HSR Health Services Research

© Health Research and Educational Trust
DOI: 10.1111/j.1475-6773.2010.01130.x
RESEARCH ARTICLE

Quality and Outcomes

The Hospital Compare Mortality Model and the Volume–Outcome Relationship

Jeffrey H. Silber, Paul R. Rosenbaum, Tanguy J. Brachet, Richard N. Ross, Laura J. Bressler, Orit Even-Shoshan, Scott A. Lorch, and Kevin G. Volpp

Objective. We ask whether Medicare's Hospital Compare random effects model correctly assesses acute myocardial infarction (AMI) hospital mortality rates when there is a volume–outcome relationship.
Data Sources/Study Setting. Medicare claims on 208,157 AMI patients admitted in 3,629 acute care hospitals throughout the United States.
Study Design. We compared average-adjusted mortality using logistic regression with average adjusted mortality based on the Hospital Compare random effects model. We then fit random effects models with the same patient variables as in Medicare's Hospital Compare mortality model but also included terms for hospital Medicare AMI volume and another model that additionally included other hospital characteristics.
Principal Findings. Hospital Compare's average adjusted mortality significantly underestimates average observed death rates in small volume hospitals. Placing hospital volume in the Hospital Compare model significantly improved predictions.
Conclusions. The Hospital Compare random effects model underestimates the typically poorer performance of low-volume hospitals. Placing hospital volume in the Hospital Compare model, and possibly other important hospital characteristics, appears indicated when using a random effects model to predict outcomes. Care must be taken to insure the proper method of reporting such models, especially if hospital characteristics are included in the random effects model.

Key Words. Hospital Compare, mortality, acute myocardial infarction, random effects model

Medicare's web-based "Hospital Compare" is intended to provide the public "with information on how well the hospitals in your area care for all their adult patients with certain medical conditions" (U.S. Department of Health & Human Services 2007a). For acute myocardial infarction (AMI) mortality in

2007, of 4,477 U.S. hospitals (many of which presumably have no experience with AMI), the Medicare Hospital Compare model asserted that 4,453 (99.5 percent) were "no different than the U.S. national rate," only 17 hospitals were "better," and 7 were "worse" than the U.S. national rate. In 2008, the Hospital Compare model suggests that of 4,311 hospitals, none were worse than average, and nine were better than average.

These evaluations are surprising. Some hospitals treat only a few AMIs every year, and others treat a few each week. One of the more consistent findings in the medical literature (Luft, Hunt, and Maerki 1987; Halm, Lee, and Chassin 2002; Gandjour, Bannenberg, and Lauterbach 2003; Shahian and Normand 2003) is that, after adjusting for patient risk factors, there is often a higher risk of death when a patient is treated at a low-volume hospital. Indeed, this pattern is unmistakable in Medicare data, the data used to construct Hospital Compare. However, Hospital Compare reports no such pattern.

THE SMALL NUMBERS PROBLEM

Over 20 years ago, in a paper published in this journal, Chassin et al. (1989) described the small numbers problem when attempting to rank hospitals by adjusted mortality rates. To account for death rate instability with low volume, the Chassin study chose to rank hospitals by the statistical significance associated with their observed and expected death rates. This, in retrospect, was

Address correspondence to Jeffrey H. Silber, M.D., Ph.D., Center for Outcomes Research, 3535 Market Street, Suite 1029, Philadelphia, PA 19104; e-mail: silberj@wharton.upenn.edu. Tanguy J. Brachet, Ph.D., Richard N. Ross, M.S., Laura J. Bressler, B.A., Orit Even-Shoshan, M.S., and Scott A. Lorch, M.D., M.S.C.E., are with the Center for Outcomes Research, The Children's Hospital of Philadelphia, Philadelphia, PA. Jeffrey H. Silber, M.D., Ph.D., and Tanguy J. Brachet, Ph.D., are with the Department of Anesthesiology & Critical Care, University of Pennsylvania School of Medicine, Philadelphia, PA. Jeffrey H. Silber, M.D., Ph.D., is with the Department of Pediatrics, University of Pennsylvania School of Medicine, Philadelphia, PA. Jeffrey H. Silber, M.D., Ph.D., and Kevin G. Volpp, M.D., Ph.D., are with the Department of Health Care Management, The Wharton School, University of Pennsylvania, Philadelphia, PA. Jeffrey H. Silber, M.D., Ph.D., Paul R. Rosenbaum, Ph.D., Tanguy J. Brachet, Ph.D., Orit Even-Shoshan, M.S., Scott A. Lorch, M.D., M.S.C.E., and Kevin G. Volpp, M.D., Ph.D., are with the Leonard Davis Institute of Health Economics, University of Pennsylvania, Philadelphia, PA. Paul R. Rosenbaum, Ph.D., is with the Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA. Scott A. Lorch, M.D., M.S.C.E., is with the Department of Pediatrics, Division of Neonatology, The Children's Hospital of Philadelphia, Philadelphia, PA. Kevin G. Volpp, M.D., Ph.D., is with the Center for Health Equity Research and Promotion, Veteran's Administration Hospital, Philadelphia, PA. Kevin G. Volpp, M.D., Ph.D., is with the Department of Medicine, University of Pennsylvania School of Medicine, Philadelphia, PA.

not a good decision. Large hospitals could have extreme ranks because their size allowed for statistical significance, whereas small hospitals could not reach significance and would be forced to be ranked near the middle.

Twenty years later, a new solution for the small numbers problem in AMI has been introduced by Medicare's Hospital Compare. The Hospital Compare model for AMI is based on a random effects model published by Krumholz et al. (2006b) as well as a technical report funded by a contract from Medicare (Krumholz et al. 2006a). Consistent with the technical report, in a section titled *Adjusting for Small Hospitals or a Small Number of Cases* (U.S. Department of Health & Human Services 2007b), the Hospital Compare web page says:

> The [Medicare] hierarchical regression model also adjusts mortality rates results for . . . hospitals with few heart attack . . . cases in a given year . . . . This reduces the chance that such hospitals' performance will fluctuate wildly from year to year or that they will be wrongly classified as either a worse or better performer . . . .In essence, the predicted mortality rate for a hospital with a small number of cases is moved toward the overall U.S. National mortality rate for all hospitals. The estimates of mortality for hospitals with few patients will rely considerably on the pooled data for all hospitals, making it less likely that small hospitals will fall into either of the outlier categories. This pooling affords a "borrowing of statistical strength" that provides more confidence in the results.

