ORIGINAL ARTICLE

A Comparison of statistical methods for hospital performance assessment

Xiaoting Wu*[1], Min Zhang[2], Ruyun Jin[3], Gary L. Grunkemeier[3], Charles Maynard[4], Ravi S. Hira[5], Todd MacKenzie[6], Morley Herbert[7], Chang He[1], Sari D. Holmes[8], Michael P. Thompson[1], Donald S. Likosky[1] on behalf of the National Cardiac Surgery Quality IMPROVE Network

[1] *Department of Cardiac Surgery, University of Michigan, Ann Arbor, Michigan, United States*
[2] *Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, Michigan, United States*
[3] *Center for Cardiovascular Analytics, Research and Data Science, Providence Heart Institute, Providence St. Joseph Health, Portland, Oregon, United States*
[4] *Department of Health Services, University of Washington and the Foundation for Health Care Quality, United States*
[5] *Division of Cardiology, University of Washington, Seattle, Washington. Foundation for Health Care Quality, Seattle, Washington, United States*
[6] *Geisel School of Medicine at Dartmouth College, Lebanon, New Hampshire, United States*
[7] *HCA Healthcare, Medical City Dallas Hospital, Dallas, Texas, United States*
[8] *Maryland Cardiac Surgery Quality Initiative (MCSQI), Baltimore, Maryland; Division of Cardiac Surgery, University of Maryland School of Medicine, Baltimore, Maryland, United States*

Received: March 3, 2021 **Accepted:** May 5, 2021 **Online Published:** May 24, 2021
DOI: 10.5430/jha.v10n3p32 **URL:** https://doi.org/10.5430/jha.v10n3p32

ABSTRACT

During hospital quality improvement activities, statistical approaches are critical to help assess hospital performance for benchmarking. Current statistical approaches are used primarily for research and reimbursement purposes. In this multi-institutional study, these established statistical methods were evaluated for quality improvement applications. Leveraging a dataset of 42,199 patients who underwent coronary artery bypass grafting surgery from 2014 to 2016 across 90 hospitals, six statistical approaches were applied. The non-shrinkage methods were: (1) indirect standardization without hospital effect; (2) indirect standardization with hospital fixed effect; (3) direct standardization with hospital fixed effect. The shrinkage methods were: (4) indirect standardization with hospital random effect; (5) direct standardization with hospital random effect; (6) Bayesian method. Hospital performance related to operative mortality and major morbidity or mortality was compared across methods based on variation in adjusted rates, rankings, and performance outliers. Method performance was evaluated across procedure volume terciles: small (< 96 cases/year), medium (96-171), and large (> 171). Shrinkage methods reduced inter-hospital variation (min-max) for mortality (observed: 0%-10%; adjusted: 1.5%-2.4%) and major morbidity or mortality (observed: 2.6%-35%; adjusted: 6.9%-17.5%). Shrinkage methods shrunk hospital rates toward the group mean. Direct standardization with hospital random effect, compared to fixed effect, resulted in 16.7%-38.9% of hospitals changing quintile mortality ranking. Indirect standardization with hospital random effect resulted in no performance outliers among small and medium hospitals for mortality, while logistic and fixed effect methods identified one small and three medium outlier hospitals. The choice of statistical method greatly impacts hospital ranking and performance outlier' status. These findings should be considered when benchmarking hospital performance for hospital quality improvement activities.

Key Words: Quality improvement, Statistical methodologies, Cardiac surgery

*Correspondence: Xiaoting Wu, Ph.D.; Email: xiaotinw@med.umich.edu; Address: Section of Health Services Research and Quality, Department of Cardiac Surgery (5346 CVC), University of Michigan Medical School, Ann Arbor, MI 48109-5864, United States.

1. Introduction

Investigators have developed several statistical methods to address differences in patient demographics and health status to support fair comparisons when evaluating hospital performance.[1,2] These hospital comparisons have multiple uses, including: (1) supporting patient decision-making, (2) public reporting, and (3) healthcare reimbursement.[3,4] Evaluating the role of each of these identified statistical methods for hospital ranking[5] is important insight for foundationally supporting quality improvement activities.

Distinct from reimbursement purposes that result in penalty or rewards to hospitals identified as low and high-performance outliers, quality improvement seeks to learn and subsequently disseminate best practices from hospitals identified as high-performance outliers.[6] Established statistical methods to rank hospital performance include indirect standardization with each hospital's own case-mix and direct standardization with a same reference case mix for all hospitals.[2,7–9] These standardization methods are used in conjunction with various statistical models, including standard logistic regression, fixed effect or random effect modeling for risk adjustment.[2] For example, the Society of Thoracic Surgeons (STS) risk model for Coronary Artery Bypass Grafting (CABG) applies logistic regression models in conjunction with indirect standardization for hospital performance comparisons.[10] The STS also advocates a Bayesian approach that ultimately ranks hospitals with a typical case-mix for all hospitals.[11]

On the other hand, the Hospital Compare program from the Centers for Medicare and Medicaid Services uses hierarchical logistic regression models (random effect models) to calculate ratios of predicted to expected outcomes for a given hospital.[12,13] While random effect models result in shrinkage of hospital rates towards the group mean, especially for low volume hospitals,[14–17] fixed effect models may be more likely to identify performance outliers (i.e., false positive).[18,19] However, the application of these alternative modeling approaches may lead to divergent hospital ranking for the same hospital,[2,5,20] thus adversely impacting efforts to identify high and low-performance outliers to support quality improvement.

