By Laurent G. Glance, Caroline P. Thirukumaran, Yue Li, Shan Gao, and Andrew W. Dick

DOI: 10.1377/hlthaff.2019.00778
HEALTH AFFAIRS 39,
NO. 5 (2020): 862–870

Improving The Accuracy Of Hospital Quality Ratings By Focusing On The Association Between Volume And Outcome

Laurent G. Glance (laurent_glance@urmc.rochester.edu) is vice chair for research and a professor in the Department of Anesthesiology and Department of Public Health Sciences, University of Rochester School of Medicine and Dentistry, in New York.

Caroline P. Thirukumaran is an assistant professor in the Department of Orthopaedics and Department of Public Health Sciences, University of Rochester School of Medicine and Dentistry.

Yue Li is a professor in the Department of Public Health Sciences, University of Rochester School of Medicine and Dentistry.

Shan Gao is an associate in the Department of Biostatistics and Computational Biology, University of Rochester School of Medicine and Dentistry.

Andrew W. Dick is a senior economist at RAND Health, RAND Corporation, in Boston, Massachusetts.

ABSTRACT The Centers for Medicare and Medicaid Services (CMS) uses hierarchical modeling to stabilize its hospital quality star ratings by shrinking the performance of low-volume hospitals toward the performance of average hospitals. Responding to criticism that the methodology may distort the performance of low-volume hospitals, a CMS expert panel recommended that the agency consider using "shrinkage targets" to more accurately classify hospital quality performance. To test the "shrinkage targets" approach, we created two parallel sets of performance measures. We found that there was moderate-to-substantial agreement between the standard CMS approach and the approach based on shrinkage targets in hospital star ratings for all but the lowest-volume hospitals. These findings suggest that the standard CMS risk-adjustment methodology does not distort the star ratings of hospitals as long as case volumes exceed the current cutoff (twenty-five cases) used by CMS for public reporting.

Performance measurement is central to the efforts of the Centers for Medicare and Medicaid Services (CMS) to make health care safer and more affordable and to allow patients to make informed choices.[1] Performance measures must be valid to avoid misleading regulators and the public.[2] Two of the most important challenges to creating valid quality measures are that serious adverse outcomes are rare and that many hospitals have low case volumes. Reporting the performance of low-volume hospitals using the well-known observed-to-expected ratio based on standard logistic regression can lead to wild fluctuations in hospital ratings from year to year.[3] Instead, CMS uses hierarchical regression modeling to minimize the large year-to-year fluctuations for low-volume hospitals that are due to chance alone.[4] This technique estimates a hospital's performance as the weighted average of its own outcomes and the performance of average hospitals.[5] The smaller a hospital's volume, the more this weighted estimate is tilted toward the performance of average hospitals.

But it has long been widely understood that low-volume hospitals have higher mortality rates for surgical procedures and common medical conditions and that the performance of low-volume hospitals is frequently below average.[6–9] Ten years ago Jeffrey Silber and coauthors showed that Medicare's Hospital Compare model for acute myocardial infarctions strikingly underestimates the risk-adjusted mortality rates of low-volume hospitals.[3] The authors showed that adding hospital volume to the acute myocardial infarction model resulted in much more accurate mortality estimates for low-volume hospitals. Citing the work of Silber and coauthors, a recent white paper commissioned by CMS and the Committee of Presidents of Statistical Societies recommended that measure developers consider incorporating shrinkage targets in hierarchical modeling to address this problem.[10] When

shrinkage targets based on hospital case volume are used, a hospital's performance is calculated as the weighted average of its own outcomes and the performance of other hospitals with similar case volumes. This approach shrinks the performance of low-volume hospitals to the overall performance of other low-volume hospitals, instead of shrinking low-volume hospitals' performance to that of average hospitals.

The goal of this exploratory analysis was to examine the changes in hospital star rankings for aortic valve replacement when shrinkage targets based on hospital case volume are used instead of the standard CMS approach. We used Medicare data to create two sets of hospital star ratings for thirty-day mortality, one based on conventional hierarchical modeling and the other using hospital case volume as a shrinkage target. CMS uses a star-rating system (one to five stars) to publicly report the overall performance of hospitals based on a composite outcome measure that includes mortality and readmissions.[11] Although aortic valve replacement mortality is not publicly reported on Hospital Compare, we selected this surgery because it is one of the most common cardiac surgeries, is often performed in low-volume hospitals, and has a strong volume-outcome association.[8] For simplicity, we chose to base our analysis on a single outcome, mortality, instead of a composite outcome. We hypothesized that shrinkage targets would shift the measured performance of many low-volume hospitals from average to below average, given the strong volume-outcome association for this surgery. Our study goes beyond Silber and coauthors' seminal report[3] by illustrating the impact of using shrinkage targets on hospital quality rating. Our findings may prove useful to CMS, the National Quality Forum, measure developers, and other stakeholders that seek to address the potential low-case-volume bias inherent in hierarchical modeling that treats low-volume hospitals as if they were average when, for many conditions for which there is a strong volume-outcome association, their performance may be very much below average.

