Association between coronary artery bypass graft center volume and year-to-year outcome variability: New York and California statewide analysis

Makoto Mori, MD,[a,b] Gabe A. Weininger, BS,[a] Michael Shang, BS,[a] Cornell Brooks II, BA,[a] Clancy W. Mullan, MD,[a] Michael Najem, BS,[a] Magdalena Malczewska, MD,[a] Prashanth Vallabhajosyula, MD,[a] and Arnar Geirsson, MD[a]

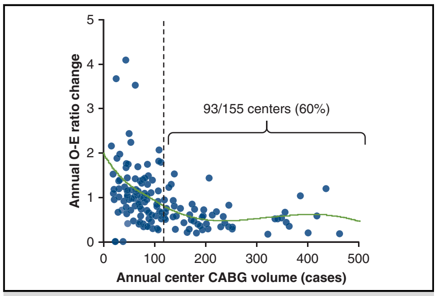
ABSTRACT

Objective: We evaluated whether volume-based, rather than time-based, annual reporting of center outcomes for coronary artery bypass grafting may improve inference of quality, assuming that large center-level year-to-year outcome variability is related to statistical noise.

Methods: We analyzed 2012 to 2016 data on isolated coronary artery bypass grafting using statewide outcome reports from New York and California. Annual changes in center-level observed-to-expected mortality ratio represented stability of year-to-year outcomes. Cubic spline fit related the annual observed-to-expected ratio change and center volume. Volume above the inflection point of the spline curve indicated centers with low year-to-year change in outcome. We compared observed-to-expected ratio changes between centers below and above the volume threshold and observed-to-expected ratio changes between consecutive annual and biennial measurements.

Results: There were 155 centers with median annual volume of 89 (interquartile range, 55-160) for isolated coronary artery bypass grafting. The inflection point of observed-to-expected ratio variability was observed at 111 cases/year. Median year-to-year observed-to-expected ratio change for centers performing less than 111 cases (62 centers) was greater at 0.83 (0.26-1.59) compared with centers performing 111 cases or more (93 centers) at 0.49 (022-0.87) ($P < .001$). By aggregating the outcome over 2 years, centers above the 111-case threshold increased from 93 centers (60%) to 118 centers (76%), but the median observed-to-expected change for all centers was similar between annual aggregates at 0.70 (0.26-1.22) compared with observed-to-expected change between biennial aggregates at 0.54 (0.23-1.02) ($P = .095$).

Conclusions: Center-level, risk-adjusted coronary artery bypass grafting mortality varies significantly from one year to the next. Reporting outcomes by specific case volume may complement annual reports. (J Thorac Cardiovasc Surg 2021;161:1035-41)



Year-to-year variability in risk-adjusted CABG mortality is associated with center volume.

CENTRAL MESSAGE

Risk-adjusted CABG mortality varies significantly from one year to the next in low-volume centers. Reporting based on volume may improve inference.

PERSPECTIVE

Annual outcomes of low-volume centers may be susceptible to irregular clustering of adverse events. We analyzed statewide isolated CABG data and showed that center-level, risk-adjusted mortality varied significantly between consecutive years, and variation was more pronounced at centers with annual volume less than 111 cases. Volume-based aggregation of outcomes may provide a better inference of quality.

See Commentaries on pages 1042, 1043, and 1046.

Profiling outcomes of small hospitals and low-volume surgeons is challenging, because erratic concentration of adverse events could make the inference of true quality challenging for low-volume centers and providers.[1-3] Public reporting of such performance profiles has implications to patients and provider care choices, as well as guiding quality-improvement efforts.

From the [a]Section of Cardiac Surgery, Yale School of Medicine, New Haven, Conn; and [b]Center for Outcomes Research and Evaluation, Yale-New Haven Hospital, New Haven, Conn.

Funding: M.M. is a PhD student in the Investigative Medicine Program at Yale University, which is supported by CTSA Grant Number UL1 TR001863 from the National Center for Advancing Translational Sciences, a component of the National Institutes of Health. Its contents are solely the responsibility of the authors and do not necessarily represent the official view of National Institutes of Health.

Read at the 100th Annual Meeting of The American Association for Thoracic Surgery: A Virtual Learning Experience, May 22-23, 2020.

Received for publication April 12, 2020; revisions received July 1, 2020; accepted for publication July 12, 2020; available ahead of print Sept 22, 2020.

Address for reprints: Arnar Geirsson, MD, Section of Cardiac Surgery, Yale School of Medicine, BB204, 330 Cedar St, PO Box 208039, 06510, New Haven, CT 06511 (E-mail: arnar.geirsson@yale.edu).