This study leveraged a clinical dataset reflecting isolated CABG procedures performed between 2014 and 2016 across 90 hospitals from 11 states. The aim of this study was to evaluate the existing statistical models in terms of their role to support hospital ranking and identification of low- and high-performance outliers for quality improvement.

2. Methods

This study was approved by the Institutional Review Board at the University of Michigan on 2/28/2017.

2.1 Data example and patient population

Clinical data were collected through each of the following quality collaboratives participating in the National Cardiac Surgery Quality IMPROVE Network:[21] The Cardiac Care Outcomes Assessment Program from Washington, Maryland Cardiac Surgery Quality Initiative, Michigan Society of Thoracic and Cardiovascular Surgeons, Northern New England Cardiovascular Disease Study Group, Heart Institute at Providence St. Joseph Health, and the Texas Quality Initiative. Data sharing was permitted through a data use agreement between the collaborative coordinating centers. Data were collected according to the STS Adults Cardiac Surgery Registry specifications.[10]

Missing data were handled following the previous STS risk models.[10] Statistical models were applied among a dataset representing isolated CABG procedures performed across 90 centers from the IMPROVE Network. Model development was performed among 42,199 procedures in 2014-2016, while hospital ranking was evaluated among 14,211 procedures in 2016.

2.2 Measures

Analyses leveraged two established outcome measures for hospital ranking: (1) operative mortality and (2) major morbidity or mortality.[10] The major morbidity or mortality measure was defined by STS and is composed of six component outcomes: (1) operative mortality; (2) permanent stroke (cerebrovascular accident); (3) renal failure; (4) prolonged ventilation (longer than 24 hours); (5) deep sternal wound infection; (6) reoperation for any reason. Risk factors included in the STS published mortality models were used for risk adjustment.[10] Hospitals were categorized into terciles based on their CABG volumes in 2016 (low: < 96, medium: 96-171, high: > 171) to estimate the effect of hospital procedural volume on hospital ranking.

2.3 Modeling approaches

2.3.1 *Risk-adjustment models*

Three broad categories of risk-adjustment models were implemented: (1) standard logistic regression models with no hospital effect, (2) fixed effect models accounting for hospital fixed effects, and (3) random effects models accounting for hospital random effects with empirical Bayes estimates.

Two stages of modeling were applied to address the potential instability of point estimates. We first used three years of data (2014-2016) to obtain patient-level coefficients, and subsequently used these coefficients to estimate hospital level effects in 2016 for hospital ranking. Details of the model evaluations for both outcomes are provided in Supplemental Table 1. The adjusted rates of each outcome were estimated using data from 2016.

2.3.2 *Estimating rates using direct and indirect standardization*

Rates of each outcome were estimated for each hospital using either direct or indirect standardization combined with the above risk-adjustment models.

Direct standardization assumes a same reference case-mix for every hospital, which is the combined patient sample from all 90 hospitals in our data. Two approaches for direct standardization were applied: (1) including hospital-specific fixed effect estimates (“Dir_fixed”) and (2) including hospital-specific random effect estimates, which also is known as shrinkage estimates (“Dir_random”).

Three approaches for indirect standardization were applied to generate an observed-to-expected (O/E) ratio or a predicted to expected (P/E) ratio. The expected number of events (E) for each hospital was the sum of the adjusted patients’ risk within each hospital combined with: (1) the median hospital effect from the fixed effect model (“Indir_fixed”), or (2) the mean hospital effect from the random effect model (“Indir_random”), or (3) the standard logistic model absent any hospital effect (“Indir_logit”). The observed number of events (O) in the O/E ratios was the sum of patients identified as having an outcome (operative mortality, major morbidity or mortality) within each hospital. Clopper-Pearson exact binomial confidence intervals were used to construct the 95% confidence interval (CI) for the O/E ratios. The predicted number of events (P) in the predicted to expected (P/E) ratios were calculated from the random effect models including hospital-specific random effects. Bootstrapping 95% CIs were constructed for the P/E ratio.[12] Indirect standardized rates for each hospital were subsequently calculated by multiplying the overall outcome rate with the hospital-specific O/E ratio (“Indir_fixed” or “Indir_logit”) or the hospital specific P/E ratio (“Indir_random”).

2.3.3 *Bayesian approach*

The Bayesian method (“Bayesian”) was implemented based on the STS approach.[11] Diffuse prior was specified for the parameters included in the models. Hospital performance was assigned to the average if risk-standardized rates were statistically indistinguishable from the average rate, based on the 95% Bayesian certainty criterion.

2.4 Comparison of modeling approaches

Several approaches were used to compare findings from the applied statistical models. First, the distribution and correlation of hospital-specific outcome rates were compared across statistical methods. Second, hospital rankings derived from the standardized rates were compared. Third, performance outliers were compared based on 95% confidence intervals. Specifically, a hospital was considered a “better hospital” with an outcome rate lower than the expected average rate if the 95% confidence interval of its O/E ratio or P/E ratio was lower than 1, while a hospital was considered a “worse hospital” with an outcome rate higher than the expected average rate if the 95% confidence interval of its O/E ratio or P/E ratio was above 1.

