

# Impact of Changing the Statistical Methodology on Hospital and Surgeon Ranking

## *The Case of the New York State Cardiac Surgery Report Card*

Laurent G. Glance, MD,\* Andrew Dick, PhD,† Turner M. Osler, MD, FACS,‡ Yue Li,† and Dana B. Mukamel, PhD§

**Background:** Risk adjustment is central to the generation of health outcome report cards. It is unclear, however, whether risk adjustment should be based on standard logistic regression, fixed-effects or random-effects modeling.

**Objective:** The objective of this study was to determine how robust the New York State (NYS) Coronary Artery Bypass Graft (CABG) Surgery Report Card is to changes in the underlying statistical methodology.

**Methods:** Retrospective cohort study based on data from the NYS Cardiac Surgery Reporting System on all patient undergoing isolated CABG surgery in NYS and who were discharged between 1997 and 1999 (51,750 patients). Using the same risk factors as in the NYS models, fixed-effects and random-effects models were fitted to the NYS data. Quality outliers were identified using 1) the ratio of observed-to-expected mortality rates (O/E ratio) and confidence intervals (CIs) calculated using both parametric (Poisson distribution) and nonparametric (bootstrapping) techniques; and 2) shrinkage estimators.

**Results:** At the surgeon level, the standard logistic regression model, the fixed-effects model, and the fixed-effects component of the random-effects model demonstrated near-perfect agreement on the identity of quality outliers using a quality indicator based on the O/E ratio and the Poisson distribution. Shrinkage estimators identified the fewest outliers, whereas the O/E ratios with bootstrap CI identified the greatest number of outliers. The results were similar for hospitals, except that the fixed-effects model identified more outliers than either the NYS model or the fixed-effects component of the random-effects model.

**Conclusion:** Shrinkage estimators based on random-effects models are slightly more conservative in identifying quality outliers compared with the traditional approach based on fixed-effects modeling and standard regression. Explicitly modeling surgeon provider effect (fixed-effects and random-effects models) did not significantly alter the distribution of quality outliers when compared with standard logistic regression (which does not model provider effect). Compared with the standard parametric approach, the use of a bootstrap approach to construct 95% confidence interval around the O/E ratio resulted in more providers being identified as quality outliers.

**Key Words:** outcome assessment, quality of care, quality assurance, statistical models, coronary artery bypass, health services research

(*Med Care* 2006;44: 311–319)

Health outcomes report cards are at the center of efforts to improve healthcare quality. Performance profiling has 2 primary objectives. First, to promote an efficient market economy in health care, which allows patients, referring physicians, and third-party payers to select physicians and hospitals using performance-based criteria<sup>1–5</sup>; and second, to serve as a “catalyst to stimulate and promote internal quality improvement at the level of the organizational provider.”<sup>6–8</sup>

Coronary artery disease (CAD) is the leading cause of deaths worldwide.<sup>9</sup> In the United States alone, over 500,000 people died of heart disease in 2002<sup>9</sup> and 314,000 patients underwent coronary artery bypass surgery (CABG) in 2000.<sup>10</sup> The clinical and economic burden of CAD has provided the impetus for the development of CABG surgery report cards in several states. The New York State (NYS) cardiac surgery report card<sup>11–13</sup> is one of the earliest and most highly respected health outcomes report card and is widely recognized as being methodologically rigorous. Nonetheless, important questions regarding the accuracy of this and other health outcomes report cards remain to be answered.

There is a broad consensus on the need for risk adjustment when comparing outcomes across different hospitals and surgeons. However, there is a “risk” to risk adjustment because different risk-adjustment models will not always agree on the identity of high- and low-performance hospi-

From the Departments of \*Anesthesiology and †Community and Preventive Medicine, the University of Rochester School of Medicine and Dentistry, Rochester, New York; the ‡Department of Surgery, the University of Vermont Medical College, Burlington, Vermont; and the §Center for Health Policy Research, University of California, Irvine, California.

Supported by a grant from the Agency for Healthcare and Quality Research (RO1 HS 13617).

The views presented in this manuscript are those of the authors and may not reflect those of Agency for Healthcare and Quality Research or of the New York State Department of Health or of the Cardiac Advisory Committee.

Reprints: Laurent G. Glance, MD, Department of Anesthesiology, University of Rochester Medical Center, 601 Elmwood Avenue, Box 604, Rochester, NY 14642. E-mail: Laurent\_Glance@urmc.rochester.edu

Copyright © 2006 by Lippincott Williams & Wilkins

ISSN: 0025-7079/06/4404-0311

tals.<sup>14,15</sup> Critics of performance profiling frequently focus on the question of whether existing risk-adjustment models adequately capture patient risk. This is important because model selection—determining which risk factors to include in the final model—is necessarily subjective.<sup>16</sup>

Several studies have suggested that the statistical methodology used to create health outcomes report cards may be another important source of bias.<sup>17–21</sup> Our goal in this study was to describe the implications of statistical modeling on quality assessment and to determine how robust the NYS Cardiac Surgery Report Card is to the choice of statistical methodology. Our study recreated the NYS Report Card using the original NYS risk-adjustment models and data from the NYS Cardiac Surgery Reporting System (CSRS) database. We then examined the sensitivity of the outlier status (high quality, medium quality, low quality) for hospital and surgeons to the choice of statistical model: standard logistic regression, fixed-effects and random-effects modeling. We also determined whether various approaches to constructing “outlier” confidence intervals led to hospitals and surgeons being classified differently.

