Journal of the American College of Cardiology
© 2000 by the American College of Cardiology
Published by Elsevier Science Inc.

Vol. 36, No. 7, 2000
ISSN 0735-1097/00/$20.00
PII S0735-1097(00)01022-6

Challenges in Comparing Risk-Adjusted Bypass Surgery Mortality Results

Results From the Cooperative Cardiovascular Project

Eric D. Peterson, MD, MPH, FACC,* Elizabeth R. DeLong, PHD,* Lawrence H. Muhlbaier, PHD,* Allison B. Rosen, MD, MPH,* Hope E. Buell, MS,* Catarina I. Kiefe, MD, PHD,† Timothy F. Kresowik, MD, MPH‡

Durham, North Carolina, Birmingham, Alabama, West Des Moines, Iowa

OBJECTIVES We sought to evaluate the predictive accuracy of four bypass surgery mortality clinical risk models and to examine the extent to which hospitals' risk-adjusted surgical outcomes vary depending on which risk-adjustment method is applied.

BACKGROUND Cardiovascular "report cards" often compare risk-adjusted surgical outcomes; however, it is unclear to what extent the risk-adjustment process itself may affect these metrics.

METHODS As part of the Cooperative Cardiovascular Project's Pilot Revascularization Study, we compared the predictive accuracy of four bypass clinical risk models among 3,654 Medicare patients undergoing surgery at 28 hospitals in Alabama and Iowa. We also compared the agreement in hospital-level risk-adjusted bypass outcome performance ratings depending on which of the four risk models was applied.

RESULTS Although the four risk models had similar discriminatory abilities (C-index, 0.71 to 0.74), certain models tended to overpredict mortality in higher-risk patients. There was high correlation between a hospital's risk-adjusted mortality rates regardless of which of the four models was used (correlation between risk-adjusted rating, 0.93 to 0.97). In contrast, there was limited agreement in which hospitals were identified as "performance outliers" depending on which risk-adjustment model was used and how outlier status was defined.

CONCLUSIONS A hospital's risk-adjusted bypass surgery mortality rating, relative to its peers, was consistent regardless of the risk-adjustment model applied, supporting their use as a means of provider performance feedback. Designation of performance outliers, however, can vary significantly depending on the benchmark and methods used for this determination. (J Am Coll Cardiol 2000;36:2174–84) © 2000 by the American College of Cardiology

As early as the nineteenth century, Florence Nightingale recognized the value of comparing hospital mortality rates as a means of assessing quality of care (1). Since then, others have reinforced the importance of providing caregivers with outcomes feedback as a necessary step toward continual quality improvement (2–6). Although comparing patient outcomes is important, it is clear that these results need to be adjusted for potential differences in type, or "case-mix" of patients cared for by various caregivers. To allow for such comparisons on a leveled playing field, researchers use a statistical approach known as risk-adjustment (7,8). One common risk-adjustment mechanism uses a statistical model that adjusts for individual patient risk factors while predicting the event of interest. With such a "risk-prediction" model, one can calculate a provider's expected clinical event rate (based on their patients' summated estimated risk) and compare this expected rate with observed results.

Many of the prototypic provider-level comparisons of risk-adjusted outcomes have examined mortality rates following coronary artery bypass surgery. New York State and Pennsylvania routinely compare and publish hospital- and surgeon-specific risk-adjusted bypass surgery mortality results as a means of increasing consumer awareness (9–13). Other voluntary groups of health-care providers internally share bypass surgery mortality data as a means of "benchmarking" outcomes performance results across centers and promoting quality improvement efforts (14–20).

Commonly, bypass surgery outcomes performance measures are "risk-adjusted" using one of several published surgical mortality models (9,21–23). These surgical models were developed in separate patient populations with significantly different event rates (Table 1). In part, because of these differences, individual risk factors and the "weighting" of these factors vary among models (Appendix 1). To date, few have attempted to assess and compare the predictive accuracy of these models when applied outside of the database in which they were developed (24–28). Furthermore, the impact of different risk-adjustment models on a

From *The Duke Outcomes Research and Assessment Group, Duke University Medical Center, Durham, North Carolina; †The Alabama Quality Assurance Foundation, and the University of Alabama at Birmingham Center for Outcomes and Effectiveness Research and Education, Birmingham, Alabama; and ‡The Iowa Foundation for Medical Care, West Des Moines, Iowa. Supported in part by grant HS 06503-03 Supplement 2 from the Health Care Financing Administration through the Agency for Health Care Policy and Research; and R01 HS09940-01A1 from the Agency for Health Care Policy and Research.

Manuscript received July 26, 1999; revised manuscript received June 1, 2000, accepted July 14, 2000.

Abbreviations and Acronyms

CABG = coronary artery bypass surgery
CCP = Cooperative Cardiovascular Project
O/E = ratio of observed mortality to expected mortality
RS = risk score

provider's bypass surgery performance rating has not been assessed. If a provider's performance rating shifted from superior to inferior depending on which risk-adjustment model was used, then the face validity of the risk-adjustment process would be in question. Lacking this information, clinicians have generally been skeptical of the risk-adjusted outcomes profiling efforts (29–31).