A *p*-value of less than .05 was considered significant for all two-tailed significance testing. Welch’s ANOVA was used to test the difference across hospital terciles. Pearson correlation coefficient (*r*) was used to quantify the correlation between the two rates. Detailed methods can be found in Appendix B (Supplemental Methods and Codes). Statistical analyses were conducted using SAS software, Version 9.4 (SAS Institute, Cary, NC) and *R* version 3.5.2 (The *R* Foundation for Statistical Computing, Vienna, Austria).

3. Results

3.1 Variation of outcomes across hospitals

From the 90 participating hospitals, the mean observed mortality was 2.1% (standard deviation (std): 1.8%; range: 0%–10%) mortality, and the mean observed major morbidity or mortality was 10.8% (std: 5.2%; range: 2.6%-35.0%). Small hospitals (min-max: 0%-10.0% for operative mortality; 2.6%-35% for major morbidity or mortality) had greater variation in outcomes when compared to larger hospitals (min-max: 0.3%-3.6% for operative mortality; 6.7%-15.6% for major morbidity or mortality), Supplemental Figure 1. Hospital operative mortality and, major morbidity or mortality rates did not differ across hospital volume terciles (Welch’s ANOVA *p*-value for mortality = .72; *p*-value for major morbidity or mortality = .22).

3.2 Methods with shrinkage estimates reduced hospital outcome variation

The methods that used shrinkage estimates (Indir_random, Dir_random, Bayesian) reduced the variation of standardized rates (see Figure 1). Based on the boxplot, for operative mortality, the non-shrinkage methods had a similar distribution of hospital standardized rates for each hospital tercile. For example, in the small hospital tercile, Indir_logit resulted in standardized rates ranging 0%-6.6%, while Indir_fixed had rates ranging 0%-6.6% and Dir_fixed ranging 0%-6.5%. The shrinkage methods also had a similar distribution of standardized rates for each hospital tercile, but had reduced outcome variation when compared to the non-shrinkage methods. In the small hospital tercile, the methods with shrinkage had similar distribution: Indir_random (standardized rates min-max: 1.6%-2.4%), Dir_random (1.6%-2.3%), and Bayesian (1.7%-2.3%) (see Figure 1A). Shrinkage methods also resulted in a reduction of hospital variation in major morbidity or mortality when compared to the non-shrinkage methods (see Figure 1B).

The shrinkage methods moved hospital standardized rates toward the average, particularly for small hospitals. For example, seven small hospitals had greater than 4% standardized mortality rates under the Indir_fixed method (a non-shrinkage method). After shrinkage under the Indir_random method, these hospitals had their standardized rates closer to the average rate of 2.1% (see Figure 2A). For hospitals with a zero mortality rate, their standardized rates from the Indir_random method were between 1.6% and 1.8%. Similar results were observed from direct standardization methods (Dir_fixed vs. Dir_random, Pearson correlation $r = 0.89$, see Figure 2B). A similar shrinkage effect was also observed for major morbidity or mortality which had higher event rates (see Figure 2C and 2D). Two small hospitals with the Indir_fixed standardized rates of 24% had the Inidr_random standardized rate of 15% after shrinkage. The correlation between the standardized rates remained high (Indir_random vs. Indir_fixed: $r = 0.96$; Dir_random vs. Dir_fixed: $r = 0.95$).

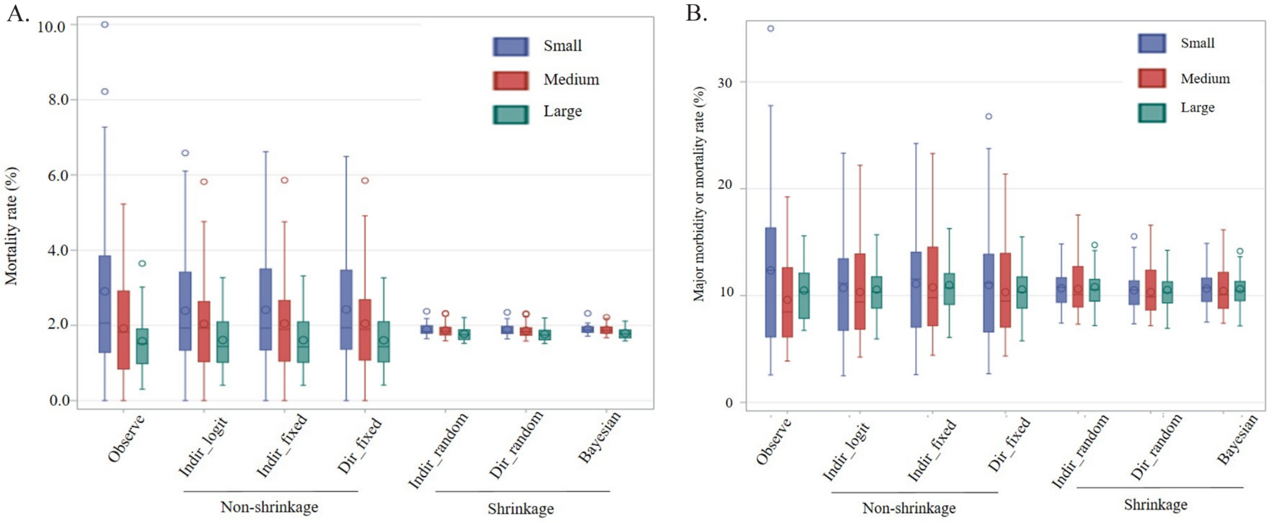

Figure 1. Distribution of standardized rates across statistical approaches
Indir_logit: indirect standardization with logistic regression models; Indir_fixed: indirect standardization with fixed effect models, Indir_random: indirect standardization with random effect models; Dir_fixed: direct standardization with fixed effect models; Dir_random: direct standardization with random effect models; Bayesian: Bayesian model. A) mortality; B) major morbidity or mortality. The line in the middle of each box is the median. The box represents the middle 50% of the data. The box edges are the 25th and 75th percentiles. The circle inside the box represents the mean. The circle outside the box represents the outliers. The colors represent the hospital size.