Study Data And Methods

DATA SOURCE This study was conducted using data for 2013–15 from the 100 percent Medicare Provider Analysis and Review (MedPAR) files and the Master Beneficiary Summary File. These databases include beneficiary demographic information; *International Classification of Diseases, Ninth Revision, Clinical Modification* (ICD-9-CM), diagnosis and procedure codes; and mortality for all fee-for-service Medicare patients.

The Institutional Review Board of the University of Rochester School of Medicine and Dentistry approved the study protocol.

STUDY SAMPLE We identified 134,144 patients who underwent aortic valve replacement in the period January 2013–September 2015. Patients who were younger than age sixty-five ($n = 7,422$) or who also underwent mitral valve replacement ($n = 3,779$) or mitral valve repair ($n = 7,603$) were excluded (for additional exclusion criteria, see online appendix exhibit A1).[12] The analytic data set consisted of 115,084 observations in 1,166 hospitals.

MODEL DEVELOPMENT We first estimated a baseline nonhierarchical multivariable logistic regression model (model 1). We adjusted for patient age, surgical urgency, concomitant coronary artery bypass grafting surgery, history of previous cardiac surgery, and coexisting diseases using the Elixhauser comorbidity algorithm.[13] We then examined the volume-outcome association (model 2), because the use of shrinkage targets based on hospital case volume would not be indicated in the absence of a clinically significant volume-outcome association (model details are in appendix exhibit A5).[12] We used robust variance estimators in models 1 and 2 to account for clustering of observations within hospitals.[14]

Based on the baseline model, we estimated a hierarchical logistic regression model (without shrinkage targets) for thirty-day mortality (model 3). We specified hospitals as random effects. We also estimated another model identical to model 3, except that it included hospital case volume as a shrinkage target (model 4). The optimal specification for the volume term was determined using fractional polynomials.[15,16]

HOSPITAL PERFORMANCE We used model 3 to calculate the hospital predicted-to-expected ratio using the standard CMS approach.[17] This ratio is a measure of hospital performance and is analogous to the hospital observed-to-expected mortality ratio based on nonhierarchical modeling. The hospital predicted mortality rate was calculated using patient-level risk factors and included the hospital contribution to outcomes. The hospital expected mortality rate was calculated using only patient-level risk factors and did not include the hospital effect (see the Methods Supplement in the online appendix).[12] We used bootstrapping to estimate 95% confidence intervals around the hospital predicted-to-expected ratios.[18] The risk-adjusted mortality rate was calculated by multiplying the hospital predicted-to-expected ratio by the overall thirty-day mortality rate for all patients.

We used model 4 to calculate hospital predicted-to-expected ratios based on shrinkage targets, as described in the CMs-commissioned white

paper.[10] The hospital predicted mortality was a function of hospital case volume, in addition to patient risk factors and the hospital contribution to outcome. The hospital expected mortality rate was calculated using only patient-level risk factors and a calibration factor (see the Methods Supplement in the online appendix).[12]

We applied k-means clustering to assign each hospital one to five stars.[11] This iterative procedure partitioned hospitals into five categories by minimizing the distance between each hospital's risk-adjusted mortality rate and the mean risk-adjusted mortality rate for each star category.[11,19] We applied this CMS algorithm separately to the distribution of risk-adjusted mortality rates based on standard shrinkage and shrinkage targets to create two sets of star ratings. We also classified hospitals as low-performance outliers if the lower limit of their 95% confidence interval was higher than the national mortality rate and as high-performance outliers if the upper limit was lower than the national mortality rate.

We also calculated the overall observed-to-expected ratio quartiles based on all of the patients in each volume quartile. We chose this ratio because it provides an estimate of the performance of all of the hospitals together as a group in each quartile. The observed mortality rate is the actual mortality rate for the patients treated by all of the hospitals in a volume quartile, whereas the expected mortality rate is the average of the predicted mortality rates for the same patients based on the baseline nonhierarchical model (model 1). While hospital observed-to-expected ratios for low-volume hospitals are frequently unstable because of small sample sizes, the overall observed-to-expected ratio based on all patients in each of the volume quartiles, including the lowest, was expected to provide reliable estimates of the overall performance of the hospitals in each volume quartile because the observed-to-expected ratios for each quartile were based on large sample sizes.

COMPARISON OF MODEL FIT We evaluated model calibration for the standard shrinkage model using calibration plots. We first ranked the observations according to predicted risk of thirty-day mortality and then divided the analytic sample into ten equal-size deciles of risk. We then plotted the mean of the observed mortality rate alongside the mean of the predicted mortality rate for each decile as a function of the decile of risk. To further evaluate model calibration, we created separate calibration plots for hospitals with very low (fewer than 25 cases), low (25–49 cases), medium (50–124 cases), and high (at least 125 cases) volumes, based approximately on hospital volume quartiles. In addition to standard calibration plots based on deciles of risk, we compared the observed and predicted mortality rates for each of the four volume quartiles using the two-sample *t*-test. We also evaluated model discrimination using the C statistic.[20] We evaluated the performance of the shrinkage targets model in a similar fashion.