We evaluated the predictive accuracy of four commonly used bypass surgery-specific risk-adjustment tools in a large, community-based elderly population. We also examined the extent to which a hospital-level risk-adjusted surgical outcome rating varied depending on which risk-adjustment model was applied. We then repeated the process above after the risk models were adjusted (recalibrated) to reflect the mortality rates in our elderly study population. Finally, we assessed whether "outlier hospitals" (providers identified as having significantly superior or inferior outcomes) changed depending on which risk model was used and how outlier performance was defined.

METHODS

Bypass surgery risk models. We considered four nonproprietary models that estimated short-term mortality risk following bypass surgery or open heart surgery. These four models will be referred to in this article by their first author's last name, including the Parsonnet, O'Connor, Higgins, and Hannan models (9,21–23). The Parsonnet model was developed on 3,500 patients undergoing coronary bypass and/or valve surgery in New Jersey between 1982 and 1987 (21). The O'Connor model was developed on data from 3,055 patients receiving isolated bypass surgery procedures at five northern New England hospitals between 1987 and 1989 (22). The Higgins model developed a risk prediction model using data from 5,051 patients undergoing bypass surgery at the Cleveland Clinic between 1986 and 1988 (23). Hannan and colleagues developed a bypass surgery risk model for New York State using a population of 57,187 patients operated on between 1989 and 1992 (9). The clinical risk factors included in each model are displayed in Appendix 1.

Patient population. The Cooperative Cardiovascular Project (CCP) Pilot Revascularization Study was a joint quality improvement effort between the Health Care Financing Administration, state peer review organizations, and several national medical societies (including the American Medical Association, American College of Physicians, American College of Cardiology and the American Academy of Family Physicians) (32,33). The study population included all patients aged 65 years or older covered by Medicare who underwent isolated bypass surgery procedures in Alabama and Iowa between June 1, 1992, and February 28, 1993. To avoid double counting patients, those who underwent more than one bypass surgery procedure during the study period were included only once as defined by their initial procedure. We also excluded those who received a procedure at an institution that performed in total

Table 1. Baseline Patient Characteristics

	Parsonnet	O'Connor	Higgins	Hannan	CCP
Year(s) of data entry	1982–1987	1987–1989	1986–1988	1989–1992	1992–1993
No. of patients	3,500	3,055	5,051	57,187	3,654
No. of hospitals	N/A	5	1	30	28
Mortality rate* (%)	8.9	4.3	2.5	3.1	5.6
Mean age (yr)	65	64	N/A	65	72
Male (%)	69	73	79	73	65
Diabetes mellitus (%)	35	18	17	24	27
Congestive heart failure (%)	7	N/A	9	15	11
Pulmonary disease (%)	4	11	8	17	12
Prior bypass surgery (%)	6	6	19	8	12
Mean LVEF (%)	46	58	N/A	45	49
Left main disease (%)	17†	21	N/A	20	18
No. of diseased coronary vessels					
One (%)	N/A	12	N/A	9	12
Two (%)		35		26	36
Three (%)		53		62	51
Surgical priority (%)					
Emergent	5	9	3	7	12
Urgent	72	56	N/A	42	5
Elective	23	36	N/A	51	83

CCP = Cooperative Cardiovascular Project Pilot Revascularization Study; LVEF = left ventricular ejection fraction; N/A = not reported in original publication.

*Mortality rate refers to in-hospital rate for Hannan, O'Connor, and CCP data and 30-day mortality rate for Parsonnet and Higgins data. †Recorded as left main stenosis >90% versus >50% for others.

less than 50 Medicare surgical procedures during the study period. Each of the models was developed using logistic regression, in which the risk of mortality for a patient with a vector X of risk factors is given as

$$p(\chi) = \frac{\exp(\beta\chi)}{1 + \exp(\beta\chi)}$$

Here β is a vector of coefficients associated with the risk factors and the linear combination β_χ is called the risk score (RS).

Data collection and mortality end points. Patients were identified using Medicare claims data (ICD-CM Codes 36.10–19), and the medical records of eligible patients were reviewed by trained nurse clinicians. Detailed clinical and demographic data were collected via chart abstraction using standardized definitions. This abstraction tool was designed prospectively to contain the main data elements used in published surgical mortality prediction models. The CCP data definitions were matched to the extent possible to those used in the prior model populations. In-hospital mortality rate was chosen as the end point of interest as most community hospitals lack the ability to track postdischarge events. Of note, while the Parsonnet and Higgins models were initially developed to predict the risk of mortality within 30 days of surgery, their predictive accuracy was higher for predicting in-hospital mortality events when tested in CCP data.

Analysis: model validation. We used two standard measures of a model's performance: discrimination and calibration. Discrimination is the ability of a mortality model to correctly distinguish those patients who will die from those who will survive. An overall measure of model discrimination can be summarized by its area under a receiver operator characteristic (ROC) curve or C-index (34,35). A model's C-index can range from 0.5 (no predictive ability) to 1.0 (perfect predictive accuracy). A second measure, calibration, examines how closely the model's predicted mortality rates match observed mortality rates for various risk groups of patients (36). To assess this, patients were rank-ordered by their predicted mortality. Patients were then grouped into five similarly sized risk groups and the average expected mortality rate for each group was compared with that actually observed.