3.3 Methods with shrinkage change hospital rankings

The absolute hospital rankings changed as different methods changed the standardized rates (see Supplemental Figure 2). To quantify the change in rankings, hospitals were classified into quintiles based on their absolute rankings. Compared to the mortality quintiles rankings under the Dir_fixed method, 16.7% to 38.9% of hospitals in each quintile rank changed their quintile ranks when the Dir_random method was applied (see Table 1A). For example, 22.2% and 5.6% of hospitals in the 1st mortality quintile ranks under the Dir_fixed method were re-classified into the 2nd and 3rd quintile rank under Dir_random methods respectively, all of which were small hospitals. For major morbidity or mortality quintile rankings based on non-shrinkage methods (Dir_fixed), 5.6% to 27.8% of the hospitals in each quintile rank changed their ranks when shrinkage methods (Dir_random) were applied (see Table 1B). The changes across quintiles for the other methods are shown in Supplemental Table 2.

Small hospitals moved toward the middle in hospital ranking when the shrinkage methods were applied, resulting in fewer small hospitals at the top 10% and bottom 10% of the ranking. For example, methods without shrinkage (Indir_logit, Indir_fixed, Dir_fixed) had 5 small hospitals and 4 medium-size hospitals in the top 10% for performance, whereas 6 small and 3 medium-sized hospitals were in the bottom 10% of hospitals. Methods with shrinkage (Indir_random, Dir_random, Bayesian) had no small, 8 large, and 1 medium-sized hospital for performing in the top 10%, whereas 4 small, 3 medium, and 2 large-sized hospitals were in the bottom 10% ranking for mortality (Supplemental Table 3).