## METHODS

### Data

This study is based on data from the NYS Cardiac Surgery Reporting System and includes all patients undergoing isolated CABG surgery in NYS and who were discharged between 1997 and 1999 (51,750 patients; 2.20% mortality). The CSRS database includes the following data fields: patient demographics, hospital and surgeon identifiers (encrypted), preoperative clinical risk factors, and in-hospital mortality. These data were collected prospectively at the hospital level and were then submitted to the NYS Department of Health. Audit mechanisms were in place to ensure the validity of the data.<sup>22</sup>

### Risk-Adjustment Models

The outcome variable was in-hospital mortality. We recognize that 30-day mortality is preferable to in-hospital mortality as the outcome variable because discharge policies vary across hospitals. Use of in-hospital mortality, however, does not limit our ability to compare the performance of different risk-adjustment methods. Three different types of models were fit to the data: standard regression, fixed-effects and random-effects models. We first reproduced the NYS Department of Health (DOH) models using standard logistic regression based on the same data used by NYS to create its Cardiac Surgery Report Card. The “hospital” model, used to create the hospital report card, was based on data from 1999 alone, whereas the “surgeon” model, used to create the surgeon report card, was based on data between 1997 and 1999. We used the same risk factors as those used by NYS DOH to create its report cards; the list of risk factors for the hospital model was available at the NYS Department of Health web site<sup>23</sup> and the list of risk factors for the surgeon model was obtained from the NYS DOH (personal communication). These models are identical to the NYS models used to produce the NYS Cardiac Surgery Report Cards. Henceforth, we refer to these models as the NYS DOH models.

Second, we estimated a “hospital” fixed-effects model using STATA SE/8.2 (STATA Corp., College Station, TX), which included a separate hospital identifier for each hospital in the dataset (the first hospital was omitted). This model contained the same risk factors as the DOH NYS hospital model and was based on data from 1999. We also estimated a separate “surgeon” fixed-effects model using data from 1997–1999 and using the risk factors in the NYS “surgeon” model.

Third, we estimated a “hospital” random intercept model using the SAS (SAS Corp, Cary, NC) macro PROC GLIMMIX.<sup>24</sup> This model contained the same risk factors as the DOH NYS hospital model and was based on data from 1999. To quantify the contribution of provider effect to outcome, we calculated the intraclass correlation coefficient (ICC).<sup>25</sup> We also estimated a separate “surgeon” random intercept model using data from 1997–1999 and using the risk factors in the NYS “surgeon” model. The surgeon models are shown in Appendix A (the hospital models are available on request).

Model discrimination was evaluated using the C statistic.<sup>26</sup> Calibration is usually assessed using the Hosmer-Lemeshow statistic.<sup>26</sup> The Hosmer-Lemeshow statistic, however, explicitly assumes that all of the observations are independent. Because it is not known how robust the Hosmer-Lemeshow statistic is to the violation of the independence assumption, we assessed model calibration by constructing calibration curves that graphically compare observed mortality rate (OMR) and expected mortality rate (EMR) across deciles of risk. The EMR was obtained by averaging the predicted probability of death for all of the patients treated by the same provider.

### Identification of Quality Outliers

We examined 5 different approaches for identifying hospital and surgeon quality outliers (Table 1).

The first approach, used by the NYS DOH, is based on the ratio of the OMR to the EMR for each provider (O/E ratio). The 95% confidence interval around the OMR is based on the Poisson distribution.<sup>27</sup> (The NYS DOH uses the Byar approximation of the Poisson distribution<sup>28</sup>; the Byar approximation of the Poisson distribution and the exact method yield nearly identical results.) The EMR is treated as constant. The lower bound and the upper bound of the 95% CI for the OMR are then divided by the EMR to obtain the 95% CI for the O/E ratio. The O/E ratio for a given provider is then multiplied by the overall mortality rate for the entire patient cohort in NYS

TABLE 1. Methods for Identifying Quality Outliers

Model Type	Approach	Methods for Identifying Quality Outliers
Standard logistic regression	1	O/E ratio and Poisson CI
	2	O/E ratio bootstrap CI
Fixed-effects model	3	O/E ratio and Poisson CI
	4	O/E ratio and Poisson CI
Random-intercept model	5	Random provider intercept term (shrinkage estimator)

CI indicates confidence interval; O/E, observed-to-expected.

to obtain the risk-adjusted mortality rate (RAMR). Providers whose RAMR is statistically different from the average mortality rate in NYS are identified as quality outliers. Equivalently, we chose to label a provider whose O/E ratio is statistically different from 1 (95% confidence interval does not include 1) as a quality outlier.

The second approach is also based on the O/E ratio (and the NYS risk-adjustment models), except that the 95% confidence interval for the O/E ratio was calculated using bootstrapping.<sup>29</sup> Sampling with replacement was used to construct the bootstrap samples. The bootstrap samples were stratified within each provider to keep the distribution of patients by provider in the bootstrap samples identical to the distribution of patients within the original dataset. The OMR and the EMR were calculated for each bootstrap sample. The NYS DOH model was refit with each bootstrap sample and then used to calculate the EMR for each provider in the bootstrap sample.<sup>30</sup> One thousand replications were used to calculate the bias-corrected confidence interval<sup>31</sup> for each provider. Unlike the parametric approach used by the NYS DOH to calculate the 95% CI, which treats the EMR as fixed, the bootstrap approach allows for uncertainty in both the OMR and EMR.