After "moving" ("shrinking") many hospitals' AMI mortality rates toward the national rate, the 2008 Hospital Compare concludes that 4,302/4,311 or 99.8 percent of hospitals are "no different than U.S. national rate" and zero hospitals are "worse than U.S. national rate." This study will attempt to explain why Hospital Compare came to these conclusions.

The "hierarchical" or "random effects" model used to construct Hospital Compare utilizes the fact that low volume at a hospital implies that its empirical mortality rate is imprecisely estimated. However, it assumes that there is no relationship between volume and mortality (Panageas et al. 2007), not a reasonable assumption, given that the literature suggests AMI mortality rates tend to be higher when volume is lower (Luft et al. 1987; Farley and Ozminkowski 1992; Thiemann et al. 1999; Tu, Austin, and Chan 2001; Halm et al. 2002). Hospital Compare could have developed a random effects model that allowed the empirical data to speak to the issue of whether a systematic volume–outcome relationship is present, but the Hospital Compare model used an overriding assumption that true hospital mortality rates are random and independent of volume. For this reason, the mortality rate at a low-volume hospital is judged by Hospital Compare to be unstable and in need of

"shrinkage" toward the national average (rather than toward the average of hospitals of similar volume).

In this paper, we first review the mechanics of the Hospital Compare random effects model and explain how Hospital Compare arrives at an adjusted hospital death rate. We then look for an empirical relationship between volume and outcome in the AMI Medicare data and find one very similar to that reported in the literature. Next, we compare the results of the Hospital Compare-adjusted death rate model with the results of a standard logistic regression model by grouping all hospitals into quintiles of size so that there is no "small numbers problem" and ask whether Hospital Compare is correctly estimating average mortality inside these large quintile groupings of hospitals. We also present a modification of the present Hospital Compare model that includes Medicare AMI volume and other hospital characteristics.

METHODS

The Hospital Compare Random Effects Model

To motivate why we are interested in the Hospital Compare model when assessing a condition or procedure for which there is a demonstrated volume–outcome relationship, we must first explain how the Medicare Hospital Compare random effects model works. To begin with, it is essential to note that the model does not evaluate hospitals based on the typical observed (O) versus expected (E) or O/E ratio generally utilized by the health services community (Iezzoni 2003). In the typical O/E model, the observed death rate is simply the raw death rate at a hospital, and the expected death rate is based on a model with only patient characteristics, not hospital characteristics. Instead, Hospital Compare produces a hospital-adjusted death rate by computing a predicted death rate (P) based on a random effects model with patient characteristics and the specific hospital's outcomes, and an expected death rate (E) based only on patient characteristics at the hospital of interest. The resultant P/E ratio of mortality is then multiplied by the national average death rate to report the adjusted death rate. In standard practice, a random effects model aims to predict correctly, and including hospital characteristics such as volume is consistent with that goal (though volume is not included in the Medicare Hospital Compare model). In sharp contrast, however, the expected term never includes hospital variables or characteristics to estimate this term.

Random effects models produce "shrinkage." If a hospital has a substantially higher (or lower) death rate than average, the hospital will be

predicted to have substantially elevated (or decreased) risk only if the sample size is adequate to substantiate the elevated (or decreased) rate; otherwise, if the sample size is small, the predicted death rate is "shrunken" or moved toward the model's sense of what is typical. In the Hospital Compare model, which does not include attributes of hospitals, each hospital is shrunken toward the mean of all hospitals.

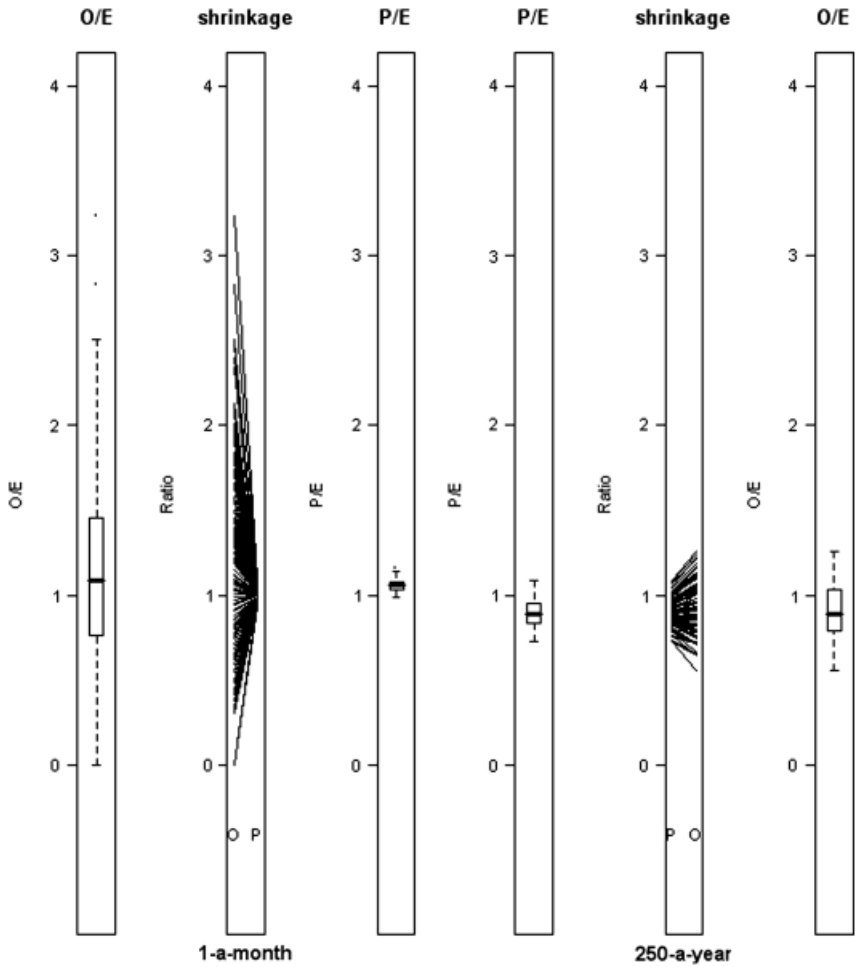
Figure 1 allows us to better understand how shrinkage works in the Medicare random effects model. It compares the 166 small hospitals with volume between 11 and 13 patients a year (1-a-month) to the 86 large hospitals with at least 250 patients a year (on the order of 1-a-day). One can observe that for small hospitals, O/E is very unstable, but the median O/E for the small hospitals is >1. The P/E for these same small hospitals, what Medicare actually reports to the public, is far less variable but now has a median near 1. For the higher volume hospitals, with 250 cases per year, we observe far less shrinkage, and the O/E and P/E ratios both appear below 1. A slightly more technical description of the random effects model used by Hospital Compare is described by equation (1) of the Appendix SA2.