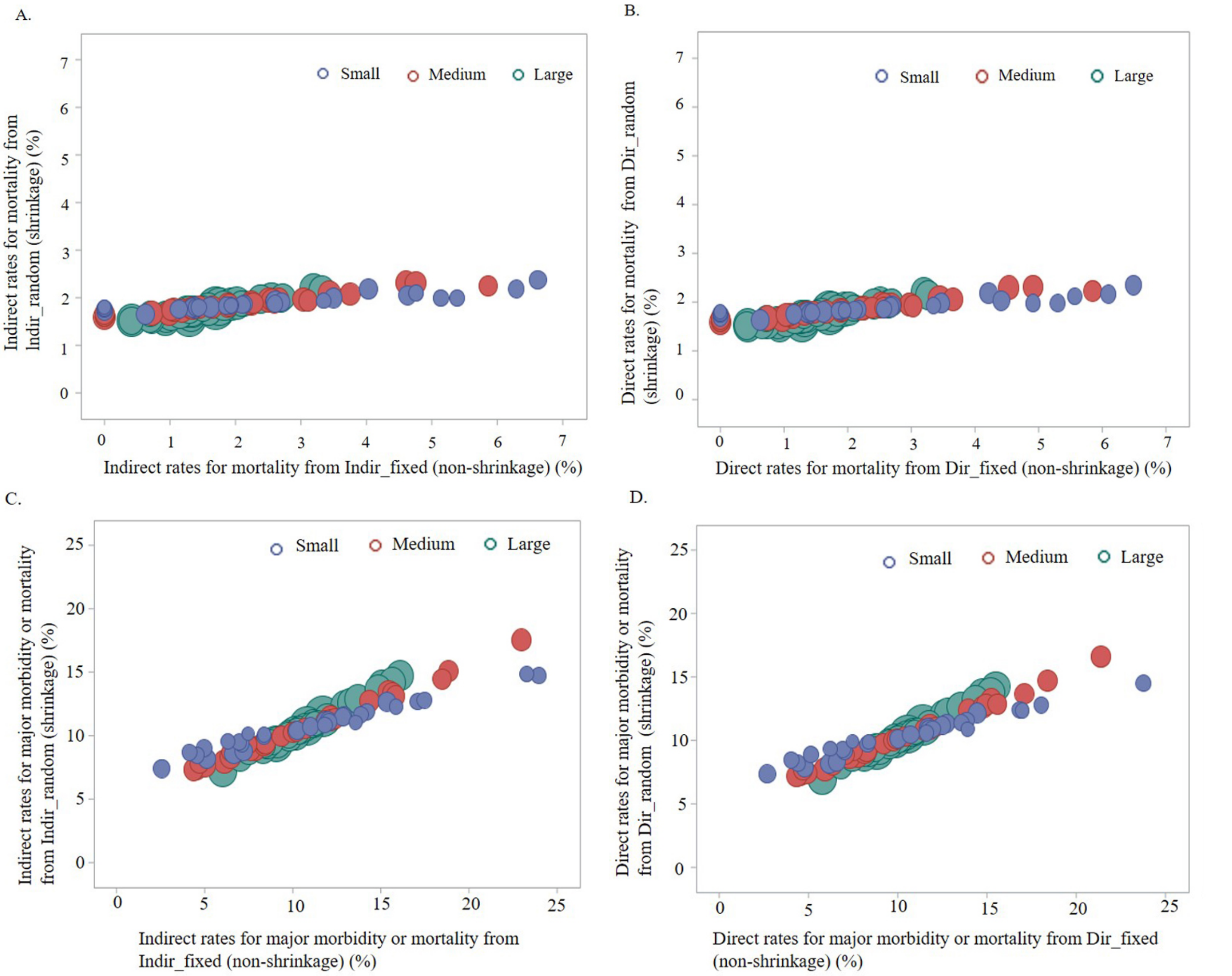

Figure 2. Shrinkage methods move small hospitals towards the average, and reduce hospital variation

Scatter plots of standardized rates across methods are shown. Each bubble represents a hospital; the bubble size reflects the hospital relative case volume. A) For mortality as the outcome, the correlation between indirect standardized rates from Indir_random (a shrinkage method) vs. Indir_fixed (a non-shrinkage method) is shown. r = 0.885 (p < .0001). B) For mortality as the outcome, the correlation between direct standardized rates from Dir_random (a shrinkage method) vs. Dir_fixed (a non-shrinkage method) is shown. r = 0.886 (p < .0001). C) For major morbidity or mortality as the outcome, the correlation between indirect standardized rates from Indir_random vs. Indir_fixed is shown. r = 0.956 (p < .0001). D) For major morbidity or mortality as the outcome, the correlation between direct standardized rates from Dir_random vs. Dir_fixed is shown. r = 0.95 (p < .0001).

3.4 Methods identified different performance outliers

To examine if different methods yielded distinct performance outliers, we compared the ratio of observed or predicted rates to the expected rates (assuming average hospital effect) in each method to identify performance outliers. For mortality, the Indir_random (P/E) method identified 5 large hospitals with significant lower predicted rates than expected (better hospitals), while the Indir_logit and Indir_fixed (O/E) methods identified 1 small hospital and 3 medium hospitals with significant higher observed rates than expected (worse hospitals) (see Supplemental Figure 3A). For the more common major morbidity or mortality outcomes, all three methods (Indir_logit, Indir_fixed, Indir_random) identified both small and large hospitals as performance outliers, but the identified hospital outliers varied by methods (Supplemental Figure 3B).

Table 1. Comparison of hospital quintile rankings based on direct standardized rates with and without shrinkage

A. Quintiles ranking changes for hospital mortality					
Quintiles based on Dir_random method with shrinkage	**Quintiles based on Dir_fixed method without shrinkage, no. (column %)**				
	1 “Low”	**2**	**3**	**4**	**5 “High”**
1 “Low”	13 (72.2%)	5 (27.8%)			
2	4 (22.2%)	11 (61.1%)	3 (16.7%)		
3	1 (5.6%)	2 (11.1%)	13 (72.2%)	2 (11.1%)	
4			2 (11.1%)	13 (72.2%)	3 (16.7%)
5 “High”				3 (16.7%)	15 (83.3%)
No. of Hospitals with changing rankings	5 (27.8%)	7 (38.9%)	5 (27.8%)	5 (27.8%)	3 (16.7%)
Small	5	2	0	2	1
Medium	0	1	0	0	2
Large	0	4	5	3	0
B. Quintiles ranking changes for hospital major morbidity or mortality					
Quintiles based on Dir_random method with shrinkage	**Quintiles based on Dir_fixed method without shrinkage, no. (column %)**				
	1 “Low”	**2**	**3**	**4**	**5 “High”**
1 “Low”	15 (83.3%)	3 (16.7%)			
2	3 (16.7%)	13 (72.2%)	2 (11.1%)		
3		2 (11.1%)	15 (83.3%)	1 (5.6%)	
4			1 (5.6%)	16 (88.9%)	1 (5.6%)
5 “High”				1 (5.6%)	17 (94.4%)
No. of Hospitals with changing rankings	3 (16.7%)	5 (27.8%)	3 (16.7%)	2 (11.1%)	1 (5.6%)
Small	3	2	0	1	1
Medium	0	1	1	0	0
Large	0	2	2	1	0

Table 2. Summary of surveyed statistical methods

Methods	**Indirect standardization with logistic model**	**Indirect standardization with fixed effect model**	**Indirect standardization with random effect model**	**Direct standardization with fixed effect model**	**Direct standardization with random effect model**	**Bayesian**
Abbreviation	Indir_logit	Indir_fixed	Indir_random	Dir_fixed	Dir_random	Bayesian
Model hospital effect	Logistic (no hospital effect)	Fixed effect	Random effect	Fixed effect	Random effect	Bayesian
Standardization	O/E, indirect	O/E, indirect	P/E, indirect	Direct	Direct	Direct
Shrinkage	No	No	Yes	No	Yes	Yes
Case-mix for standardization	Different	Different	Different	Same	Same	Same
Direct comparisons	No	No	No	Yes	Yes	Yes
Strengths	Simple	Control potential confounders between hospital and patients	Shrinkage; stable estimates for small hospitals	Direct comparisons	Direct comparisons, shrinkage; stable estimates for small hospitals	Full Bayesian, shrinkage, direct comparisons
Weakness	Hospital effect is not captured	Large standard error for small hospitals	Bias when hospital effect is correlated with patient effect; Minimize observation variation	Inaccurate representation when hospitals do not have same case mix; Large SE for small hospitals	Inaccurate representation when hospitals do not have the same case mix. Minimize observation variation	Inaccurate representation when hospitals do not have the same case mix. Minimize observation variation
Reference	STS national report[27], NYS Cardiac Surgery Report card[5]	CMS Dialysis Facility Reports[19]	CMS White Paper[12]	Pouw ME; Nicholl[9,24]	Dimick[16]	STS provider rating[11]