Four of the surgeons meeting the NYS DOH volume criteria had observed mortality rates of zero. Bootstrap confidence intervals for the O/E ratio were undefined for these surgeons because each bootstrap sample would necessarily have an O/E ratio of zero. To create bootstrap confidence intervals for these surgeons, we changed the live/die status of the patient with the highest probability of death (for each surgeon) to die before estimating the bootstrap confidence interval. Surgeons with a zero OMR for whom the upper bound of the CI was less than 1 were classified as high-performance outliers.

The third approach for identifying quality outliers is based on the fixed-effects model. However, we could not use the intercept terms in the fixed-effects model to quantify outliers because these intercept terms represent a comparison to an arbitrarily chosen reference provider. Instead, the EMR for a specific provider was calculated using a weighted average of the other providers' effects. The probability of death for patient "i" treated by provider "j" was calculated as

$$\hat{p}_{ij} = \frac{1}{1 + e^{-\left(x_{ij}\hat{\beta}' + \sum_{k=1, k \neq j}^k \hat{\alpha}_k x_{ik}\right)}}$$

where  $x_{ij}$  is the vector of values for the patient-level explanatory variables ( $x_{1ij}, x_{2ij}, \dots, x_{pij}$ ),  $\hat{\beta}'$  is the vector of estimated coefficients for the patient-level explanatory variables ( $\beta_1, \beta_2, \dots, \beta_p$ ),  $\hat{\alpha}_k$  is the coefficient associated with the dummy variable for the  $k^{\text{th}}$  provider,  $\kappa$  is the number of providers,  $f_k$  is equal to  $n_k/(N - n_j)$ ,  $n_k$  is the number of patients treated by provider "k," and  $N$  is the total number of patients in the sample. Note that for providers with no deaths, the contribution of those providers to  $\hat{p}_{ij}$  is zero. Confidence intervals around the

point estimates for the O/E ratios based on the fixed-effects model were calculated using the exact Poisson distribution.

For the fourth approach, the point estimates and the 95% CI for the O/E ratio were based on the random-effects model

$$\hat{p}_{ij} = \frac{1}{1 + e^{-(x_{ij}\hat{\beta}' + \mu_{0j})}}$$

where  $\mu_{0j}$  is a random quantity corresponding to the  $j^{\text{th}}$  provider and has mean zero and variance  $\sigma_{\mu_0}^2$ . However, only the fixed-effects coefficients were used to calculate the probability of death for patient "i" treated by provider "j":

$$\hat{p}_{ij} = \frac{1}{1 + e^{-x_{ij}\hat{\beta}'}}$$

By only using the "fixed portion" of the model, we were able to estimate the EMR for a group of patients treated by an "average" provider. The rationale for using only the fixed coefficients is that the group intercept  $\mu_{0j}$  is equivalent to quality for a specific provider. The 95% confidence interval around the O/E ratio was based on the Poisson distribution.<sup>27</sup>

Using the fifth approach, quality outliers were identified using the full random-effects model. The shrinkage coefficient for each provider was exponentiated to obtain the adjusted odds ratio for each provider. Providers with odds ratios significantly greater than 1 were labeled as low-performance outliers, whereas providers with an odds ratio significantly less than 1 were classified as high-performance outliers.

All 33 hospitals in the NYS dataset were classified as high-performance, low-performance, or average-performance using the previously mentioned methods. Only surgeons meeting the NYS DOH "volume" criteria for quality reporting were assigned a quality ranking based on these methods (138 of 189 surgeons): "Surgeons who performed 200 or more isolated CABG operations between 1997 and 1999, and/or performed at least 1 isolated CABG operation in each of the years 1997–1999."<sup>23</sup> The percentage of the cases performed by the 51 surgeons who were not ranked by the NYS Report Card was 5.15%. Surgeon case volume for surgeons meeting NYS volume criteria ranged between 17 and 1044. Hospital case volume ranged between 80 and 1804.

### Analysis

We examined the impact of the previously mentioned statistical methodologies on the identification of quality outliers. In our baseline analysis, we compared the NYS DOH approach, based on standard logistic regression and exact Poisson CI, with a random-intercept model with shrinkage coefficients.

We then compared outlier status based on a fixed-effects model and Poisson CI to outlier status based on 1) standard model and Poisson CI; 2) the fixed-effects component of the random-intercept model and Poisson CI; and 3) the random-intercept model and shrinkage coefficients. We chose the fixed-effects model as the "gold standard" in these comparisons because they provide consistent estimates of the betas even when the provider effects are correlated with the

patient risk factors. This is not the case with either standard logistic regression or random-intercept models.<sup>32</sup> Quality metrics based on conventional and random-intercept models may both be inconsistent and incorrectly identify quality outliers because patient-level risk factors may not be independent of provider effect. To illustrate this, consider the possibility that older patients are more likely to be treated by high-quality providers. Then, a standard model that includes only patient-level risk factors will result in an inconsistent estimate of the age coefficient because provider effect will be incorporated in the estimate of the age coefficient. Alternatively, using a random-intercept model, that specifies both patient-level variables (ie, age) and provider intercepts, will also be inconsistent because the use of a random-intercept model assumes that patient-level and provider-level effects are independent. This bias may be removed by using a random-coefficients model in which the betas for patient-level explanatory variables are allowed to have a random component. However, the shrinkage coefficients in a random-coefficients model can no longer be used to “measure” provider quality because provider effect is no longer completely captured by the random-intercept terms.