ANALYSIS Our analytic plan is outlined in appendix exhibit A3.[12] We first compared the distributions of hospital predicted-to-expected ratios based on the standard shrinkage and shrinkage targets to the distributions of the overall observed-to-expected ratios for each hospital volume quartile using the sign test. We then compared hospital rankings based on standard shrinkage (model 3) and shrinkage targets (model 4) by assessing the agreement of star ratings, risk-adjusted mortality rates, and performance outlier status. We assessed the agreement for the star ratings and outlier status using kappa analysis,[21] and we repeated this analysis for star ratings after we stratified hospitals into volume quartiles. We assessed agreement for risk-adjusted mortality rates using the intraclass correlation coefficient, and we repeated this analysis after stratifying hospitals into volume quartiles. Agreement was evaluated using the Landis scale: Values less than 0.00 suggest poor agreement, 0.00–0.20 slight agreement, 0.21–0.40 fair agreement, 0.41–0.60 moderate agreement, 0.61–0.80 substantial agreement, and 0.81–1.00 almost perfect agreement.[22]

All statistical analyses were performed using Stata MP, version 15.1.

LIMITATIONS This study had several limitations. First, our decision to limit this analysis to a single surgery could lead to questions regarding the generalizability of our findings. However, the extent to which shrinkage estimators distort hospital profiling is a function of the strength of the volume-outcome association, not the procedure itself. Thus, we would expect to see similar findings for other procedures and conditions that exhibit a strong volume-outcome association, such as coronary artery bypass graft surgery, mitral valve replacement, lower extremity bypass, and acute myocardial infarction.[3,8]

Second, we elected to include all hospitals in our analysis as opposed to excluding hospitals with fewer than twenty-five cases per year. We included these very-low-volume hospitals because excluding them would have caused one-fourth of the hospitals to be excluded—which arguably would have caused a large blind spot in a measurement system for aortic valve replacements. We believe that the performance of very-low-volume hospitals should be reported to ensure transparency and accountability since it is precisely these lowest-volume hospitals that, as a group, have the worst outcomes. Furthermore,

our approach was consistent with the CMS-commissioned white paper, which recommended avoiding volume cutoffs in performance reporting.[10] As recommended by the white paper's expert panel,[10] our study explored the use of case volume as a shrinkage target and provided empirical evidence that this approach reduces some of the distortion of hospital profiling caused by standard shrinkage estimators. However, the use of shrinkage targets is not expected to address the uncertainty around the point estimates for hospital predicted-to-expected ratios for very-low-volume and low-volume hospitals better than does the use of conventional shrinkage estimators.

Finally, our examination of the comparative accuracy of the standard approach versus shrinkage targets across volume quartiles compared the predictions of these models to the observed mortality rate. In doing so, we assumed that the observed mortality rate was the true rate. Although we did not know the true stochastic process that generated deaths, and thus the true mortality rate, the sample size in each quartile was large enough to provide a reasonable approximation of the true mortality rate. Other approaches to examining model goodness of fit, such as the Hosmer-Lemeshow statistic,[23] also assume that the observed mortality rate is the true mortality rate.

Study Results

VOLUME-OUTCOME ASSOCIATION Patient demographic characteristics are shown in appendix exhibit A4.[12] Patients treated in hospitals in the lowest-volume quartile (those with fewer than 25 cases) had 2.6-fold higher odds of mortality (adjusted odds ratio: 2.61; 95% confidence interval: 2.19, 3.11; $p < 0.001$), compared to hospitals in the highest-volume quartile (those with at least 125 cases) (appendix exhibit A5).[12] Patients in hospitals with case volumes of 25–49 (quartile 2) had 1.75-fold higher odds of mortality (AOR: 1.75; 95% CI: 1.51, 2.04; $p < 0.001$), whereas patients in hospitals with case volumes of 50–124 (quartile 3) had nearly 1.5-fold higher odds of mortality (AOR: 1.48; 95% CI: 1.32, 1.65; $p < 0.001$), compared to the highest-volume hospitals.