0022-5223/$0.00
Published by Elsevier Inc. on behalf of The American Association for Thoracic Surgery
https://doi.org/10.1016/j.jtcvs.2020.07.119

Abbreviations and Acronyms
CABG = coronary artery bypass grafting
O-E = observed-to-expected

Scanning this QR code will take you to the table of contents to access supplementary information. To view the AATS Annual Meeting Webcast, see the URL next to the webcast thumbnail.

Hierarchical treatment of the data by certain volume thresholds and setting the minimum volume for the measured entity to qualify for public reporting have been proposed, although such approaches harbor concerns for gaming.[2] Reliability adjustment is a potential method that could reduce the measure variability based on empirical Bayes methods,[4] although not commonly in use in public reporting. An alternative may be to prolong the measurement period to increase the denominator size for a more reliable outcome measurement. However, it remains unknown above what case volume that rare outcomes, such as operative mortality, may become a reliable metric.

In cardiac surgery, publicly reported outcomes often use 1-year worth of data to infer quality,[5-7] resulting in a wide range of hospital or surgeon-level volumes included in the measured period. One-year duration may be administratively convenient but is arbitrary and may not yield sufficient case volume for many centers to function as a predictor of future performance, even for a case type with relatively large volume such as coronary artery bypass grafting (CABG).[8] Using New York and California statewide outcome reports, we aimed to (1) characterize the relationship between annual center volume and year-to-year outcome variability for isolated CABG; (2) identify case volume above which center-level year-to-year outcome variability stabilizes; and (3) evaluate whether increasing the measurement period from 1 year (annual) to 2 years (biennial) improves the inference of quality. We hypothesize that increasing the measurement period reduces outcome variability between 2 consecutive periods, thereby making such publicly reported outcomes more representative of the actual quality of the hospitals and providers.

MATERIALS AND METHODS

Data Source and Outcome

We used data on isolated CABG from publicly available statewide cardiac surgery outcome reports in New York and California for cases performed from 2012 to 2016.[5,6] Of 171 centers, we included centers performing 1 or more isolated CABG every year between 2012 and 2016 with the only exclusion criterion being centers not performing at least 1 isolated CABG every year between 2012 and 2016 (Figure 1). Annual case volume for each center was calculated as the average of annual volume from 2012 to 2016. The Yale Institutional Review Board exempted the study, and individual consent was waived.

Because the risk models developed by New York and California differed slightly in their model specifications,[5,6] we used observed-to-expected (O-E) operative mortality ratio as the standardized metric of risk-adjusted outcome. The differences between the 2 states' approaches to data collection, model specifications, and response to poor performing outliers are summarized in the Appendix E1. Regardless of the differences in variable definitions, both states' models had similar discriminatory ability with c-statistics ranging from 0.80 to 0.82.[9,10] Additionally, use of different risk models to risk adjust outcomes have yielded a high correlation between different models, especially at the center level, provided that important risk factors are captured in both models.[9,10] This has been demonstrated between clinical and claims datasets, which differ significantly in the nature of variables included. On the basis of such reports, we elected to combine the outcomes of 2 states that were generated by slightly different risk models. O-E ratio is defined as the ratio of observed and expected mortality rates for each center, with O-E ratio of 1 representing expected outcomes given the case-mix at the center. Expected mortality is calculated from state-specific risk models that accounted for patient factors.[5,6]

Measure of Outcome Stability and Analysis

Annual changes in O-E ratio for each center represented the stability of year-to-year outcomes. We interpreted a small O-E ratio to represent stable outcome measures, meaning that the measures were likely a reliable quality metric for that center. Annual changes in O-E ratio were calculated by taking the absolute value of the O-E ratio difference in consecutive years. To define the case-volume threshold above which year-to-year changes in outcomes stabilize, we obtained restricted cubic spline fit of the annual O-E ratio change and center volume distribution, plotting the center-level mean O-E ratio changes and mean center volumes. The knots were defined at 50, 100, and 200 cases arbitrarily. Second-order derivative of the spline curve identified the inflection point of volume, above which the O-E ratio change is interpreted to be stable. We chose to use the inflection point in a spline curve to objectively define a threshold, as adopted by prior studies.[11,12]

To examine whether the biennial aggregate of outcomes serves as a more reliable metric of performance, we calculated O-E ratio changes between 2 consecutive biennial aggregates (2012-2013 vs 2014-2015). We then compared the magnitude of O-E ratio differences between annual and biennial aggregates. Values were reported as median with interquartile range. Continuous variables were compared using the Wilcoxon rank-sum test. Considering the difference between the distribution of case volumes in New York and California, we fitted linear regression models for case volume and O-E ratio variability to evaluate whether a similar relationship between the 2 variables exists in both states. Analyses were conducted using SAS 9.4 (SAS Institute, Inc, Cary, NC) and Python 3.6 (Python Software Foundation, Wilmington, Del).