If a given model is not well calibrated (i.e., it significantly underpredicts or overpredicts mortality) when applied in a new population, one can consider recalibrating the model. There are several mechanisms for implementing such prevalence corrections (37). We used logistic regression to fit an intercept term (α) and a multiplier term (β^*) to the original risk score (RS). The statistical formulation for this secondary logistic regression model can be summarized as follows: $\ln [p^*/1 - p^*)] = \alpha^* + \beta^*$ RS, where p^* is the revised predicted probability for mortality, RS is the original risk score and α^* and β^* are estimated when the model is applied in the current population.

Hospital-level risk-adjusted outcomes measures. We calculated an "expected mortality rate" for each of the 28 hospitals by aggregating their patients' individual estimated mortality risk and dividing by the total number of patients treated at that hospital. We then calculated a hospital's ratio of observed mortality rate to its expected (O/E). Hospitals with O/E ratios <1 were institutions with lower (better) observed bypass surgery mortality than predicted. Conversely, hospitals with O/E ratios >1 reflected higher mortality than predicted. We repeated this process using each of the four risk adjustment models, producing four O/E ratios for each hospital. We also repeated this process after each of the original risk models was revised and recalibrated in the CCP patient population. Confidence intervals surrounding hospitals' O/E ratios were computed based on the normal approximation to the binomial distribution.

We also created risk-adjusted mortality rates from each model by multiplying the O/E ratios by the overall average CCP mortality rate (9). The correlation between each of these four risk-adjusted hospital mortality rates and the hospital's unadjusted mortality rate was assessed graphically and using Spearman correlation coefficients. A hospital was considered to have outlier performance if the 95% confidence interval around its O/E ratio excluded 1.0. As an alternative method for identifying outliers, we estimated the individual effect of hospital performance on outcome using a random effects regression model (37). With this, a "shrunken estimate" of a provider's influence on outcome is determined relative to its peers and after adjusting for underlying risk (38).

RESULTS

The CCP revascularization database included 4,152 Medicare patients undergoing isolated bypass surgery at 32 hospitals in Alabama and Iowa between June 1, 1992 and February 28, 1993. From this cohort, we excluded 390 patients who were <65 years old. We also excluded 108 patients who received bypass surgery at any of the four institutions that performed fewer than 50 surgical procedures on Medicare patients during the study period. Thus, the final CCP analysis cohort consisted of 3,654 bypass surgery patients from 28 separate institutions. The mean and median number of Medicare bypass patients per hospital during this nine month period was 132 and 124, respectively. Given that patients aged 65 or older make up approximately half of an average hospital's case volume, the estimated mean yearly surgical volumes for all-aged patients at these hospitals would be 352 cases.

Baseline characteristics. Baseline clinical characteristics for CCP patients were compared with those from the development populations for the Parsonnet, O'Connor, Higgins and Hannan risk-adjustment models (Table 1). The CCP cohort contained older patients, more women, a

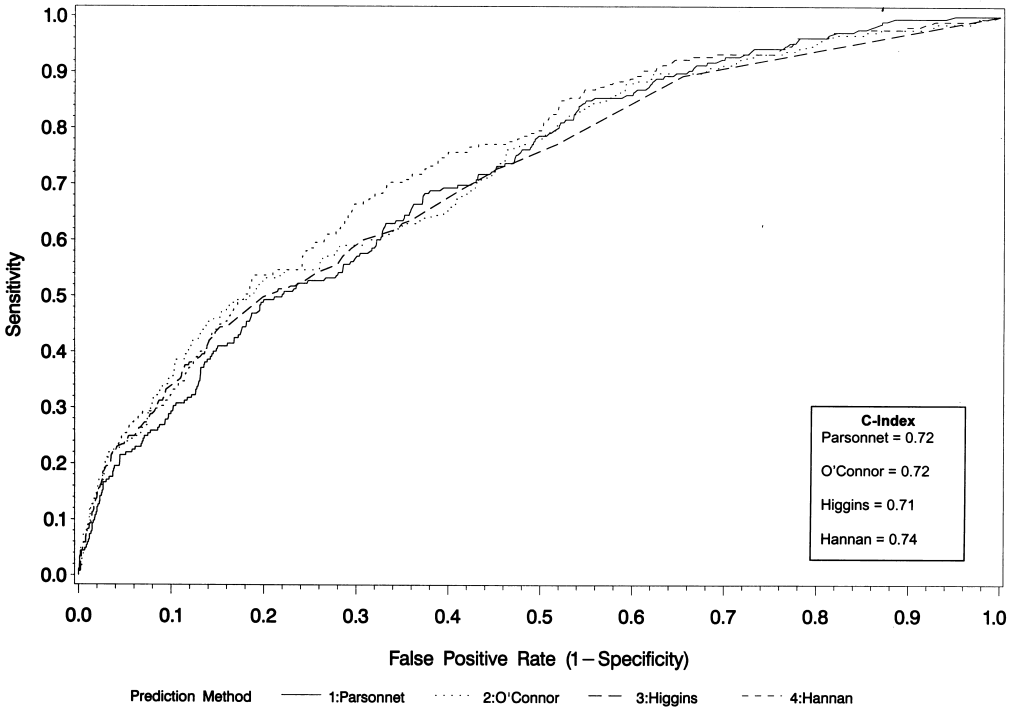

Figure 1. This figure the ROC curves for the four bypass surgery risk models. The C-index is equivalent to the area under each ROC curve.

higher percentage of those undergoing prior revascularization procedures and procedures under emergent conditions. Rates of most comorbid illnesses were similar across the four cohorts, as was the severity of underlying coronary stenoses, frequency of significant left main stenosis and degree of left ventricular dysfunction. The overall observed surgical mortality rates in these cohorts varied from 2.5% in the Higgins et al. (23) study to 8.9% in the Parsonnet et al. (21) study.