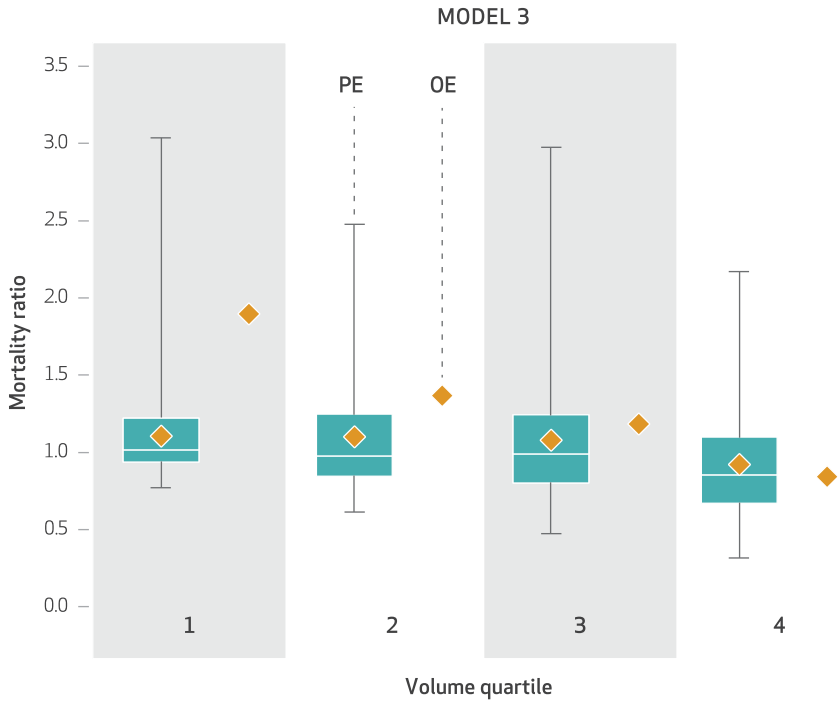
MODEL PERFORMANCE Both the standard shrinkage model (model 3) and the shrinkage targets model (model 4) demonstrated very good discrimination (C statistic: 0.80). Visual inspection of the calibration plots suggests that both models were well calibrated (appendix exhibit A6).[12] Separate calibration graphs for each of the four volume quartiles, however, suggest that the shrinkage targets model was better calibrated than the model without shrinkage targets in quartiles 1 and 2 (appendix exhibit A7).[12]

Appendix exhibit A8 examines model calibration in each volume quartile by comparing the overall observed mortality to the predicted mortality based on the models with shrinkage targets or without shrinkage targets. The exhibit shows that the risk-adjusted mortality rates based on the model without shrinkage targets substantially underestimated mortality in quartiles 1 and 2. By comparison, the risk-adjusted mortality rates based on the model with shrinkage targets much more closely approximated observed mortality rates for quartiles 1 and 2. Both models, with and without shrinkage targets, were well calibrated in quartiles 3 and 4.

DISTRIBUTION OF PREDICTED-TO-EXPECTED MORTALITY RATIOS Exhibit 1 compares the distribution of hospital predicted-to-expected ratios based on the model without shrinkage targets (model 3) to the overall observed-to-expected mortality ratio for each volume quartile. The medians of the hospital predicted-to-expected ratio for quartiles 1, 2, and 3 were substantially less than the overall observed-to-expected ratios

EXHIBIT 1

Box plot of hospital predicted-to-expected (PE) mortality ratios, based on model 3 (no shrinkage targets), and overall observed-to-expected (OE) mortality ratios for each volume quartile



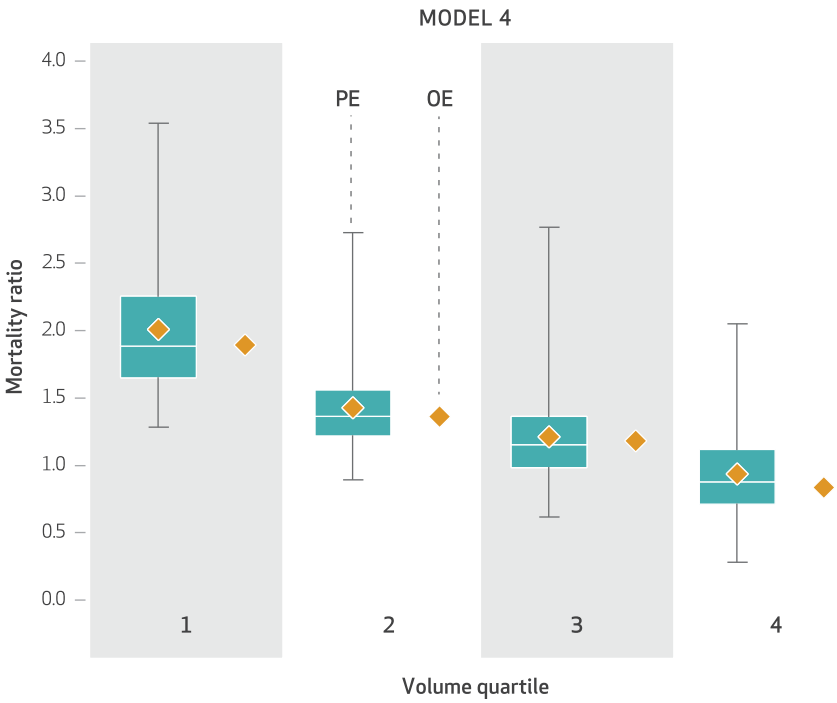
SOURCE Authors' analysis of data for 2013–15 from 100 percent Medicare Provider Analysis and Review files and the Master Beneficiary Summary File. **NOTE** The box plot displays the values for the minimum, first quartile, median, third quartile, and maximum values of the hospital PE mortality ratios.

for each of the hospital volume quartiles ($p < 0.001$): While the observed-to-expected ratios for quartiles 1, 2, and 3 were 1.90, 1.37, and 1.18, the median of the hospital predicted-to-expected ratios for quartiles 1, 2, and 3 were 1.02, 0.98, and 0.99, respectively. These findings suggest that hospital predicted-to-expected ratios based on the baseline model without shrinkage targets tended to underestimate the mortality of patients treated in low-volume hospitals and that the degree of underestimation was greatest in the lowest-volume quartiles.

