the number of low- and especially very-low-volume programs performing CABG seems unwarranted. A solution to this problem is somewhat less obvious.

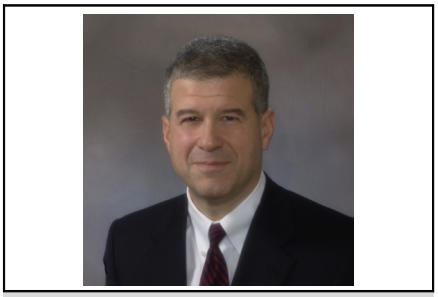
References

1. Mori M, Weininger GA, Shang M, Brooks C II, Mullan CW, Najem M, et al. Association between coronary artery bypass graft center volume and year-to-year outcome variability: New York and California statewide analysis. *J Thorac Cardiovasc Surg*. 2021;161:1035-41.e1.
2. California Office of Statewide Health Planning and Development. Hospital performance ratings for CABG surgery. Available at: https://og-production-open-data-chelseama-892364687672.s3.amazonaws.com/resources/51e78126-5217-49ab-bf4a-9d96767ce073/2017-hospital-results.pdf?Signature=e4sPcVpDjZoUZAjKumMkCScLhmQ%3D&Expires=1598389960&AWSAccessKeyId=AKIAJJIENTAPKHZMIPXQ. Accessed July 28, 2020.
3. Mori M, Shahian DM, Suter LG, Geirsson A, Lin Z, Krumholz HM. Relevance of cardiac surgery outcome reporting 3 years later in a New York and California statewide analysis. *JAMA Surg*. 2020;155:442-4.
4. Lancey RA. How valid is the quantity and quality relationship in CABG surgery? A review of the literature. *J Card Surg*. 2010;25:713-8.
5. Shahian DM, Normand S-LT. Low-volume coronary artery bypass surgery: measuring and optimizing performance. *J Thorac Cardiovasc Surg*. 2008;135:1202-9.
6. Kurlansky PA, Argenziano M, Dunton R, Lancey R, Nast E, Stewart A, et al. Quality, not volume determines outcome of coronary artery bypass surgery in a university-based community hospital network. *J Thorac Cardiovasc Surg*. 2012;143:287-93.
7. Shahian DM, O'Brien SM, Normans S-LT, Peterson ED, Edwards FH. Association of hospital coronary artery bypass volume with processes of care, mortality, morbidity, and the Society of Thoracic Surgeons composite quality score. *J Thorac Cardiovasc Surg*. 2010;139:273-82.

See Article page 1035.

Commentary: Safety in numbers

David M. Shahian, MD



David M. Shahian, MD, Massachusetts General Hospital and Harvard Medical School.

Using publicly reported data from New York and California, Mori and colleagues[1] found substantial year-to-year variation in publicly reported, hospital-level ratios of observed to expected coronary artery bypass grafting (CABG) mortality, which they interpret as measure instability related to small sample sizes. Based on inflection point analyses, they recommend adding mortality metrics derived from a hospital's most recent 111 CABG cases (ie, a standardized denominator sample size) as a complement to traditional annual or biennial reports.

CENTRAL MESSAGE

Quality measures based on small sample sizes have low statistical power and reliability. Mitigation may include multiyear samples, standardized denominator sample sizes, composite metrics, shrinkage estimators, or graphical plots.

THE CURSE OF SMALL NUMBERS

Notwithstanding its methodological issues (eg, admixture of 2 states with markedly different cardiac surgery structures and oversight; attribution of all year-to-year variation in observed to expected mortality as random "noise") and obvious implementation challenges (eg, confusion related to simultaneous time- and sample size–specified measures), this study illustrates a pervasive challenge in health care quality measurement—small sample size.[2-7] Annual hospital discharge volumes for individual conditions and procedures are often relatively low (eg, <100 discharges), which limits accurate performance measurement. For binary outcomes such as mortality, the

From the Division of Cardiac Surgery, Department of Surgery, and Center for Quality and Safety, Massachusetts General Hospital, Boston, Mass.

Disclosures: The author reported no conflicts of interest.

The *Journal* policy requires editors and reviewers to disclose conflicts of interest and to decline handling or reviewing manuscripts for which they may have a conflict of interest. The editors and reviewers of this article have no conflicts of interest.