We performed 3 different pairwise analyses to compare the different methodologies for identifying quality outliers. First, we calculated the kappa statistic for each pairwise comparison. Second, we calculated the intraclass correlation coefficient to assess the level of agreement between the point estimates of the O/E ratio obtained using the different methods. The derivation of the intraclass correlation coefficient<sup>33</sup> was based on a 2-way repeated-measures analysis of variance (ANOVA) model in which the O/E ratio was a function of 1) the provider and 2) of the approach used to estimate the O/E ratio. Finally, we compared the size of the confidence interval as a function of the method used to construct the confidence interval: 1) Poisson distribution, 2) normal distribution, and 3) bootstrapping; pairwise comparisons were performed using the sign test. This latter analysis was limited to comparing 1) the standard model/Poisson CI versus standard model/bootstrap CI; and 2) hierarchical model/Poisson CI versus hierarchical model/shrinkage estimator/normal distribution.

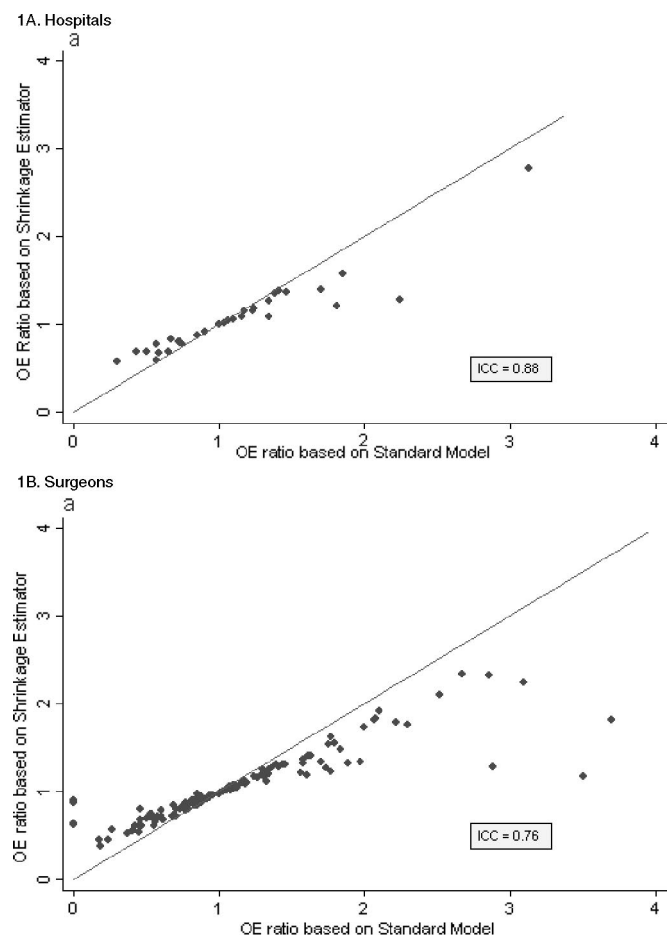
## RESULTS

All the models exhibited good discrimination (C statistic): NYS DOH hospital (standard) model, 0.794; fixed-effects hospital model, 0.812; random-intercept hospital model, 0.811; NYS DOH (standard) surgeon model, 0.800; fixed-effects surgeon model, 0.824; and random-intercept surgeon model, 0.821. Visual inspection of the calibration curves showed that these models also exhibited good calibration (available on request). The ICC for the random-intercept hospital and surgeon models were 0.049 and 0.059, respectively.

The plots (available on request) comparing O/E ratio based on the fixed-effects model to the O/E ratios based on 1) the standard model and 2) random-effects model (O/E ratio based on the patient-level variables only) shows a high level of agreement (nearly all the points fell on the 45° line); intraclass correlation coefficients, based on the ANOVA models, were very high (0.99–1.00). The plots for the 1)

baseline analysis (standard model versus shrinkage coefficients) (Fig. 1) and 2) fixed-effects model versus shrinkage estimators (available on request) reveal significant departures from the identity line; intraclass correlation coefficients ranged between 0.75 and 0.88. Point estimates for the O/E ratios based on the shrinkage coefficients are shifted upward for providers with O/E ratios less than 1 and are shifted downwards for providers whose O/E ratios are greater than 1. (We have assumed that the exponentiation of the shrinkage coefficient, which yields adjusted odds ratio for each provider, is approximately equal to the O/E ratio.)

Tables 2 and 3 show the outlier status of the hospitals and surgeons as a function of the method used to determine outlier status. Only those hospitals and surgeons whose outlier status varied as a function of the methodology used are displayed. Tabulation of the results for the surgeons of the pairwise comparisons for the different methods and the results of the kappa analysis are presented in Appendix B (hospital results are available on request).



**FIGURE 1.** Scatter plots of observed-to-expected (O/E) ratios as a function of the statistical model. The 45° line represents the line of identity. Note: point estimates for the O/E ratios based on the shrinkage coefficients are shifted upward for providers with O/E ratios less than 1 and are shifted downwards for providers whose O/E ratios are greater than 1. ICC indicates intraclass correlation coefficient.

**TABLE 2.** Surgeon Quality Outliers as a Function of the 5 Methods\*

Surgeon	Standard Model		Fixed Effects Model	Random Intercept	
	Poisson CI	Bootstrap CI	Poisson CI	Poisson CI	Shrinkage Estimators
1	L	L	L	L	
2		H			
3	L		L	L	
4		H			
5		H			
6		L			
7		H			
8		H	H	H	H
9		L			
10	H	H	H	H	
11		H			

\*Only surgeons for which 2 or more methods disagreed on outlier status are displayed. The 5 methods agreed on the outlier status of the other 127 surgeons (16 of which were identified as quality outliers).