In thinking about shrinkage, the notation of Morris (1983) and more recently Dimick et al. (2009) and Mukamel et al. (2010) is helpful. One can describe the extent of shrinkage by writing $P = \lambda O + (1 - \lambda)E$ where for each hospital, we solve for λ after being given P, O, and E. This is a descriptive tool and not the mathematical formula used to create the shrinkage (see Gelman et al. 1997, chapter 14 for details). If we divide both sides of the equation by E, we get $P/E = \lambda\ O/E + (1 - \lambda)E/E$. To express adjusted rates as calculated in Hospital Compare random effects model, one need to only multiply both sides of the equation by the national AMI death rate (16.6 percent mortality). We graph λ versus AMI volume in Figure 2. When volume is low, the Hospital Compare model places almost no emphasis on O/E (λ is near 0), and almost all emphasis on $E/E = 1$, making the prediction offered to the public (the Hospital Compare-adjusted rate) almost identical to the national rate, regardless of what is observed.

Data

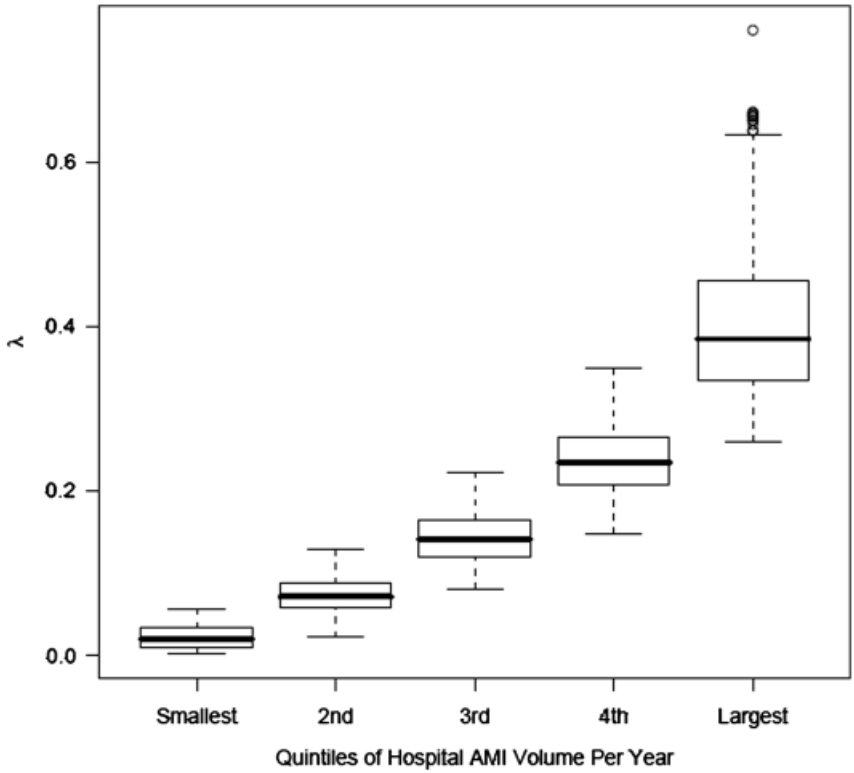
Using the Medicare Provider Analysis and Review File from July 1, 2004 to June 30, 2005, we selected all cases of AMI based on ICD9 codes starting with "410" excluding those with fifth digit = "2" for "subsequent care" (as per Krumholz et al. 2006a, b). We included only hospitals we knew did not open or close in the study year. Starting with 377,515 AMI admissions, we selected

Figure 1: Example of How Hospital Compare Shrinks Predictions Based on Acute Myocardial Infarction Volume



Notes. This figure compares the 166 hospitals with between 11 and 13 patients a year (1-a-month) and the 86 hospitals with at least 250 patients a year (on the order of 1-a-day). The three pictures on the left are for 1-a-month. The three pictures on the right are for 250-a-year. *O/E* is on the outside. *P/E* is on the inside. *O/E* and *P/E* appear worse for 1-a-month than for 250-a-year. Medicare reports to the public the shrunken *P/E* rate times the national rate, not the *O/E* rate. For the higher volume hospitals, with 250 cases per year, we observe far less shrinkage, and the *O/E* and *P/E* ratios both appear below 1. For the 1-a-month hospitals, we observe a median $O/E > 1$ but a *P/E* very near to 1.

Figure 2: How Much Does the Hospital Compare Model Emphasize an Individual Hospital's Observed Mortality Rate When Calculating the Predicted Mortality That Is Provided to the Public?