Received for publication July 17, 2020; revisions received July 17, 2020; accepted for publication July 17, 2020; available ahead of print July 22, 2020.

Address for reprints: David M. Shahian, MD, Division of Cardiac Surgery, Department of Surgery, and Center for Quality and Safety, Massachusetts General Hospital, 55 Fruit St, Boston, MA 02114 (E-mail: dshahian@partners.org).

J Thorac Cardiovasc Surg 2021;161:1043-5
0022-5223/$36.00

https://doi.org/10.1016/j.jtcvs.2020.07.058

confidence and prediction intervals around point estimates widen dramatically when the number of observations is small, which is the basis for so-called "funnel plots."[8,9] This random sampling variation creates greater uncertainty regarding a hospital's true underlying performance when sample sizes for the measured condition or procedure are small.

This statistical issue has important health policy implications. Because of the volume-outcome association, the low-volume programs for whom it is most difficult to reliably measure outcomes are also those more likely to have marginal performance and for whom accurate monitoring is most crucial.[10,11]

STATISTICAL POWER AND MEASURE RELIABILITY

Small sample size is associated with low statistical power to detect outliers, a core function of risk-adjusted performance measures.[12,13] It also compromises one of the most important characteristics of any performance measure—reliability, or the proportion of measured performance variation that is due to true differences in quality.[14] Because reliability is a function of sample size and within- and between-provider variance, when sample sizes are small and within-provider random sampling error increases, reliability is lower.

Many surgical performance measures have reliabilities well below 0.40 to 0.50 at the hospital level,[15,16] which is generally considered a minimally acceptable lower limit. For its composite measures,[17-20] the Society of Thoracic Surgeons (STS) insists on average reliabilities of at least 0.50 and will not assign a performance classification to providers whose volumes are inadequate to ensure this.

MITIGATING SAMPLE SIZE CONCERNS

In addition to a fixed, minimum number of observations as advocated by Mori and colleagues,[1] numerous other strategies have been implemented to address sample size issues. For example, *multiyear time windows* that annually update by 1 year have been used by many health care report cards. The STS uses 3-year sampling periods for all its adult cardiac surgery composite measures except CABG,[21,22] currently based on 1 year of data but expanding to 3 years in 2021. Periods longer than 3 years are generally not recommended, because more remote data may not be representative of current practice. If Mori and colleagues[1] had also examined 3-year time windows, most hospitals would likely have exceeded the authors' recommended 111-case sample size threshold, and it would not have been necessary to propose a separate and possibly confusing measure.

Composite measures encompassing multiple outcomes are extremely useful and have been the basis of STS performance measurement since 2007.[17-22] Composite measures are multidimensional (eg, mortality and morbidity) and thus more comprehensive in scope, and they effectively increase the number of end points, making it possible to more reliably discriminate performance. In the development of the original STS CABG composite measure, mortality alone could classify only 1% of STS participants as better or worse then expected outliers, whereas the composite measure identified 23% as outliers.[22]

Shrinkage estimation[4,23-29] (referred to by some as "reliability adjustment"[30-32]) is a statistical technique that provides more accurate estimates when sample sizes are small, analogous to what happens naturally with regression to the mean as more observations become available. Extreme values are "shrunk" closer to the overall provider population mean, with greater shrinkage for providers with the lowest volumes and less shrinkage for those with larger volumes, whose estimates are inherently more reliable. Shrinkage estimation reduces the likelihood of false-positive outlier identification but may result in more false-negatives.[29]

Graphical methods are particularly useful to monitor performance in low-volume programs. Funnel plots explicitly demonstrate the increasing random sampling variation of point estimates derived from smaller samples and typically include specific alarm and outlier control limits.[8,9] Risk-adjusted CUSUM or VLAD plots[33-38] allow near real-time, case-by-case monitoring of observed versus expected outcomes, facilitating more timely identification of deteriorating trends in performance for low-volume programs.