RESULTS

There were 155 centers (119 in California and 36 in New York) with median annual isolated CABG volume of 89 (55-160) cases per center (Table 1). Restricted cubic spline curve demonstrated an inverse relationship between volume and outcome variability, with the inflection point of O-E ratio variability occurring at 111 cases (Figure 2). Median year-to-year O-E ratio change for centers performing less than 111 cases (62 centers) was 0.83 (0.26-1.59), and those with 111 cases or more (93 centers) was 0.49 (0.22-0.87) ($P < .001$).

By aggregating the outcome over 2 years, centers above the 111-case threshold increased from 93 centers (60%) to 118 centers (76%), but the median O-E change for all centers was similar between biennial aggregates at 0.54

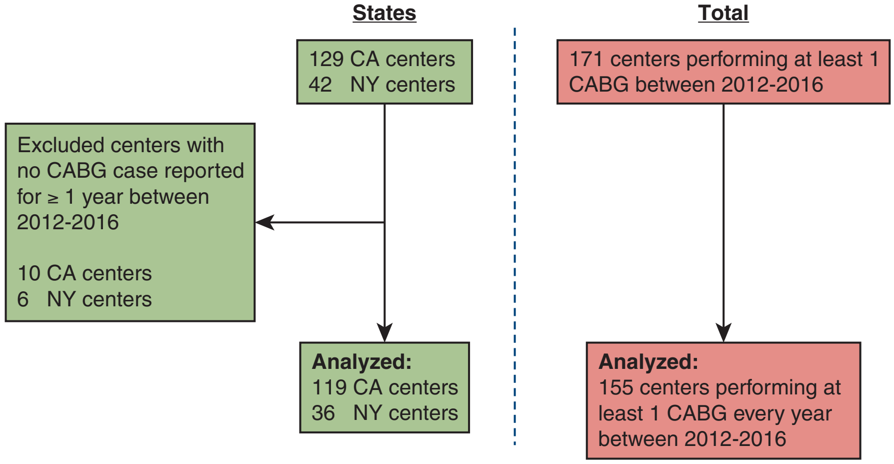


FIGURE 1. Flow diagram for included centers. The figure summarizes the exclusion of centers for each state. *CABG*, Coronary artery bypass grafting.

(0.23-1.02) compared with O-E change between annual aggregates at 0.70 (0.26-1.22) ($P = .095$).

Comparing the relationship of annual O-E variability and case volume in the 2 states, the linear regression model demonstrated that California centers had a 0.13 (0.03-0.22; $P = .012$) decline in O-E ratio variability and New York centers had 0.18 (0.06-0.30, $P = .003$) decline in O-E ratio variability per 50-case increase.

DISCUSSION

Using statewide datasets for 2 large states, we demonstrated that year-to-year risk-adjusted CABG mortality varied widely and was associated with center volume. The year-to-year variation in mortality measure stabilizes at center volume above 111 cases per year. Extending the measurement duration from 1 year to 2 years increased the number of centers above the volume threshold, but the O-E ratio difference was similar between 2 consecutive biennial measurement periods compared with 2 annual measurement periods (Figure 3). These findings suggest that standardizing the reporting frame by cumulative volume, rather than time elapsed, may reduce outcome variability related to small sample size and improve inference of quality. Additionally, the results imply larger year-to-year variability in surgeon-level measurements, which have a lower volume as a denominator.

The findings are important, because although low center volume has been known to limit the accuracy of estimates,[3,13] to our knowledge this is the first study to characterize the relationship between year-to-year outcome variations and case volume specific to CABG. The quantification of the case threshold below which the outcome varies dramatically from year to year also provides a tangible case volume above which the risk-adjusted mortality measure may become a more reliable quality metric. These results argue for complementarily aggregating the data for hospital and surgeon performance based on cumulative volume in addition to the current approach, which is based on timeframe (ie, per year) and is applied indiscriminately to both high- and low-volume entities. By estimating all entities' performances based on the last 111 CABG cases performed, the profiling may represent the quality more accurately. This approach, which standardizes the size of the denominator, may also improve profiling of hospitals based on nonmortality outcomes that may have poor reliability in discriminating hospital performances.[14] We suggest this approach to supplement but not replace the conventional time-based aggregation of the data, which has appealing features such as a more intuitive interpretation. The conventional approach likely provides reliable outcome estimates for high-volume centers, although it is difficult to objectively define above which volume should be considered to be high-volume. This challenge also supports concurrent use of these measures.