Notes. The figure provides the value of λ where $P/E = \lambda\,(O/E) + (1-\lambda)(E/E)$. Because Hospital Compare does not include volume or any hospital characteristics in the model, the $(1-\lambda)$ term is multiplied by 1, which represents the national average or typical hospital death rate. As hospital acute myocardial infarction (AMI) volume increases, λ increases, suggesting an increasing emphasis on the hospitals own O/E. As hospital AMI volume decreases, λ decreases, and the Hospital Compare model emphasizes the national mortality rate (E/E) rather than the hospitals own observed mortality rate (O/E) when making the prediction P/E.

the first AMI admission for a patient between ages 66 and 100 years at acute care hospitals (262,578 patients at 3,694 acute care hospitals). We excluded 30,984 patients transferred in from another acute care hospital, and 21,183 patients admitted for "AMI" but discharged alive within 2 days; 43 patients were deleted because of date of death mistakes, and finally we excluded 2,211

Table 1: Relationship between the Hospital Volume Quintile and the Odds of Mortality

		AMI Mortality in the United States	
Hospital Volume by Quintile	*Number of Hospitals*	*Logit Model*	*Random Effects Model*
Quintile 1 (volume $\leq$ 8 per year)	734	1.54 (1.41, 1.68)	1.53 (1.40,1.67)
Quintile 2 (volume 9 – 22 per year)	753	1.29 (1.23, 1.36)	1.29 (1.22 1.36)
Quintile 3 (volume 23 – 47 per year)	696	1.13 (1.10, 1.18)	1.14 (1.09, 1.18)
Quintile 4 (volume 48 – 95 per year)	732	1.08 (1.05, 1.11)	1.07 (1.04, 1.11)
Quintile 5 (volume 96 – 735 per year)	714	1 Reference	1 Reference

Notes. In this table, we present results for both a standard logistic model and the random effects model. For all models, hospitals in the lowest volume quintile are associated with the highest odds of mortality. Each cell provides the odds ratio adjusted for patient characteristics, and the 95% confidence interval.

All adjusted for same variables as included in the Hospital Compare model (Krumholz et al. 2006b) (see Table 1 of Appendix SA2 for full model).

patients enrolled in managed care plans due to incomplete data. This left 208,157 unique patients admitted with AMI in 3,629 hospitals.

Statistical Analysis

The Risk-Adjustment Model. We duplicated as best we could (based on the available definitions used in the Hospital Compare model) the risk adjustment model of Krumholz et al. (2006b), which is the model cited by Medicare in their website. When we implemented the Medicare random effects model, we used the *SAS* program PROC GLIMMIX (SAS Institute Inc. 2008), as suggested by Krumholz et al. (2006a) in the technical report associated with Medicare Hospital Compare. When we fit the logit model without random effects, we used the same patient parameters as in the Krumholz model, but used PROC LOGISTIC from *SAS* (SAS Institute Inc.). Implicitly, the *GLIMMIX* model takes into account hospital-level clustering and the logistic model does not, although pertinent results were nearly identical (see Table 1); see Freedman (2006) and Gelman (2006) for general discussion clustering in logistic and related models.

The random effects Model. The random effects model has patient characteristics and hospital indicators based on the model developed by Krumholz et al. (2006a, b), using the Hierachical Condition Categories (Centers for Medicare and Medicaid Services). Our model, like that of Krumholz et al. (2006a),

includes 16 comorbidity groups, 10 cardiovascular variables, as well as age and sex as used in the Medicare mortality model. In the Medicare random effects model, hospital volume is viewed as fixed, and hospital quality parameters (random effects) are assumed to be sampled from a single Normal distribution independently of hospital volume (Gelman 2006; SAS Institute Inc. 2008).

RESULTS

Hospital Volume and AMI Mortality in the United States

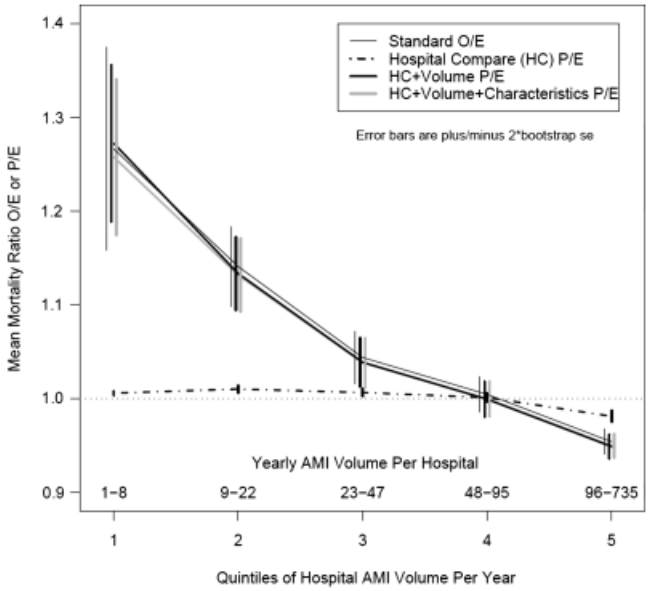
Two columns of Table 1 display the relationship between hospital AMI volume and AMI mortality in the United States, adjusting for patient risk factors using variables from the Medicare Hospital Compare model. The odds ratios compare mortality at hospitals in five quintiles of size to mortality at the largest hospitals. These models are based on 208,157 Medicare patients admitted for AMI to 3,629 hospitals. The model includes 24 variables describing patient characteristics as developed by Krumholz et al. (2006a), which is the basis of the CMS Hospital Compare model for AMI. The *C*-statistic for the logit model without volume was 0.73 (almost identical to the Krumholz random effects model *C*-statistic of 0.71). Details of the risk-adjustment model are provided in the Table 1 of Appendix SA2.

Hospitals in the lowest volume quintile had the highest risk-adjusted mortality. The volume–outcome relationship is substantial in magnitude, highly significant, with narrow confidence intervals (CIs). Together, the four quintile variables were significant using the likelihood ratio test ($p < .0001$). The logit model estimates hospitals at the lowest 20th percentile of AMI volume to have an odds of death that is 1.54 times higher than hospitals at the highest 20th percentile (95 percent CI: 1.41, 1.68). The random effects model similarly finds an unambiguously strong volume–outcome relationship. This volume–outcome relationship is descriptive and predictive but not explanatory: it is unambiguous that risk-adjusted mortality is higher at lower volume hospitals; however, why this is so and what it might mean for public policy are not immediate consequences of the existence of a volume–outcome relationship.