SURGEON-LEVEL MEASURES

All the preceding concerns are magnified when measuring performance at the surgeon level, for which sample sizes are smaller than for hospitals. The STS individual surgeon composite measure for adult cardiac surgery[39] addresses this challenge by combining the results for 5 common procedures (isolated CABG, isolated aortic valve replacement, isolated mitral procedures, aortic valve replacement + CABG, and mitral procedures + CABG); 3 years of data; and 2 outcomes (risk-adjusted mortality and morbidity). Because of these multiple strategies to increase sample size and endpoints, it has the highest average reliability ever measured for a STS performance measure (0.81). An STS participant-level (hospital or group practice) multiprocedural composite measure with equally high reliability has been developed and will be published in 2021.

CONCLUSIONS

Small sample sizes and relatively low adverse event rates are among the greatest challenges to health care performance measurement, and numerous mitigation strategies have been suggested. Although the fixed, minimum sample size recommendation of Mori and colleagues[1] is a reasonable complement to existing time-based approaches,

longer fixed time periods (eg, 3 years of data) would likely achieve similar reliability with a single, simple, familiar, and easily understood approach.

References

1. Mori M, Weininger GA, Shang M, Brooks C II, Mullan CW, Najem M, et al. Association between coronary artery bypass graft center volume and year-to-year outcome variability: New York and California statewide analysis. *J Thorac Cardiovasc Surg*. 2021;161:1035-41.e1.
2. Normand SLT, Shahian DM. Statistical and clinical aspects of hospital outcomes profiling. *Stat Sci*. 2007;22:206-26.
3. Shahian DM, Normand SL. Low-volume coronary artery bypass surgery: measuring and optimizing performance. *J Thorac Cardiovasc Surg*. 2008;135:1202-9.
4. Shahian DM, Normand SL, Torchiana DF, Lewis SM, Pastore JO, Kuntz RE, et al. Cardiac surgery report cards: comprehensive review and statistical critique. *Ann Thorac Surg*. 2001;72:2155-68.
5. McNeil BJ, Pedersen SH, Gatsonis C. Current issues in profiling quality of care. *Inquiry*. 1992;29:298-307.
6. Normand SL, Wolf RE, Ayanian JZ, McNeil BJ. Assessing the accuracy of hospital clinical performance measures. *Med Decis Making*. 2007;27:9-20.
7. 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.
8. Spiegelhalter DJ. Funnel plots for comparing institutional performance. *Stat Med*. 2005;24:1185-202.
9. Seaton SE, Manktelow BN. The probability of being identified as an outlier with commonly used funnel plot control limits for the standardised mortality ratio. *BMC Med Res Methodol*. 2012;12:98.
10. Shahian DM, Normand SL. What is a performance outlier? *BMJ Qual Saf*. 2015;24:95-9.
11. Shahian D. Improving cardiac surgical quality: lessons from the Japanese experience. *BMJ Qual Saf*. 2020;29:531-5.
12. 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.
13. Walker K, Neuburger J, Groene O, Cromwell DA, van der Meulen J. Public reporting of surgeon outcomes: low numbers of procedures lead to false complacency. *Lancet*. 2013;382:1674-7.
14. Adams J. *The Reliability of Provider Profiling: A Tutorial*. Santa Monica, CA: Rand Health, prepared for the National Committee for Quality Assurance; 2009.
15. Krell RW, Hozain A, Kao LS, Dimick JB. Reliability of risk-adjusted outcomes for profiling hospital surgical quality. *JAMA Surg*. 2014;149:467-74.
16. Krell RW, Staiger DO, Dimick JB. Reliability of surgical outcomes for predicting future hospital performance. *Med Care*. 2014;52:565-71.
17. Badhwar V, Rankin JS, He X, Jacobs JP, Gammie JS, Furnary AP, et al. The Society of Thoracic Surgeons mitral repair/replacement composite score: a report of the Society of Thoracic Surgeons quality measurement task force. *Ann Thorac Surg*. 2016;101:2265-71.