Exhibit 2 compares the distribution of hospital predicted-to-expected ratios based on the model with shrinkage targets (model 4) to the overall observed-to-expected mortality ratio for each volume quartile. In this case, the medians for the hospital predicted-to-expected ratios for all four quartiles were not significantly different than the observed-to-expected ratios ($p > 0.05$). Together, these findings suggest that predicted-to-expected ratios based on shrinkage targets more accurately reflect hospital performance than predicted-to-expected ratios based on the baseline model without shrinkage targets.

EXHIBIT 2

Box plot of the hospital predicted-to-expected (PE) mortality ratios, based on model 4 (with shrinkage targets), and overall observed-to-expected (OE) mortality ratios for each volume quartile



SOURCE Authors' analysis of data for 2013–15 from 100 percent Medicare Provider Analysis and Review files and the Master Beneficiary Summary File. **NOTE** The box plot displays the values for the minimum, first quartile, median, third quartile, and maximum values of the hospital PE mortality ratios.

DISTRIBUTION OF HOSPITAL STAR RATINGS Appendix exhibit A8 displays the range of risk-adjusted mortality rates for hospitals across the five star-rating categories.[12] Exhibit 3 displays hospital star ratings based on risk-adjusted mortality rates estimated using shrinkage targets versus ratings estimated with no shrinkage targets. Overall, kappa analysis revealed moderate agreement (kappa = 0. 51) between the two sets of star ratings, which disagreed 14.3 percent of the time. Exhibit 4 and appendix exhibit A10[12] also display hospital star ratings based either on shrinkage targets or on no shrinkage targets, but stratified by hospital volume quartile. Hospital star ratings based on shrinkage targets exhibited only slight agreement with ratings based on no shrinkage targets in quartile 1 (kappa = 0.089), and moderate-to-substantial agreement for quartiles 2 (kappa = 0.64), 3 (kappa = 0.61), and 4 (kappa = 0.55) (appendix exhibit A10).[12] Depending on whether shrinkage targets or no shrinkage targets were used to classify hospitals, star ratings were discordant 29.1 percent of the time in quartile 1, 8.5 percent in quartile 2, 10.4 percent in quartile 3, and 11.3 percent in quartile 4. Of the ninety-three very-low-volume hospitals classified as three stars using standard shrinkage, thirteen were classified as one star and forty-three as two stars using shrinkage targets (data not shown).

COMPARISON OF RISK-ADJUSTED MORTALITY RATES AND OUTLIER STATUS Overall, there was substantial agreement between hospital risk-adjusted mortality rates based on standard shrinkage versus shrinkage targets (intraclass correlation coefficient: = 0.64) (appendix exhibit A11).[12] Agreement was only slight for quartile 1 (ICC: 0.085) but was substantial for quartile 2 (ICC: 0.64) and nearly perfect for quartiles 3 (ICC: 0.91) and 4 (ICC: 0.99).

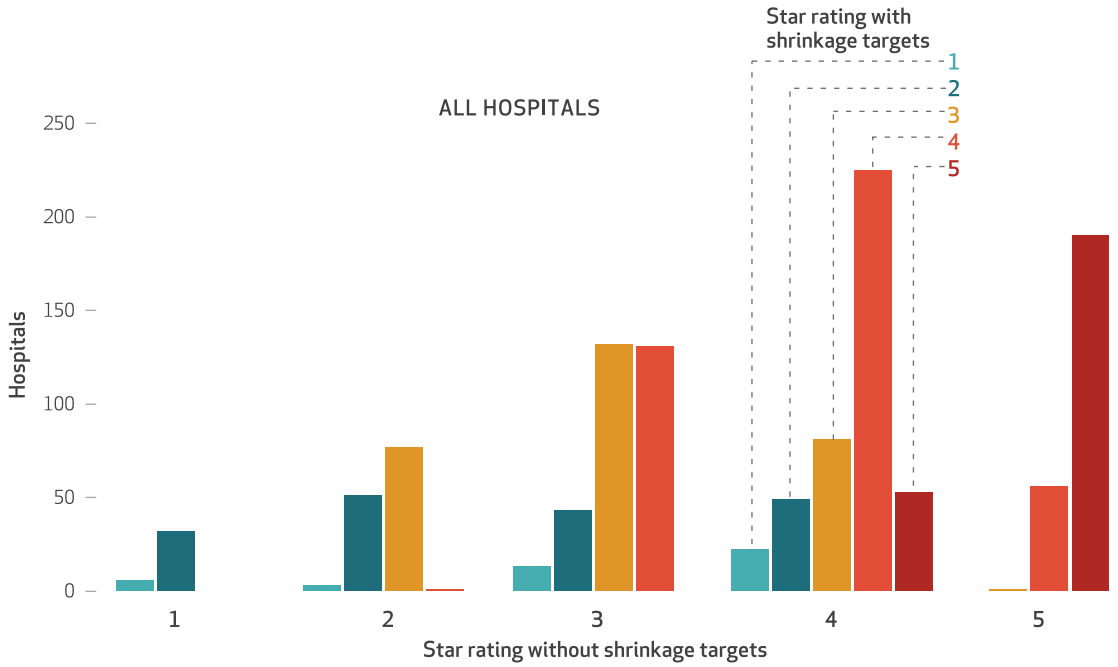
Very few hospitals were identified as performance outliers: 99.1 percent of hospitals were classified as average using standard shrinkage, and 98.7 percent were classified as average using shrinkage targets. The level of agreement between these two approaches was substantial (kappa = 0.77).