Reliability adjustment is a statistical technique based on empirical Bayes methods that can "shrink" the variations in risk-adjusted outcomes as a function of sample size.[4] Although such approaches to shrinking the variability may be appealing for the purpose of tempering erratic extreme outcomes, the technique suffers from the inherent problem

TABLE 1. Annual center-level outcomes from 2012 to 2016

Year	2012	2013	2014	2015	2016
Total isolated CABG volume	19,212	19,415	19,350	19,987	20,716
Median center volume (IQR)	87 (56-149.5)	86 (56.5-161)	89 (56-160.5)	92 (57-168)	97 (54.5-170.5)
Observed mortality rate	1.88%	2.11%	1.70%	2.15%	2.09%
Median O-E ratio difference from prior year (IQR)	-	0.56 (0.24-1.25)	0.7 (0.26-1.20)	0.81 (0.27-1.44)	0.59 (0.21-1.09)

CABG, Coronary artery bypass grafting; *IQR*, interquartile range; *O-E*, observed-to-expected.

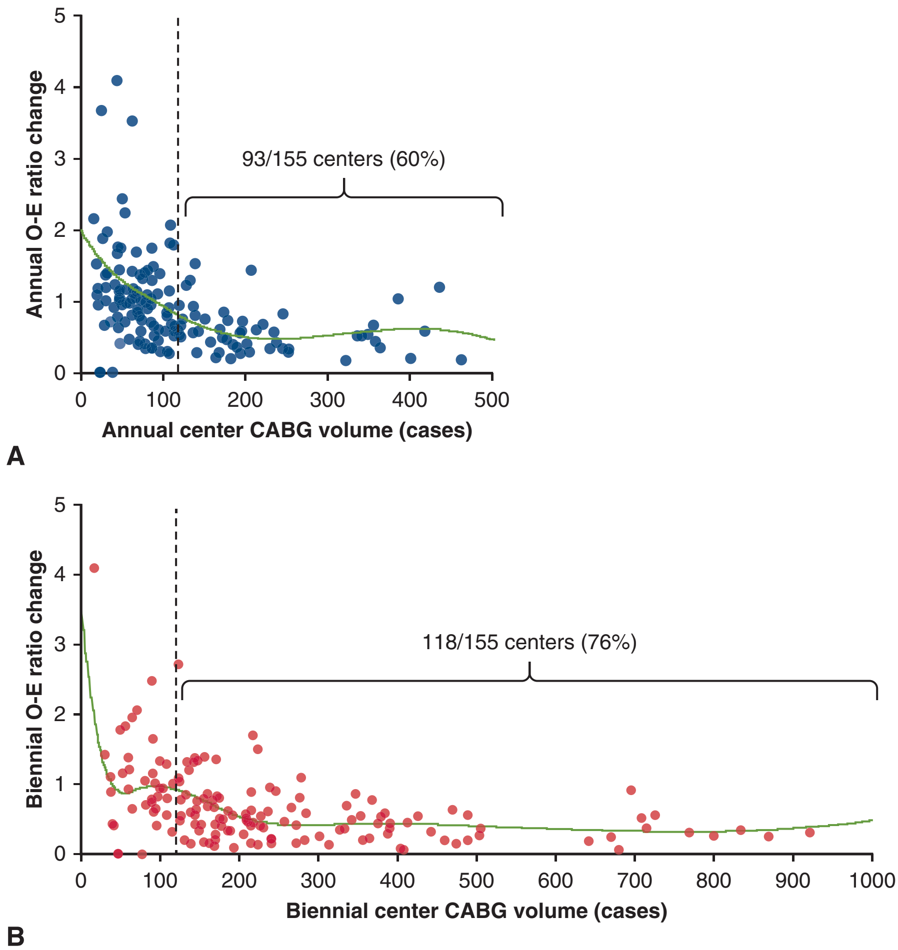


FIGURE 2. Center-level annual and biennial O-E ratio changes by CABG volume. Scatter plot of center-level annual (A) and biennial (B) O-E ratio changes by center CABG volume. *Each circle* represents a center. Volume and O-E ratios change were center-level means during the study period. *Solid lines* are the cubic spline fits. *Dotted vertical line* indicates 111-case threshold where inflection of the annual spline curve occurred. The figure shows that increasing the measurement period from annual to biennial increased the number of cases above 111-case threshold from 93 (60%) to 118 (76%) centers. *Blue triangles* indicate inflection points derived from the biennial data. *O-E*, Observed-to-expected; *CABG*, coronary artery bypass grafting.

that the inference of quality at low volume is poor, and the apparent null difference in performance measures due to such shrinkage of variability may not be informative because it makes many centers look alike in terms of performance.[15]

Because the stability of outcome relates not only to the total case volume but also to the incidence of event, it is likely that the 111-case threshold that we identified is unique to isolated CABG when the mean operative mortality is approximately 2%. This threshold is likely different for case types with a higher mortality rate, such as mitral valve replacement with a mean operative mortality of 6%,[16] or when evaluating an event of a higher incidence, such as composite morbidity or sternal wound infection rate.