H indicates high-quality outlier; L, low-quality outlier.

At the surgeon level, the NYS DOH (standard) model, fixed-effects model, and the fixed-effects component of the random-effects model demonstrated near-perfect agreement on the identity of quality outliers when the 95% CIs were based on the Poisson distribution (Table 2). Pairwise comparisons of the ICC for these models yielded ICC = 1. Shrinkage estimators identified the fewest outliers, whereas the O/E ratios based on the bootstrap CI identified the greatest number of outliers (Table 2). The finding that the shrinkage estimators led to fewer surgeons identified as outliers is a result of the “regression toward the mean” seen in Figure 1. Of note, shrinkage estimators identified the fewest outliers despite the fact that confidence intervals around these estimators (Fig. 2) were narrower than the CI based on the Poisson distribution or bootstrapping. The finding that bootstrapping resulted in the greatest number of outliers follows from the fact that the bootstrap CI were narrower than CI based on the Poisson distribution (sign test,  $P < 0.001$ ) (Fig. 2).

Interpretation of the level of agreement between the different methodologies when these were applied at the hospital level is difficult due to the smaller sample size of hospitals ( $n = 33$ ) as compared with surgeons (138). Despite excellent agreement on the point estimates for the O/E ratios, the fixed-effects model identified more outliers than either the

standard model or the random-effects model when the 95% CIs were based on the Poisson distribution (Table 3). Use of bootstrap confidence intervals resulted in slightly more hospitals being flagged as outliers, whereas shrinkage coefficients led to fewer hospitals being flagged as outliers (Table 3).

### DISCUSSION

Healthcare quality report cards will play a critical role in the effort to reshape healthcare delivery systems in this country. Although many physicians view performance profiling with skepticism,<sup>3</sup> it is likely that in the future third-party payers and patients will increasingly rely on report cards for choosing hospitals and physicians.<sup>4,34-36</sup> Therefore, it is imperative that the methodology used to produce quality report cards be as robust as possible.

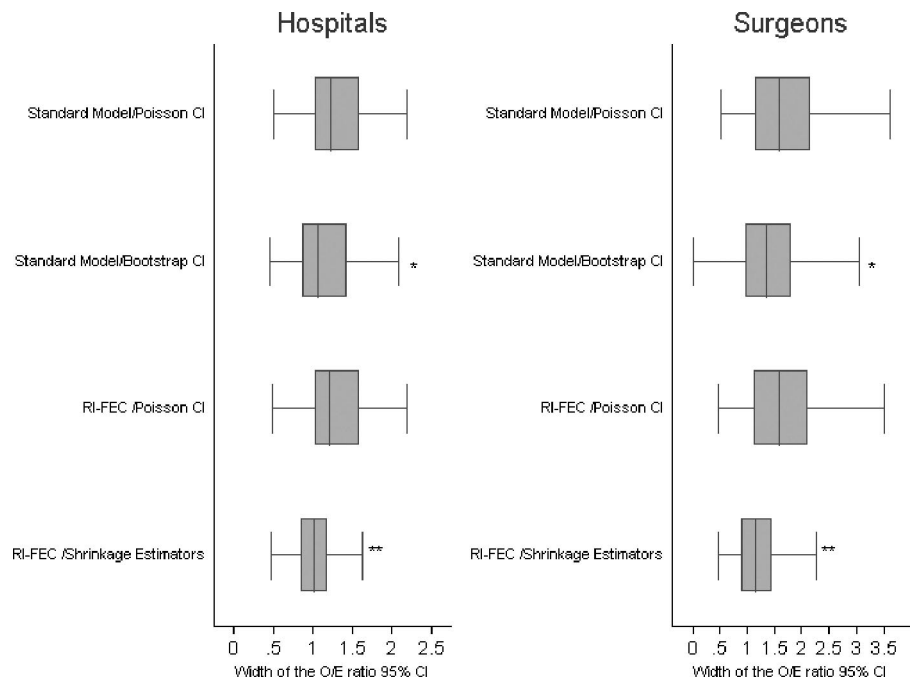
In this study, we examined whether different statistical methodologies affect which surgeons and hospitals are classified as high- and low-performance outliers in the publicly available NYS DOH CABG Surgery Report Card. Because of the relatively small sample size of hospitals ( $n = 33$ ) compared with surgeons ( $n = 138$ ), we base our discussion primarily on the findings for the surgeons. When quality outliers were identified based on the O/E ratio and Poisson

**TABLE 3.** Hospital Quality Outliers as a Function of the 5 Methods\*

Hospital	Standard Model		Fixed Effects Model	Random Intercept	
	Poisson CI	Bootstrap CI	Poisson CI	Poisson CI	Shrinkage Estimators
1			L		
2		H			
3	H	H	H	H	
4		L	L		

\*Only hospitals for which 2 or more methods disagreed on outlier status are displayed. The 5 methods agreed on the outlier status for the other 29 hospitals (2 of which were identified as quality outliers).

H indicates high-quality outliers; L, low-quality outliers.