Model performance. The discrimination abilities for each of the external bypass surgery models is displayed in Figure 1. The area under the ROC curve or C-index for these models was 0.72 for the Parsonnet model, 0.71 for Higgins, 0.72 for O'Connor and 0.74 for Hannan. For comparison, the C-indexes for these models in their original populations were 0.74 for Higgins, 0.74 for the O'Connor model, and 0.79 for Hannan (note: Parsonnet's C-index was not published).

Figure 2A demonstrates how well calibrated each original model was when applied in the CCP patient population. This figure displays observed versus expected in-hospital bypass surgery mortality results for each of the models by quintiles of patient risk. The diagonal line in this figure represents perfect agreement. The predicted mortality rates based on the Hannan and O'Connor models were quite close to those actually observed for nearly all risk groups. For example, using the Hannan model, the lowest and highest risk groups had predicted versus observed mortality rates of (1.4% vs. 1.2%) and (13% vs. 14%), respectively. In contrast, the Parsonnet and Higgins models consistently overpredicted mortality rates, particularly in higher risk patients. Among the highest risk group, the Parsonnet model predicted mortality rate was nearly twice that actually observed (23% vs. 13%).

Figure 2B displays these same risk estimates after the models were individually internally recalibrated within the CCP population (see Methods section). After recalibration, each of the models was better able to accurately estimate surgical mortality rates across a wide range of patient risk categories.

Risk-adjusted outcomes. Figure 3A displays the rank ordering of the 28 hospitals by their unadjusted bypass mortality rates (dash) versus their risk-adjusted mortality rates based on the Parsonnet (square) and Hannan risk models (circle). As noted, when certain risk models (e.g., Parsonnet) are used, the majority of hospitals' "risk-adjusted" mortality rates appeared lower than those actually observed. However, the hospitals' relative performance (compared with peers) were generally consistent regardless of which risk-adjustment model was used. This consistency between relative risk-adjusted hospital performance results becomes even more marked after the models are internally recalibrated (Figure 3B).

Table 2 displays the formal association between the various risk-adjusted mortality rates. Hospital risk-adjusted mortality rates using any of the four models were highly correlated, with Spearman correlation coefficients ranging from 0.96 for Higgins-Hannan comparison to 0.99 for Parsonnet-O'Connor comparison (Table 2). Additionally, the correlation between any two risk-adjusted mortality