A potential use case may be that such supplementary reporting based on a standardized cumulative case volume may be updated on a rolling basis, continuously reporting the average of the last 111 cases, for example. Our data indicated that it would take longer than 2 years for 25% of the centers to accrue 111 cases, but reporting the rolling average would circumvent this potential delay of more than 2 years. As the current 3-year delay in publication already makes such reports less relevant and actionable,[17] timely processing and dissemination of the data are important. Additionally, the high number of relatively low-volume centers and difficulty in inferring the quality for such centers may be additional angles to consider in the ongoing debate surrounding volume regionalization. Any implementation of such a reporting approach is likely to be highly context-specific, considering the available resources and the acceptance by the public.

In terms of reporting frame, there are organizations such as the Scientific Registry of Transplant Recipients that have successfully reported time-based (12-month) rolling average of outcomes,[18] and examination of the efficacy of

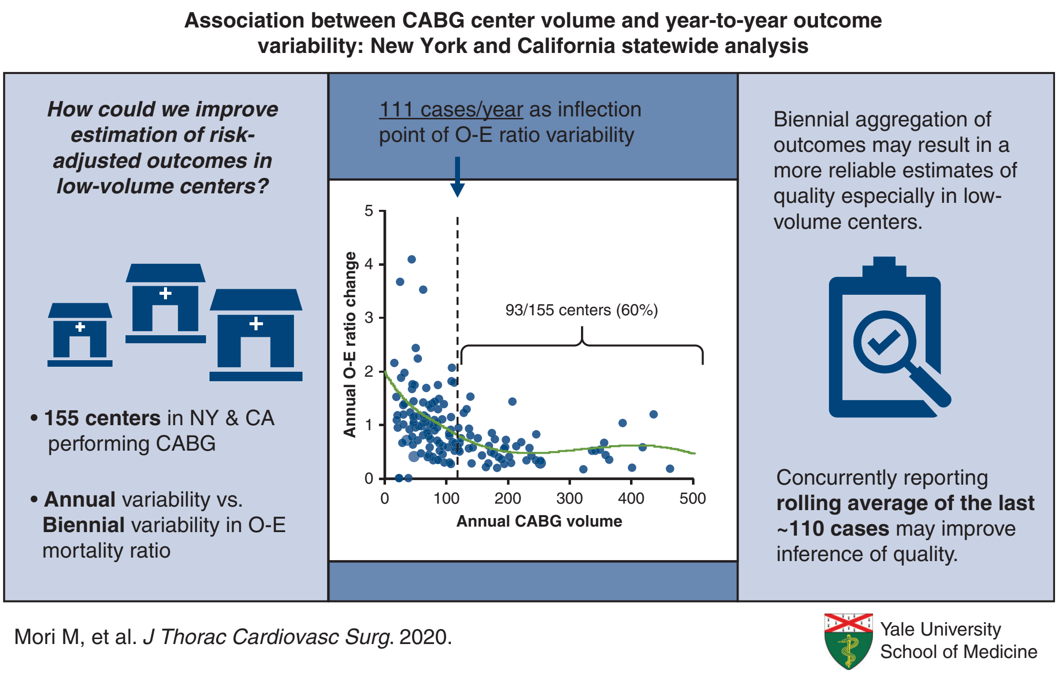


FIGURE 3. Summary of the study using statewide data from 2 large states to evaluate the relationship between outcome variability and center case volume. *CABG*, Coronary artery bypass grafting; *O-E*, observed-to-expected.

reporting the rolling frame based on case volume is needed. The variable frequencies of report updates by different organizations suggest that the delay may be related to logistic issues rather than something inherent to public reporting itself. Validation of the data is critical, because there are reports suggesting that hospitals tend to underreport mortalities.[19] As for possible ways to shorten this interval, data mapping and validation based on artificial intelligence and machine learning algorithms have reported promising results,[20] and larger-scale implementation of such approaches may enable timely reporting of such results. Although controversial and costly,[21] quality measures and public reporting mechanisms will continue to exist. It is imperative to improve confidence in such measures because the implications are far-reaching, influencing referral patterns, surgeons' risk aversion behavior,[22] patient preference,[23] and hospital payment.[24]

Study Strengths and Limitations

This study used statewide data from 2 large states in the United States representing approximately 15% of the US population. The inclusion of a wide range of patient population afforded by the statewide data, in contrast to age-restricted payer-specific claims data such as Medicare data, increases generalizability of the results. The risk-adjusted mortality rate reported in the publicly available documents was estimated by state-specific models that were rigorously developed and validated.[5,6]