**FIGURE 2.** Box plot of the width of 95% confidence interval around the observed-to-expected ratio as a function of the method used to identify hospital quality outliers (n = 33 hospitals) and surgeon quality outliers (n = 138 surgeons). \*Standard model/Poisson CI versus standard model/bootstrap CI,  $P < 0.001$ . \*\*Random intercept model/Poisson CI versus hierarchical model/shrinkage estimator,  $P < 0.001$ . RI-FEC indicates fixed-effects component of random intercept model; CI, confidence interval.

confidence intervals, we found nearly perfect agreement between the fixed-effects model, which we consider to be the gold standard, the standard model (used by the NYS DOH), and the fixed-effects component of the random-effects model. This finding is expected because surgeons and hospitals were found to have a small impact on CABG mortality. Using shrinkage estimators based on the random-effects model identified the fewest number of quality outliers, whereas using bootstrap confidence intervals resulted in the greatest number of providers identified as quality outliers.

There is an ongoing debate regarding whether risk adjustment should be based on random-effects modeling and shrinkage estimators. Currently available health outcomes report cards rely on risk-adjustment models constructed using standard logistic regression models to calculate the ratio of the expected mortality rate to the observed mortality rate. However, many experts have argued that this methodology is flawed and that risk adjustment should be based on random effects models.<sup>17,19,37-39</sup>

However, using random-effects modeling and shrinkage estimators, provider effects are “shrunk” back toward the overall mean for the cohort. Providers with low volumes or whose performance deviates more from the mean are “shrunk” more than high-volume providers or providers whose performance deviates less from the mean.<sup>40</sup> Shrinkage estimators have smaller standard errors and are thus more precise.<sup>40</sup> However, greater precision is achieved at the cost of introducing bias.<sup>40</sup> High-performing providers are reported too negatively and low-performance providers are reported too positively.<sup>40</sup> Thus, the use of shrinkage estimators to assess provider quality may also have substantial limitations.<sup>40-42</sup>

However, standard logistic regression and random-intercept models share an important limitation. Both assume

that patient risk factors are independent of provider effects. Because it is unlikely that this assumption is correct, standard logistic regression and random-effects models may be inconsistent and provide poor estimates of provider quality. Fixed-effects models, on the other hand, will result in unbiased estimates of provider effects when provider effects are correlated with observable patient risk factors.<sup>32</sup>

Marshall and Spiegelhalter<sup>43</sup> also compared random-effects modeling with the standard approach used by NYS to identify surgeon quality outliers. They found that 9 of the 16 surgeons identified by NYS as quality outliers were not classified as outliers using shrinkage estimators based on random-effects modeling. However, they did not have access to patient-level data for their analysis and instead used publicly available expected mortality rates for each surgeon as an index of patient severity. Thus, all of the patients treated by a particular surgeon were assigned an identical severity of illness. An earlier study by the same group looking at the NYS CABG Report Card was also based on pooled data.<sup>17</sup> DeLong et al<sup>18</sup> used data obtained from chart review to compare risk-adjustment methods for provider profiling in CABG surgery. In this study, based on 3654 patients treated by 28 surgeons, 3 surgeons were classified as outliers using a traditional approach based on standard logistic regression and the O/E ratio compared with 2 surgeons using shrinkage estimators and random-effects modeling.

Our study also explored the impact of using a nonparametric approach, bootstrapping, to construct confidence intervals around the point estimates of the O/E ratios. The nonparametric approach accounts for the uncertainty in both the observed mortality rate and in the estimated mortality rate and does not assume that the O/E ratio fits any particular statistical distribution. This is in contrast to the “conventional” approach, which assumes that the distribution of the

observed mortality rate can be approximated by either a Poisson or a normal distribution and which treats the expected mortality rate as constant. The bootstrap approach resulted in narrower confidence intervals and was consequently less conservative in classifying providers as quality outliers; a greater number of providers were identified as quality outliers using bootstrap CI compared with the method based on the Poisson distribution.

It should be emphasized that the goal of this study was not to determine whether the methodology used by the NYS DOH and other groups is sufficiently robust to be used to produce quality report cards for other medical conditions. Rather, this study provides a framework that can be used by analysts producing healthcare report cards to explore the sensitivity of their quality “grades” to the choice of statistical methodology. The extent to which the standard and random-effects models lead to inaccurate quality measures will depend on the magnitude of the provider contribution to outcome and the correlation between provider quality and patient risks. In this study, the small size of the ICC for the random-intercept models suggests that the provider contribution was relatively minor. Therefore, it is not surprising that quality metrics based on standard model, random-effects model, and the fixed-effects model showed a high level of agreement. This may not be the case for other medical conditions. Exploring the variability in “quality” as a function of statistical methodology is important given the increasing emphasis on “quality” as the basis for selective referral to high-quality providers,<sup>44</sup> selective avoidance of low-quality providers,<sup>45</sup> and financial incentives in pay-for-performance measures.

## CONCLUSION

Assigning outlier status to providers has potentially profound implications. Before publicly releasing quality report cards, it is incumbent on the analyst to explore the extent to which their findings vary using different statistical approaches. Further work is necessary to establish “best practices” for constructing report cards to ensure the validity of quality reporting.