Discussion

Because information on hospital performance is at the center of efforts to redesign the US health care system, the accuracy of performance measurement is of paramount importance. One of the principal criticisms of the CMS approach is that it "masks [the] performance of small hospitals"[10[p2]] and may provide misleading information to patients, referring physicians, and third-party payers.[2] CMS uses a statistical methodolo-

EXHIBIT 3

Numbers of hospitals by star ratings with and without shrinkage targets



SOURCE Authors' analysis of data for 2013–15 from 100 percent Medicare Provider Analysis and Review files and the Master Beneficiary Summary File. **NOTES** The exhibit shows the numbers of hospitals, all four volume quartiles combined, that received each star rating with standard shrinkage (that is, no shrinkage targets). The colors indicate which of those hospitals had different star ratings with shrinkage targets. For example, 319 hospitals had three stars with standard shrinkage. Of these, 56 were reclassified as having one or two stars, and 131 as having four stars with shrinkage targets. Kappa = 0.51.

gy that calculates a hospital's performance as the weighted average of its own outcomes and the performance of average hospitals. The weight assigned to a hospital's actual outcomes in this calculation decreases as its case volume decreases. Although this approach is less likely to result in extreme values for hospital performance due to chance alone compared to older approaches based on nonhierarchical modeling, it shrinks low-volume providers to the national mean, ignoring the fact that for many conditions and surgeries, hospitals with smaller case volumes have worse outcomes than higher-volume hospitals.

We created two parallel sets of performance measures—one that assumed that low-volume hospitals are average (no shrinkage targets), and one that incorporated prior knowledge that low-volume hospitals have worse outcomes than high-volume hospitals (shrinkage targets). We included very-low-volume hospitals in our analyses because the CMS expert panel recommended that low-volume hospitals not be excluded from performance reporting.[10] We found that hospital predicted-to-expected mortality ratios based on hierarchical modeling without shrinkage targets were clustered around 1 for very-low-volume and low-volume hospitals because hierarchical modeling shrinks their performance to the performance of the average hospital. In theory, hospitals with predicted-to-expected ratios close to 1 should have mortality outcomes similar to those of an average hospital. However, the overall observed-to-expected ratios for patients undergoing aortic valve replacement surgery in very-low-volume and low-volume hospitals were not close to 1: Instead, they were 1.90 and 1.37, respectively. By comparison, the hospital predicted-to-expected ratios based on shrinkage targets for very-low-volume and low-volume hospitals were clustered around the overall observed-to-expected ratio. In other words, the hospital predicted-to-expected ratios based on shrinkage targets were consistent with the overall outcomes of patients treated in very-low-volume and low-volume hospitals, while hospital predicted-to-expected ratios based on the model without shrinkage targets provided an overly optimistic assessment of the performance of very-low-volume and low-volume hospitals.

CMS assigns hospitals an overall quality star rating of one to five stars.[11,24] This approach provides patients, physicians, and other stakeholders with information that allows them to differ-

EXHIBIT 4

Numbers of hospitals in the very-low-volume quartile by star ratings with and without shrinkage targets

SOURCE Authors' analysis of data for 2013–15 from 100 percent Medicare Provider Analysis and Review files and the Master Beneficiary Summary File. **NOTES** The exhibit shows the numbers of hospitals in the lowest quartile by volume (fewer than twenty-five cases) that received each star rating with standard shrinkage (that is, no shrinkage targets). The colors indicate which of those hospitals had different star ratings with shrinkage targets. There were no hospitals in this group with five stars. Kappa = 0.089.

entiate between high-performing hospitals (those with four or five stars) and average (three stars) and low-performing (one or two stars) hospitals. After grouping hospitals into star categories using the same cluster algorithm used by CMS,[11] we found that star ratings based on shrinkage targets exhibited moderate-to-substantial agreement for all but the lowest-volume hospitals. Star ratings for hospitals with case volumes of fewer than twenty-five were different nearly 30 percent of the time, depending on whether or not the star ratings were based on shrinkage targets. For example, of the ninety-three very-low-volume hospitals assigned three stars using the standard approach without shrinkage targets, fifty-six were classified as having one or two stars using shrinkage targets. Together, these findings suggest that the star ratings for all but the lowest-volume hospitals are not distorted when shrinkage targets are not used. Thus, the use of shrinkage targets may be important if CMS choses to measure the performance of hospitals with case volumes of fewer than twenty-five.