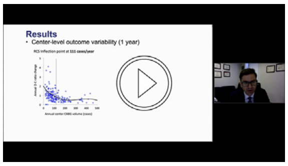
Limitations of this study include the lack of patient-level data, which may have provided additional insights into case-mix variations across centers. Additionally, it is possible that centers truly changed in their qualities from one year to the next, and that the year-to-year variation in the reported outcomes mostly represents the change in quality with relatively small contribution of the statistical noise. However, it is unknown that rather sizable changes, in terms of true performance surrogated by the O-E ratio in this study, occur on a year-to-year basis in a large number of centers. Furthermore, the nature of the dataset precluded the assessment of nonclinical changes occurring at centers, such as large turnover of surgeons during the study period. Analyses at the surgeon level were not conducted, and the potential interactions between surgeon-level outcome variability and surgeons operating at multiple hospitals require further investigation. The inflection point in the center volume identified through the cubic spline function is dependent on the dataset. Although our analysis included a wide distribution of center volume from 2 large states to improve generalizability, the volume threshold reported may be more appropriately interpreted as a range around which the outcome stabilizes. Two states differ in the states' reaction to poor reported outcomes, in that New York requires regular review of outcomes and response from the program describing improvement efforts where needed, whereas such response is not required in California. Although evidence is limited to determine whether such difference in the reaction to the report affects subsequent outcomes differently between the 2 states, this difference may have confounded the relationship between case volume and variability, as California centers tended to be lower volume.

CONCLUSIONS

Center-level, risk-adjusted CABG mortality varies significantly from one year to the next, especially in centers performing less than approximately 110 cases per year. Aggregating the outcome data by volume, instead of time period, may improve the reliability of performance measures.

Webcast

You can watch a Webcast of this AATS meeting presentation by going to: https://aats.blob.core.windows.net/media/20AM/Presentations/Association%20Between%20CABG%20Center%20Volu.mp4.



Conflict of Interest Statement

The authors 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.

References

1. Weintraub WS, Garratt KN. Challenges in risk adjustment for hospital and provider outcomes assessment. *Circulation*. 2017;135:317-9.
2. Christiansen CL, Morris CN. Improving the statistical approach to health care provider profiling. *Ann Intern Med*. 1997;127:764-8.
3. Shahian D, Normand S-L. Statistical and clinical aspects of hospital outcomes profiling. *Stat Sci*. 2007;22:206-26.
4. Dimick JB, Ghaferi AA, Osborne NH, Ko CY, Hall BL. Reliability adjustment for reporting hospital outcomes with surgery. *Ann Surg*. 2012;255:703-7.
5. New York State Department of Health. Cardiovascular Disease Data and Statistics; 2019. Available at: https://www.health.ny.gov/statistics/diseases/cardiovascular/. Accessed July 1, 2019.
6. California Hospital Performance Ratings for Coronary Artery Bypass Graft (CABG) Surgery; 2019. Available at: https://oshpd.ca.gov/data-and-reports/healthcare-quality/cabg-reports/. Accessed July 1, 2019.
7. State of New Jersey Department of Health. Hospital Performance Report; 2019. Available at: https://web.doh.state.nj.us/apps2/hpr/usingcabg.aspx. Accessed July 1, 2019.
8. Glance LG, Dick AW, Mukamel DB, Li Y, Osler TM. How well do hospital mortality rates reported in the New York State CABG report card predict subsequent hospital performance? *Med Care*. 2010;48:466-71.
9. Krumholz HM, Wang Y, Mattera JA, Han LF, Ingber MJ, Roman S, 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.
10. Krumholz HM, Wang Y, Mattera JA, Han LF, Ingber MJ, Roman S, 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.
11. Chikwe J, Toyoda N, Anyanwu AC, Itagaki S, Egorova NN, Boateng P, et al. Relation of mitral valve surgery volume to repair rate, durability, and survival. *J Am Coll Cardiol*. 2017;69:2397-406.
12. Mori M, Bin Mahmood SU, Zhuo H, Yousef S, Green J, Mangi AA, et al. Persistence of risk of death after hospital discharge to locations other than home after cardiac surgery. *J Thorac Cardiovasc Surg*. March 6, 2019 [Epub ahead of print].
13. Thomas N, Longford NT, Rolph JE. Empirical Bayes methods for estimating hospital-specific mortality rates. *Stat Med*. 1994;13:889-903.
14. Krell RW, Hozain A, Kao LS, Dimick JB. Reliability of risk-adjusted outcomes for profiling hospital surgical quality. *JAMA Surg*. 2014;149:467-74.
15. Silber JH, Rosenbaum PR, Brachet TJ, Ross RN, Bressler LJ, Even-Shoshan O, et al. The Hospital compare mortality model and the volume-outcome relationship. *Health Serv Res*. 2010;45(5 Pt 1):1148-67.
16. Badhwar V, Rankin JS, Thourani VH, D'Agostino RS, Habib RH, Shahian DM, et al. The Society of Thoracic Surgeons adult cardiac surgery database: 2018 update on research: outcomes analysis, quality improvement, and patient safety. *Ann Thorac Surg*. 2018;106:8-13.
17. 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.
18. Scientific Registry of Transplant Recipients. Available at: https://www.srtr.org/. Accessed April 5, 2020.
19. Hannan EL, Samadashvili Z, Cozzens K, Chikwe J, Adams DH, Sundt TM, et al. Out-of-hospital 30-day deaths following cardiac surgery are often under-reported. *Ann Thorac Surg*. 2020;110:183-8.
20. McPadden J, Durant TJ, Bunch DR, Coppi A, Price N, Rodgerson K, et al. Health care and precision medicine research: analysis of a scalable data science platform. *J Med Internet Res*. 2019;21:e13043.
21. Casalino LP, Gans D, Weber R, Cea M, Tuchovsky A, Bishop TF, et al. US physician practices spend more than $15.4 billion annually to report quality measures. *Health Aff (Millwood)*. 2016;35:401-6.
22. Omoigui NA, Miller DP, Brown KJ, Annan K, Cosgrove D, Lytle B, et al. Outmigration for coronary bypass surgery in an era of public dissemination of clinical outcomes. *Circulation*. 1996;93:27-33.
23. Romano PS, Marcin JP, Dai JJ, Yang XD, Kravitz RL, Rocke DM, et al. Impact of public reporting of coronary artery bypass graft surgery performance data on market share, mortality, and patient selection. *Med Care*. 2011;49:1118-25.
24. CMS.gov. Hospital-level 30-day All-Cause Unplanned Readmission Following Coronary Artery Bypass Graft Surgery; 2014. Available at: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/Downloads/Coronary-Artery-Bypass-Graft-CABG-Readmission.zip. Accessed February 10, 2020.