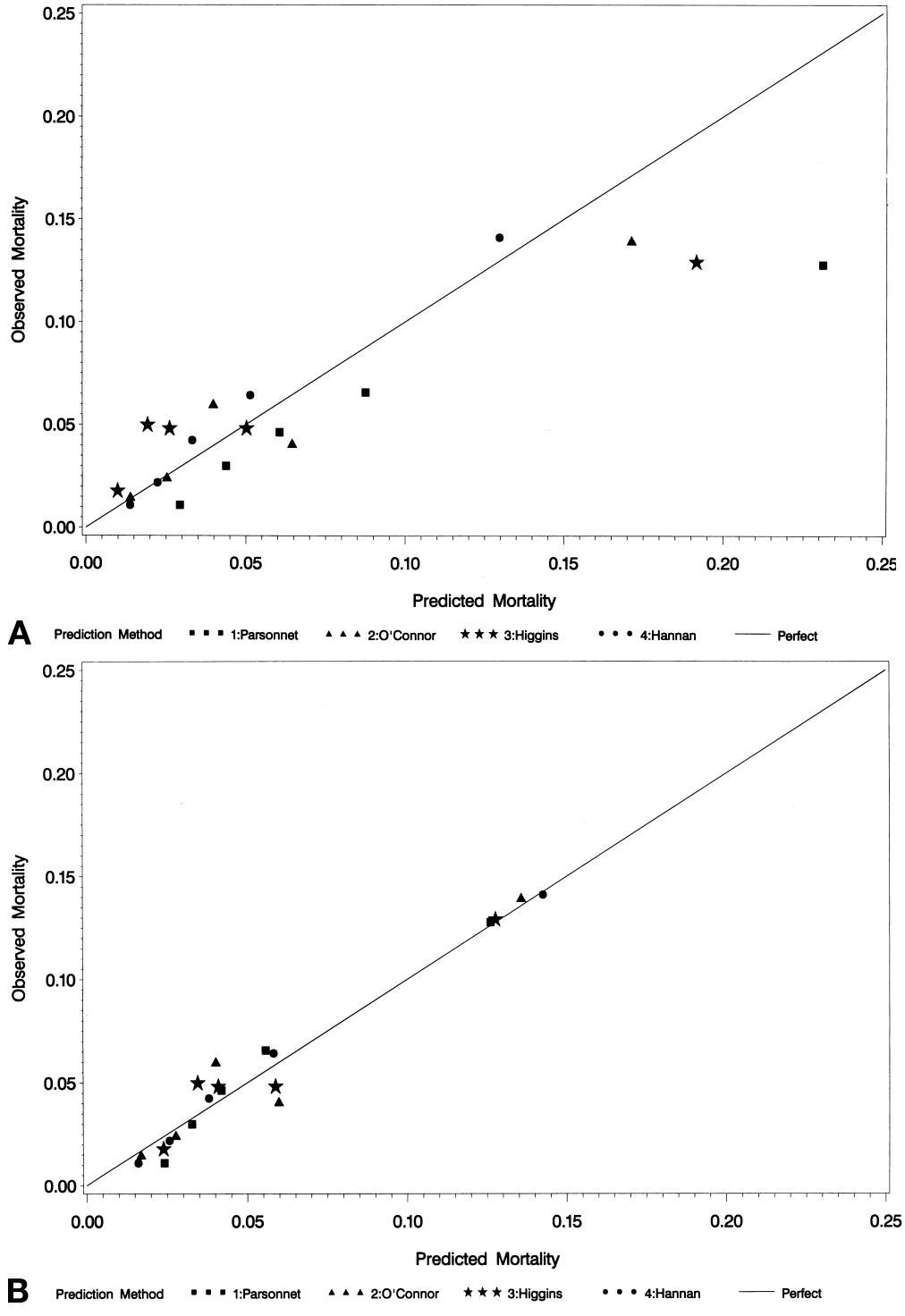

Figure 2. A, The observed to expected mortality rates for each quintile of patient risk. Each risk quintile contains approximately 750 patients. The diagonal line represents perfect agreement between observed and expected mortality estimates. **B,** The same information after the models have been internally recalibrated in the CCP database.

rates was consistently greater than the correlation between these risk-adjusted mortality rates and unadjusted mortality outcomes.

Hospital outlier status. Besides comparing relative performance, hospital-specific risk-adjusted outcomes are often used to identify "superior or inferior performers." Outlier performance, however, can be assessed by different metrics. Table 3 displays those hospitals for which the observed bypass mortality rates were significantly higher or lower than those predicted by each of the risk models (i.e., 95% confidence intervals for an O/E ratio excluding 1.0). Using this performance measure, the original Parsonnet model identified 10 significantly superior hospitals, but no hospitals as inferior performers. In contrast, the original Hannan