Contrasting Hospital Compare's Predictions to AMI Mortality in the United States

Figure 3 compares the results of the Hospital Compare average-adjusted death rates to the actual observed average hospital death rates for each of the five

Figure 3: Comparison of Standard Logistic Regression Observed/Expected Mortality Ratios versus Three Versions of the Hospital Compare Random Effects Model Predicted/Expected Mortality Ratios



Notes. Both *O/E* and *P/E* would be multiplied by the national acute myocardial infarction mortality rate to get an adjusted rate which could be reported to the public. Note that for the lower two quintiles of hospitals by volume (the smallest 40 percent of hospitals), there is a great discrepancy between the average *O/E* mortality rate ratios based on the standard *O/E* logit model (thin black line) and the average *P/E* mortality rate ratios based on the random effects model used by Medicare's Hospital Compare (the dashed line). However, when size (the thick black line) or size and other hospital characteristics (the thick gray line) are added to the present Hospital Compare model, the *P/E* values become almost identical to the standard *O/E* results.

hospital quintiles sorted by volume. The *x*-axis displays the volume quintile, lowest to highest. The *y*-axis depicts the adjusted death rates (without multiplying by the average national death rate—a constant term). The *y*-axis graphs four rates, three *P/E* rates, and the *O/E* rate. The *P/E* rate depicted by the dashed line (at the bottom of the graph) represents the average predictions based on the Hospital Compare model (without volume) that shrinks results of individual hospitals according to the random effects model for all hospitals in the quintile, divided by the expected rate based only on patient characteristics. The standard *O/E* rate (depicted by the thin black line) represents the actual

average observed rate of death at each individual hospital in the quintile divided by the expected rate based on the same model used by Hospital Compare when only patient characteristics are in the model. Differences between Hospital Compare's *P/E* and the *O/E* rates are considerable. The average of the *P/E* rates by Hospital Compare suggest little deviation from expected (except for a very slight improvement below expected in the largest quintile). However, the average *O/E* rates show very substantial differences between observed and expected, with the lowest quintile displaying approximately a 25 percent increase in mortality from expected and the largest quintile displaying approximately a 5 percent reduction in observed versus expected.

Taken together, there is quite a bit of evidence about how the smaller hospitals perform as a group—their risk-adjusted mortality is above average. The Hospital Compare model does not look at the small hospitals as a group, but rather views them individually, each with unstable mortality rates which are moved one at a time toward the national rate, ignoring the fact that as a group, the small hospitals have a mortality rate above the national average. That is, the Hospital Compare model sets *P/E* near 1 for hospitals with small volume, then "discovers" that *P/E* is near 1. For large hospitals, Hospital Compare does less shrinkage but does nonetheless predict worse outcomes than actually observed for these largest hospitals. For shrinkage to improve estimation, it is important to shrink unstable estimates toward a reasonable model, and the model which says that volume is unrelated to risk is not a reasonable model.

The metaphorical phrase "borrow strength" is widely used to describe and motivate the mathematics of shrinkage, but one is not borrowing strength if one is shrinking toward an unreasonable model. To say this is not to criticize shrinkage but rather to criticize its use with an unreasonable model. See Cox and Wong (2010) for related discussion.

Adding Hospital Characteristics into the Hospital Compare Model

We next studied whether predictions could be improved if we added the AMI volume quintile to the model. This would allow shrinkage to a hospital's own volume quintile rather than to the nation as a whole. This is displayed as the thick black curve of Figure 3. If we were to add in hospital volume and other hospital characteristics (resident-to-bed ratio, nurse-to-bed ratio, nurse mix, and technology status) as well as AMI volume, we obtain the thick gray curve in Figure 3. Comparing the *O/E* curve (thin black) to both the *P/E* curves that

include volume or volume plus hospital characteristics, we see virtual overlap, in distinction to the dashed curve representing the present Hospital Compare model without volume or hospital characteristics. Adding volume or volume and other hospital characteristics improves the Hospital Compare predictions.

Understanding the Hospital Compare Model: Examining Individual Hospitals

To better understand the present Hospital Compare random effects model, and what would happen when volume and possibly also other important hospital variables are added to the random effects model, we present data from individual hospitals in Table 2. To help illustrate that a random effects model of whatever kind evaluates small hospitals together as a group and not individually, it is informative to study five quantities (λ, O/E, $1-\lambda$, F/E, and P/E) which indicate the degree to which an individual hospital is evaluated by the group of hospitals to which it belongs. In random effects models with hospital and patient characteristics, define F to be the mortality rate forecasted for that hospital using the fixed effects that describe attributes of its patients (as in E) and also its hospital characteristics such as volume. We can then write $P=\lambda O+(1-\lambda)F$, and dividing by E, we get the more familiar $P/E=\lambda(O/E)+(1-\lambda)F/E$. When no hospital characteristics are included in the random effects model, then $F=E$, and the equation describing P/E becomes identical to that introduced earlier (see Figure 2), where P is a linear combination of the observed rate and the national rate. Again, λ describes the extent of shrinkage but is not a part of the mathematics of shrinkage.

When λ is near 0, this reflects the fact that the observed performance at the hospital, namely O/E, is given little emphasis in the random effects model prediction, and the predominant emphasis is on the group to which the hospital belongs, namely the hospital characteristics described by F/E. Table 2 illustrates results from three models: the original Hospital Compare model (Model 1) and a model that also adds in volume (Model 2) and volume plus hospital characteristics (Model 3). As hospital volume increases, more emphasis is placed on the hospitals own O/E and less on the hospitals own characteristics (F/E). However, even with very high volume of about one case-a-day (see hospitals G and H), the random effects models still place only about 60 percent the emphasis on O/E, while F/E still contributes 40 percent to the P/E estimates. For hospitals with low volumes (A, B, C, D, E, and F), the value of λ is quite low, reflecting the fact that all the random effects models displayed in Table 2 place most emphasis on F/E (the hospitals characteristics adjusting