Key Words: CABG, outcome reporting, center volume, volume-outcome relationship

Discussion

Presenter: Dr Arnar Geirsson

Dr Clifford W. Barlow *(Southampton, United Kingdom)*. My first question relates to combining data from California and New York. There are important differences between the 2 states including the number of centers, but also their methodology, analysis, and reporting of data. Could you comment on that? Second, do the results still apply when the states are analyzed separately?

Dr Arnar Geirsson *(New Haven, Conn)*. The 2 states have several key differences in factors pertinent to the studies. First, the number of centers is lower in New York—36 versus 119 in California. We didn't really focus on the difference in the article. The mean annual case volume is also different; there were 60 cases per year in California versus 220 in New York.

Perhaps this relates to the differences in geographical concentration of the 2 states. Both states use clinical registry, but the variables included differ slightly between the 2 states, as well as the inclusion criteria in the model. For example, New York excludes patients with preoperative shock, although it's a small portion of the patients. California also excludes what they define as salvage status undergoing active cardiopulmonary resuscitation.

In considering these differences, we elected to use O/E ratio rather than metrics of performance such as risk-standardized mortality rates, which is common, meaning in both states the O/E ratio of 1 indicates as expected performance giving patients risk, whereas risk-standardized mortality rate of, for example, 2%, may be above the state average in one or below in the other. In other words, the risk-standardized mortality rate would be at risk to the overall mortality rates within the state, and using this value would not allow us to compare the outcomes of centers between 2 different states. In short, we certainly acknowledged the differences and accounted for them, we believe, by using the ratio measures.

Dr Barlow. That's a very thorough answer. This is such an important and topical subject because public reporting of performance profiles has such major implications to hospitals and surgeons. One of the positive implications, for example, is that suboptimally performing centers are sometimes inspired and learn from the presentation of public data. On the other hand, of course, low-volume centers are protected when time-based reporting takes place by the presence of wide confidence intervals if the data is potentially wrong.

So, if you perceive using your method with volume reporting, will the protection provided by wide confidence intervals potentially disappear? Could hospitals and surgeons then be wrongly labeled as being poorly performing for many of the other reasons that outcome data are frequently wrong?

Dr Geirsson. I'd say that a relevant analogy may be that you would consider it unethical to conduct an underpowered experiment by sacrificing 50 animals 3 times and not achieving meaningful interrogation of the result. Whereas if you do a single 150-animal experiment that would be adequately powered to provide adequate power of analytic statistics and get meaningful results.