## REFERENCES

- Schneider EC, Epstein AM. Influence of cardiac-surgery performance reports on referral practices and access to care. A survey of cardiovascular specialists. *N Engl J Med*. 1996;335:251–256.
- Schneider EC, Epstein AM. Use of public performance reports: a survey of patients undergoing cardiac surgery. *JAMA*. 1998;279:1638–1642.
- Hannan EL, Stone CC, Biddle TL, et al. Public release of cardiac surgery outcomes data in New York: what do New York state cardiologists think of it? *Am Heart J*. 1997;134:1120–1128.
- Mukamel DB, Mushlin AI. Quality of care information makes a difference: an analysis of market share and price changes after publication of the New York State Cardiac Surgery Mortality Reports. *Med Care*. 1998;36:945–954.
- Romano PS, Zhou H. Do well-publicized risk-adjusted outcomes reports affect hospital volume? *Med Care*. 2004;42:367–377.
- Marshall MN, Shekelle PG, Leatherman S, et al. The public release of performance data: what do we expect to gain? A review of the evidence. *JAMA*. 2000;283:1866–1874.
- Khuri SF, Daley J, Henderson W, et al. The Department of Veterans Affairs’ NSQIP: the first national, validated, outcome-based, risk-adjusted, and peer-controlled program for the measurement and enhancement of the quality of surgical care. National VA Surgical Quality Improvement Program. *Ann Surg*. 1998;228:491–507.
- Khuri SF, Daley J, Henderson WG. The comparative assessment and improvement of quality of surgical care in the Department of Veterans Affairs. *Arch Surg*. 2002;137:20–27.
- Mackay J, Mensah G. *Atlas of Heart Disease and Stroke*. Geneva: World Health Organization; 2004.
- Health Care in America: Trends in Utilization*. National Center for Health Statistics; 2004.
- Hannan EL, Kilburn H Jr, O’Donnell JF, et al. Adult open heart surgery in New York State. An analysis of risk factors and hospital mortality rates. *JAMA*. 1990;264:2768–2774.
- Hannan EL, Kilburn H Jr, Racz M, et al. Improving the outcomes of coronary artery bypass surgery in New York State. *JAMA*. 1994;271:761–766.
- Ghali WA, Ash AS, Hall RE, et al. Statewide quality improvement initiatives and mortality after cardiac surgery. *JAMA*. 1997;277:379–382.
- Iezzoni LI. ‘Black box’ medical information systems. A technology needing assessment. *JAMA*. 1991;265:3006–3007.
- Iezzoni LI. The risks of risk adjustment. *JAMA*. 1997;278:1600–1607.
- Hosmer DW, Lemeshow S. Model-building strategies and methods for logistic regression. *Applied Logistic Regression*. 2000:91–142.
- Goldstein H, Spiegelhalter DJ. League tables and their limitations: statistical issues in comparisons of institutional performance. *J R Stat Soc*. 1996;159:385–443.
- DeLong ER, Peterson ED, DeLong DM, et al. Comparing risk-adjustment methods for provider profiling. *Stat Med*. 1997;16:2645–2664.
- Shahian DM, Normand SL, Torchiana DF, et al. Cardiac surgery report cards: comprehensive review and statistical critique. *Ann Thorac Surg*. 2001;72:2155–2168.
- Localio AR, Hamory BH, Fisher AC, et al. The public release of hospital and physician mortality data in Pennsylvania. A case study. *Med Care*. 1997;35:272–286.
- Faris PD, Ghali WA, Brant R. Bias in estimates of confidence intervals for health outcome report cards. *J Clin Epidemiol*. 2003;56:553–558.
- Hannan EL, Kilburn H Jr, Bernard H, et al. Coronary artery bypass surgery: the relationship between inhospital mortality rate and surgical volume after controlling for clinical risk factors. *Med Care*. 1991;29:1094–1107.
- Coronary Artery Bypass Surgery in New York State: 1997–1999*. New York State Department of Health; 2002.
- Littel R, Millikan GA, Stroup WW, et al. *SAS System for Mixed Models*. Cary, NC: SAS Institute; 1996.
- Snijders T, Bosker R. *Discrete Dependent Variables. Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. London: Sage Publications; 2003:207–234.
- Hosmer DW, Lemeshow S. *Applied Logistic Regression*, 2nd ed. New York: Wiley-Interscience Publication; 2000.
- Dobson AJ, Kuulasmaa K, Eberle E, et al. Confidence intervals for weighted sums of Poisson parameters. *Stat Med*. 1991;10:457–462.
- Breslow N, Day N. *Statistical Methods in Cancer Research*, vol II. Oxford University Press; 1987.
- Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. New York: Chapman & Hall; 1993.
- Hosmer DW, Lemeshow S. Confidence interval estimates of an index of quality performance based on logistic regression models. *Stat Med*. 1995;14:2161–2172.
- Carpenter J, Bithell J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Stat Med*. 2000;19:1141–1164.
- Greene WH. *Models for Panel Data. Econometric Analysis*. Upper Saddle River, NJ: Prentice Hall; 2003:283–338.
- Fleiss JL. *Reliability of Measurement. Design and Analysis of Clinical Experiments*. New York: John Wiley & Sons, Inc; 1999:1–31.
- Mukamel DB, Mushlin AI. The impact of quality report cards on choice of physicians, hospitals, and HMOs: a midcourse evaluation. *Jt Comm J Qual Improv*. 2001;27:20–27.
- Mukamel DB, Mushlin AI, Weimer D, et al. Do quality report cards play a role in HMOs’ contracting practices? Evidence from New York State. *Health Serv Res*. 2000;35:319–332.
- Mukamel DB, Weimer DL, Zwanziger J, et al. Quality report cards,

selection of cardiac surgeons, and racial disparities: a study of the publication of the New York State Cardiac Surgery Reports. *Inquiry*. 2004;41:435–446.