4. Discussion

In this study, we leveraged a clinical database of 90 hospitals participating in the IMPROVE network to evaluate the ability of different statistical methods to rank hospital performance for benchmarking quality improvement. Unlike previous studies that either used simulation data or administrative data to explore a limited number of methods for hospital comparisons, this study used real-world clinical data, ensured the same case-mix adjustment, and performed a thorough examination of existing methods for hospital comparisons in order to support quality improvement. Importantly, we revealed significantly different rankings and performance outliers from shrinkage versus non-shrinkage methods (see Figure 3). Under non-shrinkage methods, the hospital outcome variation is large. The small hospitals tend to rank as the "worse" hospitals with the standardized rates derived from the non-shrinkage methods. Under shrinkage methods, hospital outcome variation shrinks. Both small and large hospitals move toward the average, while the small hospitals typically move more than the large hospitals. We further summarized the pros and cons of these existing methods in Table 2.

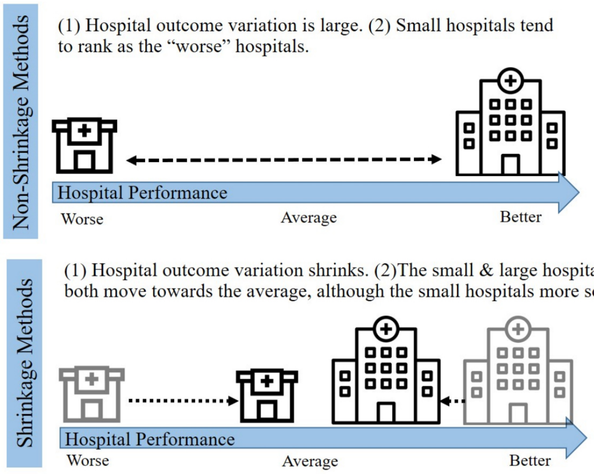
Non-Shrinkage Methods
(1) Hospital outcome variation is large. (2) Small hospitals tend to rank as the "worse" hospitals.
Hospital Performance
Worse
Average
Better
Shrinkage Methods
(1) Hospital outcome variation shrinks. (2)The small & large hospitals both move towards the average, although the small hospitals more so.
Hospital Performance
Worse
Average
Better

Figure 3. Summary of the impact on hospital rankings by methods

Non-shrinkage methods including Indir_logit, Indir_fixed, Dir_fixed have the following impact on hospital rankings: (1) the hospital outcome variation is large; (2) small hospitals tend to rank as the "worse" hospitals. Shrinkage methods including Indir_random, Dir_random, Bayesian have the following impact on hospital rankings: (1) hospital outcome variation shrinks; (2) the small & large hospitals both move towards the average, although the small hospitals more so.

Among the three types of statistical models we explored, the fixed effect and random effect models perform better, as assessed by the c-statistics, than standard logistic regression models. The fixed effect and random effect models additionally suggest that some variation in hospital outcomes may be attributed to the hospital level, given that patients are clustered within hospitals. In the case when there is a correlation between hospital effects and patient risk factors, random effect models cannot provide accurate estimates, while fixed effects models can yield unbiased estimates.[12,22,23] Unlike prior literature,[19] we obtained consistent results of patient-level effects from both the fixed and random effect models. For example, the model estimates from random-effects models were quantitatively similar to those from fixed effects models, which indicates that the hospital-level effects were unlikely confounding with the patient-level factors in our data.

Among the methods we surveyed, the P/E ratio derived from the random effect model (Indir_random), direct standardization rate from the random effect model (Dir_random), and the STS Bayesian method (Bayesian) yielded shrinkage of hospital variation. Indir_random and Dir_random used empirical Bayes estimators, and Bayesian used fully Bayesian estimators. These shrinkage methods provide substantially different hospital rankings compared to non-shrinkage methods. Shrinkage methods have a greater impact on small hospitals compared to large hospitals, and thus identify fewer small hospitals as performance outliers and results in fewer small hospitals ranking in the top and bottom 10%.

These results are consistent with the findings in the literature about shrinkage methods.[5,13,14] Dimick et al., utilizing the American College of Surgeons' National Surgical Quality Improvement Program, demonstrated the value of utilizing reliability adjustment to remove statistical noise in adjusted hospital outcomes reporting.[14] Glance et al. used data from the New York State Cardiac Surgery Database to evaluate different risk adjustment models for identifying performance outliers, including standard logistic regression, fixed effects or random effects modeling.[5] Glance found that the random effects models identified fewer performance outliers than the fixed effects or standard logistic regression approaches, in part due to the use of shrinkage estimators. Our study demonstrated a reduction of variability of direct standardized hospital outcome rates when using shrinkage estimates from random effects models. This shrinkage effect may underestimate the difference between hospital performance, particularly for smaller hospitals.