Table 2: Understanding the Hospital Compare Random Effects Model—Some Examples

			Model 1: Model Based on Hospital Compare					*Model 2: Hospital Compare Model Adding in AMI Volume*					*Model 3: Hospital Compare Model Adding in Volume and Hospital Characteristics*				
Hosp.	*Vol.*	*O*	λ	*O/E*	$1-\lambda$	*F/E*	*P/E*	λ	*O/E*	$1-\lambda$	*F/E*	*P/E*	λ	*O/E*	$1-\lambda$	*F/E*	*P/E*
A	1	0	0.004	0	0.996	1	0.996	0.005	0	0.995	1.331	1.277	0.004	0	0.996	1.282	1.277
B	1	1	0.008	2.775	0.992	1	1.014	0.007	2.775	0.993	1.226	1.237	0.006	2.775	0.994	1.221	1.231
C	8	0	0.043	0	0.957	1	0.957	0.042	0	0.958	1.295	1.241	0.039	0	0.961	1.302	1.251
D	8	0.750	0.044	3.253	0.956	1	1.099	0.041	3.253	0.959	1.260	1.342	0.025	3.253	0.975	1.175	1.251
E	24	0.042	0.106	0.209	0.894	1	0.916	0.092	0.209	0.908	1.036	0.961	0.087	0.209	0.913	1.070	0.995
F	24	0.417	0.127	1.631	0.873	1	1.080	0.108	1.631	0.892	1.034	1.098	0.103	1.631	0.897	1.083	1.140
G	362	0.127	0.580	0.866	0.420	1	0.922	0.525	0.866	0.475	0.947	0.904	0.501	0.866	0.499	0.919	0.893
H	362	0.157	0.612	0.929	0.388	1	0.957	0.558	0.929	0.442	0.947	0.937	0.534	0.929	0.466	0.914	0.922

Notes. P/E, O/E, F/E are quantities obtained from the random effects model. We can describe P/E as a linear combination of O/E and F/E, where $P/E = \lambda(O/E) + (1-\lambda)F/E$. Model 1 results are based on the present Hospital Compare model without hospital characteristics. Model 2 adds in volume and Model 3 adds in both volume and five other hospital characteristics. We have chosen eight hospitals to illustrate the results: each pair of hospitals has the same reported acute myocardial infarction (AMI) volume. For Model 1, $F/E = E/E = 1$ because there are no hospital characteristics used to estimate F. For Models 2 and 3, F/E is generally different from 1, as a hospital may have better or worse characteristics compared with the typical hospital. For the small hospitals, there are great differences between the P/E estimates for Model 1 versus Models 2 and 3, because Model 1 is shrinking the prediction to the national rate ($F/E = 1$), while Models 2 and 3 shrink toward F/E, which is higher because smaller hospitals as a group perform worse in the data set. As hospital volume increases, the models place increasing emphasis on O/E, with λ values over 0.50 for hospitals G and H.

E, expected; *F*, forecasted rate (estimated rate based on fixed effects parameters using hospital and patient characteristics); Hosp., hospital name; *O*, observed rate; *P*, predicted; vol., volume of Medicare AMI patients.

for patient characteristics) rather than O/E (the observed outcomes adjusting for patient characteristics).

A very clear message emerges from Table 2. In the present Hospital Compare model, predictions for small volume hospitals are shrunk back to the national average (P/E terms are near 1). However, when viewing models with volume added to the Hospital Compare model, we see that low-volume hospitals are now predicted to have higher death rates than average (P/E terms are above 1), because the predictions are shrunken back to the average death rates of the small hospitals. Because the volumes are low, there may be little evidence to suggest that the individual small hospitals are worse or better than their size cohort (in other words λ is low so P/E will resemble F/E, not O/E). If a hospital treated one AMI patient during a year, its O is zero or one and its O/E does virtually nothing to characterize its individual performance; moreover, no mathematical model can change that. When volumes are low, and the random effects models will be shrinking P/E toward F/E, it makes better sense to shrink these low-volume hospitals to their respective groups, rather than the national average, knowing that on average the low-volume hospitals have higher mortality.

Discussion

In 2007, Medicare introduced a web-based model to aid the public in selecting hospitals (Francis 2007). While motivated by legitimate concerns about the instability of estimates of quality in small hospitals, the random effects model used by Medicare assumes hospital quality is a random variable unrelated to hospital volume, in contrast to a substantial empirical literature that shows strong relationships between hospital volume and AMI mortality.

The assumption in the CMS Hospital Compare random effects model developed by Krumholz et al. (2006a, b) is that there is no association between the volume of the hospital and hospital characteristics associated with quality of care. By using such a random effects model to evaluate hospital quality, Medicare has, in effect, assumed away any volume–outcome effect and reassigned (recalculated or shrunken) adjusted mortality rates in low-volume hospitals thereby reducing any volume–outcome association that may be present. This recalculation is inherent in the decision to exclude volume and other hospital characteristics from the shrinkage model. A better model (see Figure 3) would allow the data to speak to the issue of the association between AMI risk and hospital attributes, such as hospital volume and other facility

characteristics, rather than assuming there is no such association. If one estimates using shrinkage, one must shrink toward a reasonable model.

As a group, small hospitals are performing below average, yet in the Hospital Compare model, one by one, these small hospital outcomes are shrunken to the expected mortality rate of the entire population and are reported to be no different from average. See Tables 3 and 4 of the Appendix SA2 for simulation studies displaying this finding.

In a recent report, Dimick et al. (2009) display the usefulness of using both mortality and volume in predictive models in surgery, and Li et al. (2009) suggest that random effects models used to compare nursing home quality appear to underestimate the poor performance of smaller nursing homes. Mukamel et al. (2010) have also shown that the random effects model can incorrectly shrink to the grand mean when systematic differences occur in the population. The Leapfrog Group has also recently implemented a random effects model with volume. For some related theory, see Cox and Wong (2010).

One may argue that hospital volume or even other hospital characteristics relevant for AMI be included in the Hospital Compare random effects model. Normand, Glickman, and Gatsonis (1997) have contended that when there are systematic differences across providers, "... more accurate estimates of provider specific adjusted outcomes will be obtained by inclusion of relevant provider characteristics." In the case of the Hospital Compare random effects model, these provider characteristics should include hospital volume and possibly other factors associated with better outcomes, such as nurse-to-bed ratio, nurse mix, technology status, and resident-to-bed ratio (Silber et al. 2007; Silber et al. 2009).

