical Societies. 2011. http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/Downloads/Statistical-Issues-in-Assessing-Hospital-Performance.pdf (accessed 18 Sep 2013).
36 Normand S-LT, Glickman ME, Gatsonis C.A. Statistical methods for profiling providers of medical care: issues and applications. *J Am Stat Assoc* 1997;92:803–14.
37 Normand S-LT, Shahian DM. Statistical and clinical aspects of hospital outcomes profiling. *Stat Sci* 2007;22:206–26.
38 Goldstein H, Spiegelhalter DJ. League tables and their limitations: statistical issues in comparisons of institutional performance. *J R Stat Soc (Series A)* 1996;159:385–443.
39 Austin PC, Anderson GM. Optimal statistical decisions for hospital report cards. *Med Decis Making* 2005;25:11–19.
40 Austin PC. A comparison of Bayesian methods for profiling hospital performance. *Med Decis Making* 2002;22:163–72.
41 Christiansen CL, Morris CN. Improving the statistical approach to health care provider profiling. *Ann Intern Med* 1997;127(8 Pt 2):764–8.
42 Austin PC, Naylor CD, Tu JV. A comparison of a Bayesian vs. a frequentist method for profiling hospital performance. *J Eval Clin Pract* 2001;7:35–45.
43 Spiegelhalter D. Funnel plots for institutional comparison. *Qual Saf Health Care* 2002;11:390–1.
44 Spiegelhalter DJ. Funnel plots for comparing institutional performance. *Stat Med* 2005;24:1185–202.
45 Rogers CA, Reeves BC, Caputo M, *et al*. Control chart methods for monitoring cardiac surgical performance and their interpretation. *J Thorac Cardiovasc Surg* 2004;128:811–19.
46 de Leval MR, Francois K, Bull C, *et al*. Analysis of a cluster of surgical failures. Application to a series of neonatal arterial switch operations. *J Thorac Cardiovasc Surg* 1994;107:914–23.
47 Grunkemeier GL, Wu YX, Furnary AP. Cumulative sum techniques for assessing surgical results. *Ann Thorac Surg* 2003;76:663–7.