Direct standardization and indirect standardization have distinct interpretations in hospital ranking.[24] Direct standardization imposes the same reference population that allows direct hospital comparisons. When there is a substantial difference in case-mix across hospitals, this may not be a practical approach. Indirect standardization compares the hospital performance with its own case-mix, thus hospitals cannot compare directly with each other unless the two hospitals share a similar case-mix. Our results show that direct standardized rates have a strong correlation with indirect standardized rates (Dir_random vs. Indir_random, r = 0.998), and the hospitals ranking in the top and bottom 10% are almost identical in these two standardization methods (Supplemental Figure 4).

We acknowledge some limitations inherent in our study. First, as with any observational cohort study, our findings could be subject to unmeasured confounding at the patient and hospital level. The existing ranking methods are not able to capture hospital-level factors except for hospital volume. However, we have included all risk factors that the STS risk models consider. Second, as there is no single gold standard method for hospital ranking and true hospital performance is unknown, a simulation study would be needed to assess the ability of each method to identify true hospital performance. However, we emphasized that our findings, using real-world data, could provide guidance in real-world practice. Thirdly, we recognize that our findings may only be generalizable to our participating centers. The average hospital performance in this study was derived from models with the 90 hospitals in our data. But we consider this number of hospitals sufficient to discuss the difference in methods.

This present survey of established risk adjustment models has important implications for advancing quality improvement. It is important to recognize that there are no existing consensus criteria to guide which statistical approach to use for quality improvement benchmarking. Nonetheless, the present findings suggest that each method has both its strengths and limitations. Some methods (e.g., random effect model) are sensitive to hospital procedural volume,[25] resulting in large changes in both ranking and outlier status (see Table 1 and Figure 2). Nonetheless, other methods (e.g., fixed effect methods) are less impacted by hospital procedural volume although are susceptible to falsely identifying more performance outliers (e.g., false positives).[26] Whereas there are financial penalties associated with a hospital being identified as a low performance outlier within the setting of health policy, the consequences of site visiting a hospital mistakenly identified as high performance outlier are much smaller. We recommend that end users wishing to advance quality improvement activities consider the strengths and weaknesses of each statistical method given the potential for disparate findings, see Table 2.

5. CONCLUSIONS

Different choices of methods can result in changes to hospital rankings and misclassification of performance outliers' status. In this study, we have summarized the pros and cons of existing statistical methods in order to support quality improvement programs in choosing a suitable option for hospital benchmarking.

ACKNOWLEDGEMENTS

The authors acknowledge the contribution of Qixing Liang, PhD.

SOURCES OF FUNDING

Donald S. Likosky received extramural support from the Agency for Healthcare Research and Quality (AHRQ, R01HS026003) and the National Heart, Lung, and Blood Institute (NHLBI, R01HL146619). Opinions expressed in this manuscript do not represent those of the NHLBI, AHRQ or the US Department of Health and Human Services. Support for the MSTCVS Quality Collaborative is provided by the Blue Cross and Blue Shield of Michigan and Blue Care Network as part of the BCBSM Value Partnerships program.

CONFLICTS OF INTEREST DISCLOSURE

Although Blue Cross Blue Shield of Michigan and MSTCVS-QC work collaboratively, the opinions, beliefs and viewpoints expressed by the author do not necessarily reflect the opinions, beliefs and viewpoints of BCBSM or any of its employees.

REFERENCES

[1] Normand SLT, Shahian DM. Statistical and Clinical Aspects of Hospital Outcomes Profiling. Statistical Science: A Review Journal of the Institute of Mathematical Statistics. 2007; 22(2): 206-226. https://doi.org/10.1214/088342307000000096

[2] DeLong ER, Peterson ED, DeLong DM, et al. Comparing risk-adjustment methods for provider profiling. Statistics in Medicine. 1997; 16(23): 2645-2664. https://doi.org/10.1002/(SICI)1097-0258(19971215)16:23<2645::AID-SIM696>3.0.CO;2-D

[3] Rothberg MB, Morsi E, Benjamin EM, et al. Choosing the best hospital: the limitations of public quality reporting. Health Affairs . 2008; 27(6): 1680-1687. PMid: 18997226. https://doi.org/10.1377/hlthaff.27.6.1680

[4] Lingsma HF, Steyerberg EW, Eijkemans MJC, et al. Comparing and ranking hospitals based on outcome: results from The Netherlands Stroke Survey. QJM: Monthly Journal of the Association of Physicians. 2010; 103(2): 99-108. PMid: 20008321. https:

//doi.org/10.1093/qjmed/hcp169

[5] Glance LG, Dick A, Osler TM, et al. Impact of changing the statistical methodology on hospital and surgeon ranking: the case of the New York State cardiac surgery report card. Medical Care. 2006; 44(4): 311-319. PMid: 16565631. https://doi.org/10.1097/01.mlr.0000204106.64619.2a

[6] Rosenthal MB, Fernandopulle R, Song HR, et al. Paying for quality: providers' incentives for quality improvement. Health Affairs . 2004; 23(2): 127-141. PMid: 15046137. https://doi.org/10.1377/hlthaff.23.2.127

[7] Shahian DM, Normand SLT. What is a performance outlier? BMJ Quality & Safety. 2015; 24(2): 95-99. PMid: 25605952. https://doi.org/10.1136/bmjqs-2015-003934