We included volume in the model through the use of indicators for hospital volume. This is a simple approach and it avoids difficulties that might arise in which the shape of the volume–outcome relationship is mostly estimated from the largest hospitals. One can certainly include hospital volume as a continuous variable in the model, but one must model the relationship correctly. In particular, in looking at the matter, we found the relationship is not linear on the logit scale. One could also use various forms of local regression (see Ruppert, Wand, and Carroll 2003).

The Hospital Compare website is intended to guide patients to better hospitals. For that task, it is reasonable, perhaps imperative, to use patterns such as the volume–outcome relationship, which are only visible when data from many hospitals are combined. That task is concerned with providing good advice to most patients. For an individual small hospital, no model can

determine with precision what the mortality rate may be. However, we can say with great certainty that the typical small hospital performs worse than the typical large hospital. The Hospital Compare website would do better to provide the public with this information, rather than suggest that any individual small hospital is no different from the mean of all hospitals. To assert that each individual small hospital is average because we lack sufficient evidence to reject the null hypothesis that this individual hospital is average is to make the familiar though serious error of asserting the null hypothesis is true because one lacks sufficient evidence to show that it is false. Given small numbers, Hospital Compare can say that there is too little data available to suggest that the individual small hospital is different from its peer group of other small hospitals, while stressing that Hospital Compare knows for certain only that small hospitals have higher death rates as a group than large hospitals.

Guiding patients to capable hospitals is one task, but evaluating the performance of a particular hospital administrator at a particular small hospital is another very different task. In one single small hospital, the observed mortality rate is, again, too unstable to be used to evaluate the unique performance of that hospital, and a random effects model will of necessity lump a small hospital with other small hospitals having similar attributes (if the random effects model includes volume). Therefore, a random effects model of whatever kind could not recognize a unique superb hospital administrator who had overcome all the problems presented by low volume to produce care of superior quality. For this reason, while a random effects model may provide patients with guidance when selecting among potential hospitals, a random effects model of whatever kind should not be used to assign grades to the performance of individual providers when their volume is low, as there may be excellent small hospitals with some characteristics that typically predict poor outcomes.

When a reasonable model is used to shrink mortality rates, those rates may more accurately describe the population of hospitals as a whole, but there will always be uncertainty about whether the shrinkage has improved the estimate for an individual hospital with low volume. If the shrunken rates are used to guide policy for the population as a whole, then individual small hospitals have no basis for complaint; however, if the shrunken rates are misused to evaluate individual small hospitals based largely on the performance of many other small hospitals, then complaint is justified. Shrinkage may be useful in performing one task and useless or harmful in performing a very different task.

Care is needed in deciding which hospital attributes to use in a random effects model that shrinks predictions toward the model. One concern here is

with gaming, that is, with manipulating one's attributes, so as to be shrunken toward a better prediction, without actually improving the quality of care. Gaming is likely to occur if random effects models are misused to evaluate and reward or punish the performance of individual hospitals. Just as gaming through upcoding (Green and Wintfeld 1995) and selection (Werner and Asch 2005) can occur with patient characteristics, so too may there be gaming of hospital characteristics. Presumably, it may not be too difficult to uncover those hospitals gaming accounting practices through mergers to manipulate volume. However, a more difficult problem may occur when a random effects model includes an indicator of a potentially effective technology at the hospital, and it may be possible to acquire the technology without putting it to effective use. In this case, the quality of care at that hospital would not improve, but its prediction would improve merely because other hospitals use the same technology effectively.

In summary, there is a considerable literature on the volume–outcome relationship, which consistently shows lower AMI mortality risk at higher volume hospitals, a relationship that exists within the data used by Medicare for the Hospital Compare model. However, the Hospital Compare random effects model uses low volume to "shrink" individual small hospitals, one-by-one, back to the overall national mean; hence, the model underestimates the degree of poor performance in low-volume hospitals. If shrinkage is used to guide patients toward hospitals with superior outcomes, it is important to respect and preserve patterns in national data, such as the volume–outcome relationship, and not allow the shrinkage to remove those patterns.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: This work was supported by NHLBI grants R01 HL082637 and R01 HL094593 and NSF grant 0849370. Portions of this work were presented at the 2009 AcademyHealth Annual Meeting. We thank Traci Frank, A.A., for her assistance with preparing this manuscript.

Disclosures: None.

Disclaimers: None.

REFERENCES

Centers for Medicare and Medicaid Services. "Medicare Advantage—Rates & Statistics. Risk Adjustment. 2007 Cms-Hcc Model Software (Zip, 61 Kb)—Updated

08/15/2007" [accessed on October 10, 2007]. Available at http://www.cms.hhs.gov/MedicareAdvtgSpecRateStats/06_Risk_adjustment.asp

Chassin, M. R., R. E. Park, K. N. Lohr, J. Keesey, and R. H. Brook. 1989. "Differences among Hospitals in Medicare Patient Mortality." *Health Services Research* 24: 1–31.

Cox, D. R., and M. Y. Wong. 2010. "A Note on the Sensitivity to Assumptions of a Generalized Linear Mixed Model." *Biometrika* 97: 209–14.

Dimick, J. B., D. O. Staiger, O. Baser, and J. D. Birkmeyer. 2009. "Composite Measures for Predicting Surgical Mortality in the Hospital." *Health Affairs* 28: 1189–98.

Farley, D. E., and R. J. Ozminkowski. 1992. "Volume–Outcome Relationships and In-Hospital Mortality: The Effect of Changes in Volume over Time." *Medical Care* 30: 77–94.

Francis, T. 2007. "How to Size Up Your Hospital: Improved Public Databases Let People Compare Practices and Outcomes; The Importance of Looking Past the Numbers." *The Wall Street Journal*, July 10, New York, pp. D.1.

Freedman, D. A. 2006. "On the So-Called 'Huber Sandwich Estimator' and 'Robust Standard Errors.'" *American Statistician* 60: 299–302.

Gandjour, A., A. Bannenberg, and K. W. Lauterbach. 2003. "Threshold Volumes Associated with Higher Survival in Health Care: A Systematic Review." *Medical Care* 41: 1129–41.