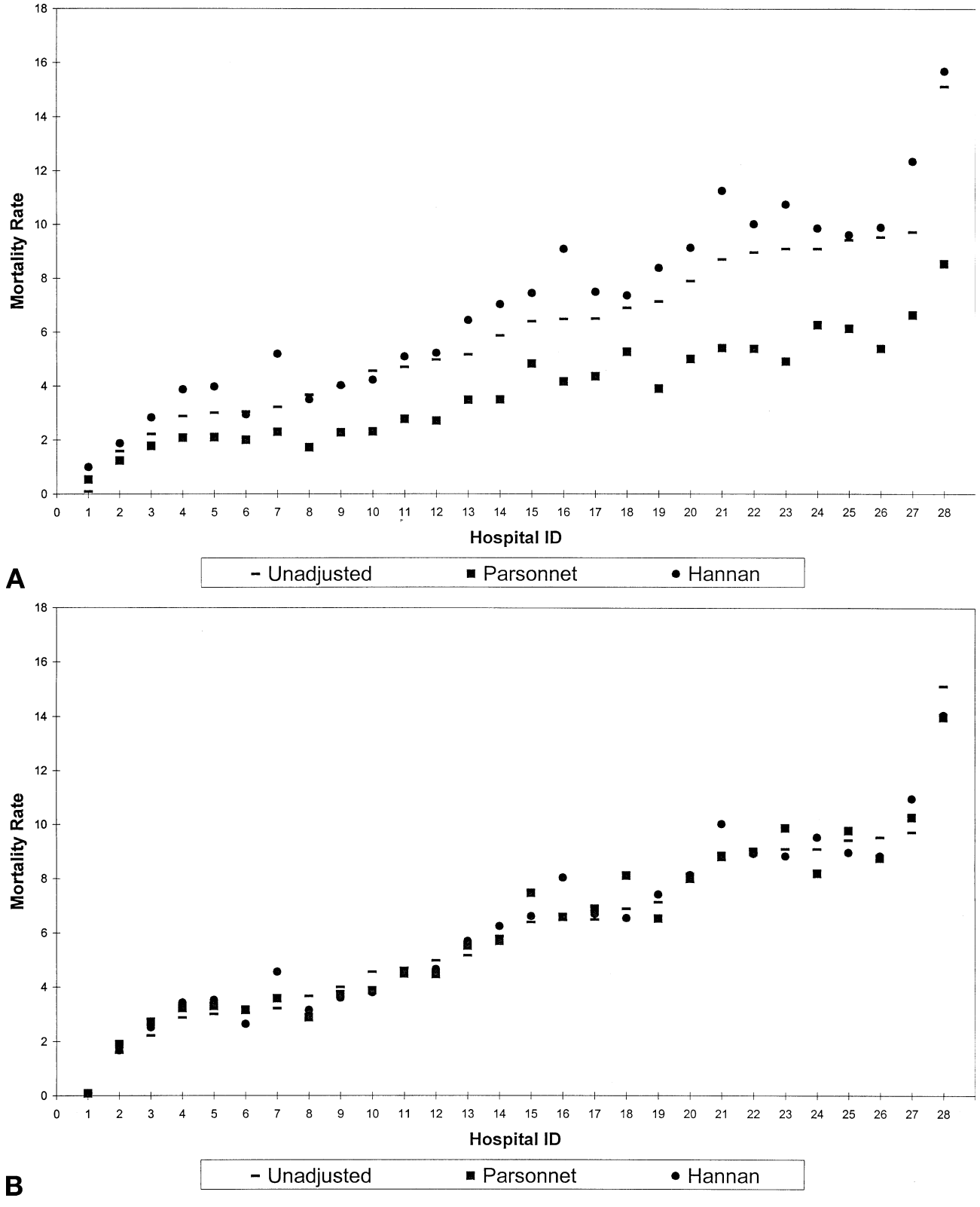

Figure 3. A, Each hospital's unadjusted mortality rates and their risk-adjusted mortality using the Parsonnet and Hannan risk models. Note: the 28 Hospitals are ordered on the x-axis by the unadjusted mortality rate. **B,** This same information after the Parsonnet and Hannan models have been internally recalibrated in the CCP database.

model identified only one significantly superior hospital and four inferior hospitals. Thus, complete agreement on outlier status using this method occurred in only one of the 28 hospitals (ID No. 1), with this identifying a superior performer. Table 4 displays similar information, but now based on their risk-adjusted mortality using the internally

Table 2. Correlation* Between Hospitals' Unadjusted and Risk-Adjusted Bypass Mortality Rates

Mortality Rate	O'Connor	Higgins	Hannan	Unadjusted
Parsonnet	0.99	0.97	0.99	0.85
O'Connor	—	0.98	0.98	0.85
Higgins		—	0.96	0.80
Hannan			—	0.83

*Correlation coefficient = Spearman test.

recalibrated risk models. After recalibration of the models, agreement in outlier status was generally consistent.

As a final means, Table 5 displays which hospital's bypass performance was deemed significantly better or worse than its peers when assessed by a random-effects logistic model. This more conservative statistical method identified few high or low outliers regardless of whether or which risk-adjustment was used. However, using this method, there was complete agreement that one hospital (ID No. 28) had significantly worse bypass outcomes by all five methods.

DISCUSSION

The era of "scorecard medicine," in which provider-specific procedure outcomes results are openly compared, is here (39). State peer review boards, insurers, corporate employers and patients are all requesting this information as a means of assessing and comparing health-care quality (40). Whether clinicians agree or not with the basic tenets of provider comparisons, nearly all agree that, if outcomes are to be compared, it should be done only after appropriately risk-adjusting the results (41,42). We studied four risk models that were specifically developed to predict procedural mortality following bypass surgery. We found that these models retained most of their discriminatory ability when applied in a community-based, elderly patient population. Most im-

Table 3. Comparison of Outlier Hospitals by Observed to Expected Ratios

	Better Than Expected Outliers*			
Hospital ID	**Parsonnet O/E (95% CI)**	**O'Connor O/E (95% CI)**	**Higgins O/E (95% CI)**	**Hannan O/E (95% CI)**
1	0.10 (0.0–0.6)	0.17 (0.0–0.9)	0.18 (0.0–0.9)	0.18 (0.0–0.9)
3	0.32 (0.0–0.8)	0.32 (0.0–0.8)	0.40 (0.0–0.9)	
4	0.37 (0.0–0.9)			
5	0.37 (0.0–0.9)	0.45 (0.0–1.0)		
6	0.35 (0.0–0.8)	0.45 (0.0–1.0)		
8	0.31 (0.0–0.6)	0.46 (0.10–0.9)	0.52 (0.1–1.0)	
9	0.41 (0.1–0.7)			
10	0.41 (0.0–0.9)			
11	0.50 (0.1–0.9)			
12	0.48 (0.1–0.9)			
	Worse Than Expected Outliers*			
Hospital ID	**Parsonnet Score (95% CI)**	**O'Connor Score (95% CI)**	**Higgins Score (95% CI)**	**Hannan Score (95% CI)**
21				1.63 (1.0–2.3)
23				1.71 (1.0–2.4)
24			2.13 (1.1–3.1)	(1.1–3.1)
27		1.77 (1.0–2.5)	1.77 (1.0–2.5)	2.20 (1.3–3.1)
28		2.46 (1.7–3.3)	2.25 (1.6–2.9)	2.80 (2.0–3.7)

Hospital ID No. corresponds to unadjusted mortality ranking from Figure 3.
*Outlier defined as hospitals whose 95% confidence interval (CI) for observed to expected mortality (O/E) ratio does not include 1.0 before rounding.