18. Rankin JS, Badhwar V, He X, Jacobs JP, Gammie JS, Furnary AP, et al. The Society of Thoracic Surgeons mitral valve repair/replacement plus coronary artery bypass grafting composite score: a report of the Society of Thoracic Surgeons quality measurement task force. *Ann Thorac Surg*. 2017;103:1475-81.
19. Shahian DM, He X, Jacobs JP, Rankin JS, Welke KF, Edwards FH, et al. The STS AVR+CABG composite score: a report of the STS quality measurement task force. *Ann Thorac Surg*. 2014;97:1604-9.
20. Shahian DM, He X, Jacobs JP, Rankin JS, Welke KF, Filardo G, et al. The Society of Thoracic Surgeons isolated aortic valve replacement (AVR) composite score: a report of the STS quality measurement task force. *Ann Thorac Surg*. 2012;94:2166-71.
21. Shahian DM, Edwards FH, Ferraris VA, Haan CK, Rich JB, Normand SL, et al. Quality measurement in adult cardiac surgery: part 1–Conceptual framework and measure selection. *Ann Thorac Surg*. 2007;83(4 Suppl):S3-12.
22. O'Brien SM, Shahian DM, DeLong ER, Normand SL, Edwards FH, Ferraris VA, 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.
23. Thomas N, Longford NT, Rolph JE. Empirical Bayes methods for estimating hospital-specific mortality rates. *Stat Med*. 1994;13:889-903.
24. Efron B, Morris C. Data analysis using Stein's estimator and its generalizations. *J Am Stat Assoc*. 1975;70:311-9.
25. Efron B, Morris C. Stein's paradox in statistics. *Sci Am*. 1977;235:119-27.
26. Normand SLT, Glickman ME, Gatsonis CA. Statistical methods for profiling providers of medical care: issues and applications. *J Am Stat Assoc*. 1997;92:803-14.
27. Shahian DM, Torchiana DF, Shemin RJ, Rawn JD, Normand SL. Massachusetts cardiac surgery report card: implications of statistical methodology. *Ann Thorac Surg*. 2005;80:2106-13.
28. MacKenzie TA, Grunkemeier GL, Grunwald GK, O'Malley AJ, Bohn C, Wu Y, et al. A primer on using shrinkage to compare in-hospital mortality between centers. *Ann Thorac Surg*. 2015;99:757-61.
29. Mukamel DB, Glance LG, Dick AW, Osler TM. Measuring quality for public reporting of health provider quality: making it meaningful to patients. *Am J Public Health*. 2010;100:264-9.
30. Dimick JB, Staiger DO, Birkmeyer JD. Ranking hospitals on surgical mortality: the importance of reliability adjustment. *Health Serv Res*. 2010;45(6 Pt 1):1614-29.
31. Osborne NH, Ko CY, Upchurch GR Jr, Dimick JB. The impact of adjusting for reliability on hospital quality rankings in vascular surgery. *J Vasc Surg*. 2011;53:1-5.
32. Staiger DO, Dimick JB, Baser O, Fan Z, Birkmeyer JD. Empirically derived composite measures of surgical performance. *Med Care*. 2009;47:226-33.
33. de Leval MR, Francois K, Bull C, Brawn W, Spiegelhalter D. Analysis of a cluster of surgical failures. Application to a series of neonatal arterial switch operations. *J Thorac Cardiovasc Surg*. 1994;107:914-23.
34. Lovegrove J, Valencia O, Treasure T, Sherlaw-Johnson C, Gallivan S. Monitoring the results of cardiac surgery by variable life-adjusted display. *Lancet*. 1997;350:1128-30.
35. Rogers CA, Reeves BC, Caputo M, Ganesh JS, Bonser RS, Angelini GD. Control chart methods for monitoring cardiac surgical performance and their interpretation. *J Thorac Cardiovasc Surg*. 2004;128:811-9.
36. Grunkemeier GL, Wu YX, Furnary AP. Cumulative sum techniques for assessing surgical results. *Ann Thorac Surg*. 2003;76:663-7.
37. Sherlaw-Johnson C. A method for detecting runs of good and bad clinical outcomes on variable life-adjusted display (VLAD) charts. *Health Care Manag Sci*. 2005;8:61-5.
38. Pagel C, Utley M, Crowe S, Witter T, Anderson D, Samson R, et al. Real time monitoring of risk-adjusted paediatric cardiac surgery outcomes using variable life-adjusted display: implementation in three UK centres. *Heart*. 2013;99:1445-50.
39. Shahian DM, He X, Jacobs JP, Kurlansky PA, Badhwar V, Cleveland JC Jr, et al. The Society of Thoracic Surgeons composite measure of individual surgeon performance for adult cardiac surgery: a report of the Society of Thoracic Surgeons quality measurement task force. *Ann Thorac Surg*. 2015;100:1315-25.