[8] Shahian DM, Normand SLT. Comparison of "risk-adjusted" hospital outcomes. Circulation. 2008; 117(15): 1955-1963. PMid: 18391106. https://doi.org/10.1161/CIRCULATIONAHA.107.747873

[9] Pouw ME, Peelen LM, Lingsma HF, et al. Hospital standardized mortality ratio: consequences of adjusting hospital mortality with indirect standardization. PloS One. 2013; 8(4): e59160. PMid: 23593133. https://doi.org/10.1371/journal.pone.0059160

[10] Shahian DM, O'Brien SM, Filardo G, et al. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 1–coronary artery bypass grafting surgery. The Annals of Thoracic Surgery. 2009; 88(1 Suppl): S2-S22. PMid: 19559822. https://doi.org/10.1016/j.athoracsur.2009.05.053

[11] O'Brien SM, Shahian DM, DeLong ER, et al. Quality measurement in adult cardiac surgery: part 2–Statistical considerations in composite measure scoring and provider rating. The Annals of Thoracic Surgery. 2007; 83(4 Suppl): S13-S26. PMid: 17383406. https://doi.org/10.1016/j.athoracsur.2007.01.055

[12] Ash AS, Fienberg SF, Louis TA, et al. Statistical issues in assessing hospital performance. 2012. Available from: http://escholarship.umassmed.edu/cgi/viewcontent.cgi?article=2117&context=qhs_pp

[13] Silber JH, Rosenbaum PR, Brachet TJ, et al. The Hospital Compare mortality model and the volume-outcome relationship. Health Services Research. 2010; 45(5 Pt 1): 1148-1167. PMid: 20579125. https://doi.org/10.1111/j.1475-6773.2010.01130.x

[14] Dimick JB, Ghaferi AA, Osborne NH, et al. Reliability adjustment for reporting hospital outcomes with surgery. Annals of Surgery. 2012; 255(4): 703-707. PMid: 22388108. https://doi.org/10.1097/SLA.0b013e31824b46ff

[15] MacKenzie TA, Grunkemeier GL, Grunwald GK, et al. A primer on using shrinkage to compare in-hospital mortality between centers. The Annals of Thoracic Surgery. 2015; 99(3): 757-761. PMid: 25742812. https://doi.org/10.1016/j.athoracsur.2014.11.039

[16] Dimick JB, Staiger DO, Birkmeyer JD. Ranking hospitals on surgical mortality: the importance of reliability adjustment. Health Services Research. 2010; 45(6 Pt 1): 1614-1629. PMid: 20722747. https://doi.org/10.1111/j.1475-6773.2010.01158.x

[17] Thomas N, Longford NT, Rolph JE. Empirical Bayes methods for estimating hospital-specific mortality rates. Statistics in Medicine. 1994; 13(9): 889-903. PMid: 8047743. https://doi.org/10.1002/sim.4780130902

[18] Kalbfleisch JD, Wolfe RA. On monitoring outcomes of medical providers. Statistics in Biosciences. 2013; 5(2): 286-302. https://doi.org/10.1007/s12561-013-9093-x

[19] He K, Kalbfleisch JD, Li Y, et al. Evaluating hospital readmission rates in dialysis facilities; adjusting for hospital effects. Lifetime Data Analysis. 2013; 19(4): 490-512. PMid: 23709309. https://doi.org/10.1007/s10985-013-9264-6

[20] Siregar S, Groenwold RHH, Jansen EK, et al. Limitations of ranking lists based on cardiac surgery mortality rates. Circulation. Cardiovascular Quality and Outcomes. 2012; 5(3): 403-409. PMid: 22592754. https://doi.org/10.1161/CIRCOUTCOMES.111.964460

[21] IMPROVE Network. (n.d.). National Cardiac Surgery Quality (IMPROVE) Network. August 4, 2020. Available from: http://www.improvenetwork.org/

[22] Austin PC, Merlo J. Intermediate and advanced topics in multilevel logistic regression analysis. Statistics in Medicine. 2017; 36(20): 3257-3277. PMid: 28543517. https://doi.org/10.1002/sim.7336

[23] Bell A, Fairbrother M, Jones K. Fixed and random effects models: making an informed choice. Quality & Quantity. 2019; 53(2): 1051-1074. https://doi.org/10.1007/s11135-018-0802-x

[24] Julious SA, Nicholl J, George S. Why do we continue to use standardized mortality ratios for small area comparisons? Journal of Public Health Medicine. 2001; 23(1): 40-46. PMid: 11315692. https://doi.org/10.1093/pubmed/23.1.40

[25] Austin PC, Reeves MJ. Effect of provider volume on the accuracy of hospital report cards: a Monte Carlo study. Circulation. Cardiovascular Quality and Outcomes. 2014; 7(2): 299-305. PMid: 24619320. https://doi.org/10.1161/CIRCOUTCOMES.113.000685

[26] Austin PC, Alter DA, Tu JV. The use of fixed- and random-effects models for classifying hospitals as mortality outliers: a Monte Carlo assessment. Medical Decision Making: An International Journal of the Society for Medical Decision Making. 2003; 23(6): 526-539. PMid: 14672113. https://doi.org/10.1177/0272989X03258443

[27] Society of Thoracic Surgeons ACSD Sample Data Analysis Report. (n.d.). Society of Thoracic Surgeons. October 21, 2020. Available from: https://www.sts.org/sites/default/files/documents/STS-Adult_SampleDAR_Blank.pdf