Table 4. Comparison of Outlier Hospitals by Observed to Expected Ratios After Recalibration

Better Than Expected Outliers*				
Hospital ID	**Parsonnet O/E (95% CI)**	**O'Connor O/E (95% CI)**	**Higgins O/E (95% CI)**	**Hannan O/E (95% CI)**
1	0.16 (0.0–0.9)	0.19 (0.0–1.0)	0.17 (0.0–0.9)	0.16 (0.0–0.9)
3		0.36 (0.0–0.9)	0.40 (0.0–0.9)	0.45 (0.0–1.0)
8	0.52 (0.1–0.9)			
Worse Than Expected Outliers*				
Hospital ID	**Parsonnet Score (95% CI)**	**O'Connor Score (95% CI)**	**Higgins Score (95% CI)**	**Hannan Score (95% CI)**
25	1.7 (1.1–2.4)			
27	1.8 (1.1–2.6)	1.9 (1.1–2.7)	1.9 (1.1–2.7)	2.2 (1.1–2.8)
28	2.5 (1.7–3.3)	2.8 (1.9–3.6)	2.5 (1.7–3.3)	2.5 (1.7–3.3)

Hospital ID No. corresponds to unadjusted mortality ranking from Figure 3.
*Outlier defined as hospitals whose 95% confidence interval (CI) for observed to expected mortality (O/E) ratio does not include 1.0 before rounding.

portantly, we found that a hospital's risk-adjusted outcomes relative to its peers was remarkably consistent, regardless of the risk model used. However, we found that model calibration varied and may markedly affect which hospitals were deemed superior or inferior performers.

While many bypass surgery risk models have been published, their comparative predictive accuracy outside of the databases in which they were developed has been rare (24–28). Iezzoni and colleagues (28) tested the predictive accuracy of five generic clinical and administrative severity of illness measures when applied in a population of bypass surgery patients. They found that the discrimination abilities of these generic risk tools were similar (C-index for the two clinical models 0.72 to 0.73 and 0.77 to 0.83 for the three administrative systems). The paradoxical better performance of the claims-based models over clinical-based ones was accounted for by the fact that administrative models often included postoperative complication data (e.g., cardiac arrest or heart failure) as preoperative risk predictors. Similar to our results, the authors also found substantial agreement in relative risk-adjusted hospital mortality rates, regardless of which risk model was applied (27).

In another study, Orr and colleagues (26) tested the predictive accuracy of four bypass-specific clinical risk models among 868 bypass patients in a single institution. Consistent with our findings, these authors also found that the discrimination abilities of the risk models in their hospital ranged from C-index 0.70 to 0.74. They also found that the Parsonnet model significantly overpredicted mortality while the Hannan model significantly underpredicted mortality rates. As this was a single-institution study, the authors were unable to examine the impact of the risk models on comparative provider performance.

Table 5. High and Low Performers Using Random Effects Model

Method	Significantly Better Hospital (ID No.)	Significantly Worse Hospital (ID No.)
Unadjusted	None	28
Risk-adjusted		
-Parsonnet	None	28
-O'Connor	3	28
-Higgins	None	28
-Hannan	1	28

Hospital ID No. corresponds to their unadjusted ranking from Figure 3.

Our study expands on this work by examining the impact of bypass surgery-specific risk models in a large multiinstitutional study. Additionally, our elderly patient population provided a more stringent test of the models' predictive accuracy as their risk profiles differed significantly from the patient samples used to originally create the models (Table 1). Despite this, we found that the discrimination ability of all four models was generally well preserved when applied in our higher risk elderly patients. Model calibration, however, was an issue for the Parsonnet and Higgins models, in which expected mortality differed significantly from observed, particularly in the high risk patient subgroups (Fig. 2). In contrast to the work of Orr et al. (26), in our data, expected mortality rates generated by the Hannan model tended to match closely those observed in all patient risk subgroups.

Impact of risk adjustment method on hospital performance measures. There are numerous reasons for comparing risk-adjusted hospital outcomes data. First, a physician, payor or patient may want a general sense of how their hospital's bypass surgery outcomes compare with community peers. In this context, our data suggest that the application of various risk-adjustment models will result in similar hospital-level relative performance measures. In

other words, if a hospital was generally a good performer (relative to its peers) using one risk-adjustment model, then it was likely to be a good performer no matter which risk-adjustment model was used. These conclusions are also consistent with the findings from other similar studies (27,37) and should decrease clinicians' concerns that their performance outcomes are somehow an artifact of the specific risk-adjustment method applied.

It should be emphasized, however, that these results do not imply that risk-adjustment is unnecessary for outcomes comparisons. In fact, we found a much stronger correlation between any two risk-adjusted outcomes methods than any risk-adjustment method and unadjusted data, indicating that some form of risk-adjustment is required for appropriate comparison. Additionally, if comparisons are to be made among hospitals, the same risk-adjustment model should be applied to all centers. For example, it would be inappropriate to compare one hospital's risk-adjusted mortality rate based on Parsonnet with another center's result based on Hannan.