Gelman, A. 2006. "Multilevel (Hierarchical) Modeling: What It Can and Cannot Do." *Technometrics* 48: 432–5.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 1997. *Bayesian Data Analysis*. New York: Chapman & Hall.

Green, J., and N. Wintfeld. 1995. "Report Cards on Cardiac Surgeons—Assessing New York State's Approach." *New England Journal of Medicine* 332: 1229–32.

Halm, E. A., C. Lee, and M. R. Chassin. 2002. "Is Volume Related to Outcome in Health Care? A Systematic Review and Methodologic Critique of the Literature." *Annals of Internal Medicine* 137: 511–20.

Iezzoni, L. I. 2003. *Risk Adjustment for Measuring Health Care Outcomes*. Chicago, IL: Health Administration Press.

Krumholz, H. M., S.-L. T. Normand, D. H. Galusha, J. A. Mattera, A. S. Rich, Y. Wang, and M. M. Ward. 2006a. *Risk-Adjustment Models for Ami and Hf 30-Day Mortality, Subcontract #8908-03-02*. Baltimore, MD: Centers for Medicare and Medicaid Services.

Krumholz, H. M., Y. Wang, J. A. Mattera, Y. Wang, L. F. Han, M. J. Ingber, S. Roman, and S.-L. T. Normand. 2006b. "An Administrative Claims Model Suitable for Profiling Hospital Performance Based on 30-Day Mortality Rates among Patients with an Acute Myocardial Infarction." *Circulation* 113: 1683–92.

Li, Y., X. Cai, L. G. Glance, W. D. Spector, and D. B. Mukamel. 2009. "National Release of the Nursing Home Quality Report Cards: Implications of Statistical Methodology for Risk Adjustment." *Health Services Research* 44: 79–102.

Luft, H. S., S. S. Hunt, and S. C. Maerki. 1987. "The Volume–Outcome Relationship: Practice-Makes-Perfect or Selective-Referral Patterns?" *Health Services Research* 22: 157–82.

Morris, C. N. 1983. "Parametric Empirical Bayes Inference: Theory and Applications." *Journal of the American Statistical Association* 78: 47–55.

Mukamel, D. B., L. G. Glance, A. W. Dick, and T. M. Osler. 2010. "Measuring Quality for Public Reporting of Health Provider Quality: Making It Meaningful to Patients." *American Journal of Public Health* 100: 264–9.

Normand, S.-L. T., M. E. Glickman, and C. A. Gatsonis. 1997. "Statistical Methods for Profiling Providers of Medical Care: Issues and Applications." *Journal of the American Statistical Association* 92: 803–14.

Panageas, K. S., D. Schrag, A. Russell Localio, E. S. Venkatraman, and C. B. Begg. 2007. "Properties of Analysis Methods That Account for Clustering in Volume–Outcome Studies When the Primary Predictor Is Cluster Size." *Statistics in Medicine* 26: 2017–35.

Ruppert, D., M. P. Wand, and R. J. Carroll. 2003. *Semiparametric Regression.* New York: Cambridge University Press.

SAS Institute Inc. 2004. "SAS/STAT 9.1 User's Guide. Chapter 42. The Logistic Procedure" [accessed on February 29, 2008]. Available at http://support.sas.com/documentation/onlinedoc/91pdf/sasdoc_91/stat_ug_7313.pdf

SAS Institute Inc. 2008. "Production Glimmix Procedure" [accessed on January 28, 2008]. Available at http://support.sas.com/rnd/app/da/glimmix.html

Shahian, D. M., and S.-L. T. Normand. 2003. "The Volume–Outcome Relationship: From Luft to Leapfrog." *Annals of Thoracic Surgery* 75: 1048–58.

Silber, J. H., P. S. Romano, A. K. Rosen, Y. Wang, R. N. Ross, O. Even-Shoshan, and K. Volpp. 2007. "Failure-to-Rescue: Comparing Definitions to Measure Quality of Care." *Medical Care* 45: 918–25.

Silber, J. H., P. R. Rosenbaum, P. S. Romano, A. K. Rosen, Y. Wang, Y. Teng, M. J. Halenar, O. Even-Shoshan, and K. G. Volpp. 2009. "Hospital Teaching Intensity, Patient Race, and Surgical Outcomes." *Archives of Surgery* 144: 113–21.

The Leapfrog Group. "White Paper Available to Explain Leapfrog's New Ebhr Survival Predictor" [accessed on April 19, 2010]. Available at http://www.leapfroggroup.org/news/leapfrog_news/4729468

Thiemann, D. R., J. Coresh, W. J. Oetgen, and N. R. Powe. 1999. "The Association between Hospital Volume and Survival after Acute Myocardial Infarction in Elderly Patients." *New England Journal of Medicine* 340: 1640–8.

Tu, J. V., P. C. Austin, and B. T. Chan. 2001. "Relationship between Annual Volume of Patients Treated by Admitting Physician and Mortality after Acute Myocardial Infarction." *Journal of the American Medical Association* 285: 3116–22.

U.S. Department of Health & Human Services. 2007a. "Hospital Compare—A Quality Tool for Adults, Including People with Medicare (December 19)" [accessed on January 28, 2008]. Available at http://www.hospitalcompare.hhs.gov/Hospital/Home2.asp?version=alternate&browser=IE%7C6%7CWinXP&language=English&defaultstatus=0&pagelist=Home

U.S. Department of Health & Human Services. 2007b. "Information for Professionals. Outcome Measures: Adjusting for Small Hospitals or a Small Number of Cases (December 12)" [accessed on January 28, 2007]. Available at http://www.hospitalcompare.hhs.gov/Hospital/Static/Data-Professionals.asp?dest=NAV|Home|DataDetails|ProfessionalInfo#TabTop

Werner, R. M., and D. A. Asch. 2005. "The Unintended Consequences of Publicly Reporting Quality Information." *Journal of the American Medical Association* 293: 1239–44.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article:

Appendix SA1: Author Matrix.
Appendix SA2: Electronic Appendix: Additional Models.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.