In addition to star ratings for overall quality, CMS separately reports the risk-adjusted outcome rates as either "no different than the national rate," "better than the national rate," or "worse than the national rate." This metric takes into account the statistical uncertainty related to the estimate of the risk-adjusted outcome, as compared to star ratings based only on the point estimates for the risk-adjusted outcome rate. But because more than 99.5 percent of hospitals are identified as average for reported conditions such as acute myocardial infarction,[3] this classification system conveys little information to patients and is not used by CMS for value-based purchasing. In our analysis, nearly 99 percent of the hospitals were classified as average either using or not using shrinkage targets. Although these two approaches exhibited substantial agreement when hospital ratings were based on outlier status, this is not surprising since both methods classified nearly all hospitals as average.

Since CMS uses risk-standardized outcome rates based on predicted-to-expected ratios for public reporting and value-based purchasing in programs such as the Hospital Readmissions Reduction Program[25] and the Comprehensive Care for Joint Replacement model,[26] we also investigated the impact of shrinkage targets on predicted-to-expected ratios. We found that these ratios showed poor agreement for very-low-volume hospitals, an intermediate level of agree-

ment for low-volume hospitals, and excellent agreement for high-volume and very-high-volume hospitals. In particular, the risk-adjusted mortality rates based on standard shrinkage consistently underestimated risk-adjusted mortality rates based on shrinkage targets for very-low-volume and low-volume hospitals. These findings are consistent with our main finding that star rankings based on standard shrinkage for very-low-volume hospitals provide an overly optimistic estimate of hospital performance.

The fact that the standard methodology used by CMS distorts the performance of low-volume hospitals has been previously reported by Silber and coauthors,[3] as well as others,[27,28] but it is not generally well understood by the medical or health care policy community. The use of shrinkage targets to address this limitation was identified as a top priority in the recent white paper commissioned by CMS.[10] However, to the best of our knowledge, measure developers have not submitted measures using shrinkage targets to the National Quality Forum for endorsement. In our study we operationalized shrinkage targets as described in the white paper and showed that the standard methodology does not introduce significant distortions in hospital rankings for most hospitals, with the exception of very-low-volume hospitals. Our study builds on prior work by Silber and coauthors,[3] as well as others,[27,28] by showing that shrinkage targets lead to major shifts in quality rankings for very-low-volume hospitals, compared to not using shrinkage targets. To the best of our knowledge, ours is the first study to demonstrate the practical implications of using shrinkage targets, compared to not using them.

In theory, one of the main advantages of hierarchical modeling is that the performance of very-low-volume hospitals can be measured. The current CMS practice of excluding these hospitals creates a blind spot in performance profiling and makes it more difficult for patients to make informed choices. Using the standard CMS methodology that does not incorporate shrinkage targets, however, distorts the performance of very-low-volume hospitals and may provide patients with potentially misleading information. Although shrinkage targets result in more accurate performance measures for very-low-volume hospitals, these hospitals may argue that incorporating hospital volume into performance profiling is unfair—since some small hospitals may deliver excellent care, and using shrinkage targets pulls their mortality rates upward. Since the true performance of individual very-low-volume hospitals may be unknowable because of sample-size issues, policy makers need to determine whether the benefit of providing patients with information on these hospitals outweighs the risk of misclassifying some of them as low quality. We believe that CMS has an obligation to the public to report the performance of very-low-volume hospitals for procedures where the risk of poor outcomes is especially high. But we also appreciate that such an approach may unintentionally misclassify some very-low-volume hospitals as low performance.

Conclusion

Our findings demonstrate the feasibility of implementing the recommendation made by CMS's expert panel tasked with addressing the criticism that the CMS methodology masks the performance of low-volume hospitals.[10] Our findings suggest that the use of shrinkage targets does not have a significant impact on the classification of hospitals with case volumes above twenty-five, which is the current cutoff used by CMS for public reporting. Our findings are particularly important in light of the ongoing controversy surrounding the use of hierarchical modeling in risk adjustment at the National Quality Forum and the strong recommendation to CMS by its expert panel to consider the option of using shrinkage targets.[10] ■

This work was supported by the National Institutes of Health (Grant Nos. R01 MD007662 and R01 MD012422). The funding source had no role in the conceptualization, design, or conduct of the study.