37. Normand SL, Glickman ME, Gatsonis CA. Statistical methods for profiling providers of medical care: issues and applications. *J Am Stat Assoc*. 1997;92:803–814.

38. Rice N, Leyland A. Multilevel models: applications to health data. *J Health Serv Res Policy*. 1996;1:154–164.

39. Moreno R, Matos R. The ‘new’ scores: what problems have been fixed, and what remain? *Curr Opin Crit Care*. 2000;6:158–165.

40. Hox J. *Multilevel Analysis: Techniques and Applications*. Mahwah, NJ: Lawrence Erlbaum Associates; 2002.

41. Krefts I, De Leeuw J. *Introducing Multilevel Modeling*. London: Sage Publications; 2000.

42. Raudenbush SW, Bryk AS. *Hierarchical Linear Models: Applications and Data Analysis Methods*, 2nd ed. Thousand Oaks, CA: Sage Publications; 2002.

43. Marshall EC, Spiegelhalter DJ. Institutional performance. In: Leyland AH, Goldstein H, eds. *Multilevel Modeling of Health Statistics*. New York: John Wiley & Sons, Ltd; 2001:127–142.

44. Dudley RA, Johansen KL, Brand R, et al. Selective referral to high-volume hospitals: estimating potentially avoidable deaths. *JAMA*. 2000; 283:1159–1166.

45. Epstein AM. Volume and outcome—it is time to move ahead. *N Engl J Med*. 2002;346:1161–1164.

**APPENDIX A. Risk-Adjustment Models for Coronary Artery Bypass Graft In-Hospital Mortality (1997–1999) Used to Create Surgeon Report Cards\***

Patient Risk Factor	Standard Logistic Model	Fixed-Effects Model	Random-Intercept Model
<b>Demographic</b>			
Age	-0.0534	-0.0533 <sup>†</sup>	-0.052 <sup>†</sup>
Age squared	0.0784	0.0805	0.0783
Female	0.7535	0.7070	0.7258
<b>Hemodynamic state</b>			
Unstable	1.2572	1.3143	1.2837
Shock	1.8868	1.9747	1.9311
Cardiopulmonary resuscitation	2.0075	2.0806	2.026
Malignant ventricular arrhythmia	0.8553	0.8682	0.8544
<b>Comorbidities</b>			
Diabetes	0.654	0.6419	0.6468
Hepatic failure	2.2032	2.1738	2.1604
Renal failure requiring dialysis	1.9646	1.9167	1.9255
Renal failure not requiring dialysis	1.1416	1.0668	1.0904
Chronic obstructive pulmonary disease	0.7256	0.8168	0.782
<b>Severity of atherosclerotic process</b>			
Previous stroke	0.7584	0.7160	0.7329
Aortoiliac disease	1.0479	1.0385	1.0379
Calcified aorta	0.8867	0.9632	0.9351
<b>Ventricular function</b>			
Ejection fraction <20	1.417	1.4064	1.4085
Ejection fraction 20–29	0.9032	0.8642	0.8781
Ejection fraction 30–39	0.7255	0.6759	0.6969
Previous myocardial infarction <6 h	1.3338	1.3548	1.3536
Previous myocardial infarction 6–23 h	0.9621	0.9377	0.9506
Previous myocardial infarction, 1–7 d	0.7512	0.7154	0.7344
Previous myocardial infarction, 8–21 d	0.5069	0.4889	0.5007
Previous open heart operations	1.2446	1.2468	1.2356
Left main coronary disease	0.5101	0.4819	0.4917
Risk squared <sup>‡</sup>	-0.0405	-0.0370	-0.0387
Intercept	-5.6066	-6.1403	-5.6783
<b>Model performance</b>			
C-statistic	0.800	0.824	0.821

\*The variance of the surgeon random effect term was 0.2067 ( $P < 0.0001$ ), yielding an intraclass correlation coefficient of 0.059.

<sup>†</sup> $P$  value  $< 0.10$ ; all other  $P$  values are  $< 0.001$ .

<sup>‡</sup>To calculate “risk squared” for a given patient, all of the risk factors (eg, female, unstable, shock) are assigned the value “1” and are then summed; the sum of these risk factors is then squared.

**APPENDIX B. Surgeon Outlier Status as a Function of Model Type and Confidence Interval (CI) Estimation Method**

Outlier Status Based on Shrinkage Estimator	Outlier Status Based on Standard Model and Poisson CI		
	High	Medium	Low
High	6	1	0
Medium	1	118	2
Low	0	0	10
Kappa = 0.88			

Outlier Status Based on Standard Model and Poisson CI	Outlier Status Based on Fixed-Effects Model and Poisson CI		
	High	Medium	Low
High	7	0	0
Medium	1	118	0
Low	0	0	12
Kappa = 0.97			

Outlier Status Based on Random-Intercept (FEC) Model and Poisson CI	Outlier Status Based on Fixed-Effects Model and Poisson CI		
	High	Medium	Low
High	8	0	0
Medium	0	118	0
Low	0	0	12
Kappa = 1.00			

Outlier Status Based on Shrinkage Estimator	Outlier Status Based on Fixed-Effects Model and Poisson CI		
	High	Medium	Low
High	7	0	0
Medium	1	118	2
Low	0	0	10
Kappa = 0.91			

Outlier Status Based on Standard Model and Poisson CI	Outlier Status Based on Standard Model and Bootstrap CI		
	High	Medium	Low
High	7	0	0
Medium	6	111	2
Low	0	1	11
Kappa = 0.77			

FEC indicates fixed-effects component.