In that sense, we propose that the protection of the low-volume centers by truncating the data annually and not powering them sufficiently to achieve a statistically significant difference is perhaps depriving the centers from knowing whatever is actually going on that could be improved once the problem is identified analytically as a problem. At the same time, your point is well taken: Let's say it takes a long time—let's say it takes 5 years for a low-volume center to achieve sufficient denominator case volume. Let's say 100 cases. Then it's possible that poor outcomes concentrating the first 2 years may lead hospitals to make changes to improve outcomes yet still be penalized for the initial poor outcomes.

So we agree that a volume-based approach is limited in that sense. We would propose that reporting both time-based and volume-based analysis may paint the better pictures. You would think that resource-wise, it would not be prohibited in addition to what's already being done, using various outcomes measures reporting this.

Dr Barlow. That makes sense but leads to my last question, which relates to the practical implications of the reporting of outcomes by volume. Reporting by time periods is understandable to the public—a 1-year period for example. How would you practically propose reporting every 110 cases? Some centers could be reporting several times a year but others only every 3 years potentially. By the same token, what would the delay be in the data release upgrade? One of the reasons that the current delay of up to 1 year, and sometimes 2, takes place before the upgrade is because we're checking the data. How would you confirm that the data are valid?

Dr Geirsson. That's another important question—to address the limitation related to the issues we propose. The practical output of the data is important. Our stance is to advocate for reporting of both the current time-based aggregation in addition to volume-based aggregation of data.

As for the mechanics, we may propose, for example, using a rolling average of the last X number of cases as an option. The question is how often one should update these data. Ideally, it is obviously a real-time update. It's probably not practical at this stage because it only takes actually 3 years for these 2 state reports to report the outcomes. The 2016 report for both of those states actually published in 2019.

There are some other studies. The Scientific Registry of Transplant Recipients updates results twice a year, and the Society of Thoracic Surgeons provides quarterly reports. We think that reporting the last X number of cases at the frequencies around that range should be practical and reasonable. Once we start using artificial intelligence–based data mapping or natural language processing, there may be a time that real-time reporting becomes a possibility. I don't think we're there quite yet, though. Ensuring the fidelity of the data is certainly important, but the variations in delay (several months vs 3 years across various organizations) make us think it's really more of a logistic and resource issue than something inherent to the process itself. So we should continue to acknowledge it as an important problem that requires attention and improvement.

Dr Barlow. Thank you, Dr Geirsson, for your presentation and thorough answers.

APPENDIX E1. DIFFERENCES AND SIMILARITIES BETWEEN CALIFORNIA AND NEW YORK STATES' RISK MODEL SPECIFICATIONS

Covariate Specifications and Model Performances

- California treats ejection fraction as a continuous variable, whereas New York categorizes ejection fraction to more than 40%, less than 30%, and 30% to 39%.
- California model uses body mass index, although it was not a significant risk factor, whereas New York model uses body surface area and found it to be a significant risk factor.
- California model incorporated liver disease and found it to be a significant risk factor, whereas the New York model does not report liver disease.
- New York model uses 5 categories of renal failure, whereas California only reports binary use of dialysis.
- New York distinguishes recent myocardial infarction between less than 6 hours and 6 to 23 hours, whereas California groups them together in less than 24 hours
    - California considers myocardial infarction more than 7 days before surgery and New York does not.
- California distinguishes between percutaneous coronary intervention less than 6 hours or more than 6 hours before surgery, whereas New York only specifies percutaneous coronary intervention in "same episode of care."
- California C-Statistic = 0.816 versus New York C-Statistic = 0.804

Data Collection, Validation, and Exclusion

- Cardiogenic shock is excluded from New York (<1% of the entire cohort).
- Salvage status is excluded in California.
- Mortality definition: New York also counts death anytime within the same hospitalization. California specifies a 90-day cutoff for death within the same hospitalization. Both states specify a 30-day cutoff for death after discharge.
- Both states review unusual reporting frequencies and validate cardiac surgery data against Department of Health databases.
- Death after discharge data:
    - New York Data on deaths after discharge obtained from Department of Health, New York City Department of Health and Mental Hygiene Bureau of Vital Statistics, and National Death Index
    - California data on deaths after discharge obtained from California Department of Public Health
- Both states conduct audits
    - In 2017, California audited 27 hospitals and a total of 2316 patient records.
        - California audits hospitals chosen randomly + hospitals, which are identified as "Better" and "Worse" performers based on preliminary risk models.
    - New York conducts review of medical records for "a selected sample of cases."
        - Does not specify how many hospitals/patient records were audited.
- New York excluded 166 patients who resided outside the United States.

Reaction to High Mortality Outlier Hospitals

- High mortality outliers New York State receive interventions from the Cardiac Advisory Committee in the form of an initial notification letter requesting a response based on raw data, whereas no such response mechanism was described in California report.