NOTES

1 Burwell SM. Setting value-based payment goals—HHS efforts to improve U.S. health care. N Engl J Med. 2015;372(10):897–9.
2 Glance LG, Joynt Maddox K, Johnson K, Nerenz D, Cella D, Borah B, et al. National Quality Forum guidelines for evaluating the scientific acceptability of risk-adjusted clinical outcome measures: a report from the National Quality Forum Scientific Methods Panel. Ann Surg. 2019 Dec 9. [Epub ahead of print].
3 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.
4 Krumholz HM, Brindis RG, Brush JE, Cohen DJ, Epstein AJ, Furie K, et al. Standards for statistical models used for public reporting of health outcomes: an American Heart Association Scientific Statement from the Quality of Care and Outcomes Research Interdisciplinary Writing Group: cosponsored by the Council on Epidemiology and Prevention and the Stroke Council. Circulation. 2006;113(3):456–62.
5 Snijders TAB, Bosker RJ. Multilevel analysis: an introduction to basic and advanced multilevel modeling. London: Sage Publications; 1999.
6 Birkmeyer JD, Siewers AE, Finlayson EV, Stukel TA, Lucas FL, Batista I, et al. Hospital volume and surgical mortality in the United States. N Engl J Med. 2002;346(15):1128–37.
7 Ross JS, Normand SL, Wang Y, Ko DT, Chen J, Drye EE, et al. Hospital volume and 30-day mortality for three common medical conditions. N Engl J Med. 2010;362(12):1110–8.
8 Reames BN, Ghaferi AA, Birkmeyer JD, Dimick JB. Hospital volume and operative mortality in the modern era. Ann Surg. 2014;260(2):244–51.
9 Chhabra KR, Dimick JB. Hospital networks and value-based payment: fertile ground for regionalizing high-risk surgery. JAMA. 2015; 314(13):1335–6.
10 Ash AS, Fienberg SE, Louis TA, Normand ST, Stukel TA, Utts J. Statistical issues in assessing hospital performance: commissioned by the Committee of Presidents of Statistical Societies [Internet]. Baltimore (MD): Centers for Medicare and Medicaid Services; [revised 2012 Jan 27; cited 2020 Mar 17]. Available from: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/Downloads/Statistical-Issues-in-Assessing-Hospital-Performance.pdf
11 Venkatesh AK, Bernheim SM, Qin L, Bao H, Simoes J, Wing M, et al. Overall hospital quality star rating on Hospital Compare methodology report (v3.0) [Internet]. New Haven (CT): Yale New Haven Health Services Corporation, Center for Outcomes Research and Evaluation; 2017 Dec [cited 2020 Mar 17]. Available for download from: https://www.qualitynet.org/files/5d0d3a1b764be766b0103ec1?filename=Star_Rtngs_CompMthdlgy_010518.pdf
12 To access the appendix, click on the Details tab of the article online.
13 Quan H, Sundararajan V, Halfon P, Fong A, Burnand B, Luthi JC, et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. Med Care. 2005;43(11):1130–9.
14 Hosmer DW, Lemeshow S. Applied logistic regression. 2nd ed. New York (NY): Wiley-Interscience Publication; 2000.
15 Royston P, Ambler G, Sauerbrei W. The use of fractional polynomials to model continuous risk variables in epidemiology. Int J Epidemiol. 1999;28(5):964–74.
16 Hospital case volume was specified in the model using two terms that each incorporated hospital case volume: log (hospital case volume/1,000) and (hospital case volume/1,000).
17 Krumholz HM, Wang Y, Mattera JA, Wang Y, Han LF, Ingber MJ, 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(13):1683–92.
18 Grady JN, Lin Z, Wang Y, Nwosu C, Keenan M, Bhat K, et al. 2013 measures updates and specifications: acute myocardial infarction, heart failure, and pneumonia 30-day risk-standardized mortality measure (version 7.0) [Internet]. Baltimore (MD): Centers for Medicare and Medicaid Services; 2013 Mar [cited 2020 Mar 17]. Available from: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/Mortality_AMI-HF-PN_Measures_Updates_Report_FINAL_06–13–2013.pdf
19 Kanungo T, Mount DM, Netanyahu NS, Piatko CD, Silverman R, Wu AY. An efficient k-means clustering algorithm: analysis and implementation. IEEE Trans Pattern Anal Mach Intell. 2002;24(7):881–92.
20 Pencina MJ, D'Agostino RB Sr. Evaluating discrimination of risk prediction models: the C statistic. JAMA. 2015;314(10):1063–4.
21 Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Fam Med. 2005;37(5): 360–3.
22 Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977; 33(1):159–74.
23 Hosmer DW Jr, Lemeshow S, Sturdivant RX. Applied logistic regression. 3rd ed. Hoboken (NJ): John Wiley and Sons; 2013. p. 153–226.
24 Chung JW, Dahlke AR, Barnard C, DeLancey JO, Merkow RP, Bilimoria KY. The Centers for Medicare and Medicaid Services hospital ratings: pitfalls of grading on a single curve. Health Aff (Millwood). 2019;38(9): 1523–9.
25 McIlvennan CK, Eapen ZJ, Allen LA. Hospital Readmissions Reduction Program. Circulation. 2015;131(20): 1796–803.
26 Centers for Medicare and Medicaid Services. Overview of CJR quality measures, composite quality score, and pay-for-performance methodology [Internet]. Baltimore (MD): CMS; [cited 2020 Mar 17]. Available from: https://innovation.cms.gov/Files/x/cjr-qualsup.pdf
27 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(2):264–9.
28 Sosunov EA, Egorova NN, Lin HM, McCardle K, Sharma V, Gelijns AC, et al. The impact of hospital size on CMS hospital profiling. Med Care. 2016;54(4):373–9.