Beyond relative performance evaluation, risk-adjusted outcomes data are often used to identify the "good and bad apples" (e.g., those institutions with exceptional performance) (8). Often a hospital's risk-adjusted performance may be compared against some standard or benchmark (43,44). One of the most common methods employs a strictly external risk-adjustment model to compare O/E mortality ratios. If this method is used, the hospitals are actually being compared with an external performance benchmark (i.e., that observed in hospitals in the original risk model study population). For example, the Parsonnet model was developed on all patients undergoing open heart surgery (including higher risk valve cases) operated on in the early 1980s. As such, these patients had high bypass mortality rates relative to contemporary, bypass-only, outcomes. When applied in our population, the Parsonnet model significantly overpredicted mortality risks, making a third of CCP hospitals appear to be superior hospitals and none being inferior (Table 3). If, instead, the Hannan model, based on more contemporary bypass-only cases from New York (a state with the lowest US bypass mortality) (45), were applied as the external benchmark, only one hospital would have been identified as a superior performer while four would be significantly inferior. While the selection of an external benchmark is arbitrary, it seems reasonable to select a risk-adjustment model that is both clinically meaningful and with overall event rates similar to those found in the study population.

Alternatively, published models can be refit or recalibrated to match event rates in a new patient group. While various methods of model recalibration have been proposed, the process, however, is quite analogous to the developing of a "new" prediction model. As such, it requires a sufficiently large study population to assure stable model performance, as well as appropriate analytical oversight. When recalibration is achieved, outlier status will be based on internal (or peer), as opposed to external performance standards. Our study demonstrated that hospital performance metrics (including designation of outlier status) were quite consistent after recalibration regardless of which of the four published risk models was used as the start-point for this process.

A final alternate method of determining outlier status is also based on internal performance standards (Table 5). In contrast to the prior method, this technique does not require model recalibration to achieve this goal. Specifically, this random-effects statistical model will determine if a hospital's surgical mortality rate differs significantly from that seen in the other comparison hospitals, after adjusting for baseline risk (based on one of the published risk models). Although this technique has some advantages (37), it remains possible that an average or better hospital could be singled out as a "poor performer" if all its comparison centers were outstanding. Additionally, this method is conservative and is less likely to identify outliers at low volume centers. To gain the clearest idea of one's surgical outcomes, it is ideal to benchmark one's results both among one's regional and national peers.

Study limitations. The performance of any surgical risk model depends in part on whether all of the variables used in the original model are collected, the degree to which variable definitions are congruent and how accurately these variables are collected. As noted, the CCP data definitions were prospectively designed to be consistent with variables needed, but slight variation between the definitions used by CCP and those used in the original patient population was unavoidable. Our study population was also limited to Medicare patients. The performance of these models may be different in all aged bypass patients. Additionally, the duration of data collection and number of cases studied per hospital was somewhat limited (average 132), which limited the power of our study to identify outlier performers.

CONCLUSIONS

Comparing hospital-specific outcome data will remain a challenging exercise. This information has the potential to give both patients and clinicians important feedback concerning a center's quality of care. However, these data can also be confounded if the results do not take into account the surgical risks of the patients treated. Using bypass surgery as a test case, we found that published surgery risk-adjustment models varied in their ability to accurately predict mortality when applied in a community-based elderly population. Despite differences in model calibration, a hospital's risk-adjusted surgical outcomes results, relative to its peers, tended to be quite consistent regardless of which risk-adjustment model was applied. The identification of outliers (with superior or inferior surgical results) varied, however, depending on which performance benchmark was used.

These data support the concept of risk-adjusting outcomes comparisons, but re-emphasize the importance of the

Appendix 1. Comparison of Covariates in Risk-Adjustment Algorithm

Risk Factor	Parsonnet Algorithm	Higgins Algorithm	O'Connor Algorithm	Hannan Algorithm
Age	x	x	x	x
Gender	x		x	x
Smoking	x			
Hypertension	x			
Diabetes mellitus	x			x
Vascular disease		x		
Chronic pulmonary disease		x		x
Anemia		x		
Renal insufficiency		x		
Dialysis dependence	x			x
Obesity*	x		x	x
Charlson comorbidity score			x	
Congestive heart failure				x
Unstable angina				x
Recent myocardial infarction				x
LV ejection fraction	x	x	x	x
LV end diastolic pressure			x	
Left main stenoses				x
Mitral valve disease	x	x		
Aortic valve disease	x			
LV aneurysm	x			
Prior bypass surgery	x	x	x	x
Priority at surgery		x	x	
Preoperative intra-aortic balloon pump	x			x
Catastrophic states	x			x

LV = left ventricular.
*O'Connor model uses the continuous variable body surface area (BSA).

risk-adjustment process. To be meaningful, consumers of these new "risk-adjusted outcomes report cards" must understand clearly how their data were analyzed and to what benchmark their results were compared.

Reprint requests and correspondence: Dr. Eric D. Peterson, Box 3236, Duke University Medical Center, Durham, North Carolina 27710.

REFERENCES

1. Nightingale F. Notes on Hospitals. 3rd ed. London: 1863.
2. Codman EA. A Study in Hospital Efficiency as Demonstrated by the Case Report of the First Five Years of a Private Hospital. Boston: Thomas Todd Company Printers, 1917.
3. Donabedian A. The Methods and Findings of Quality Assessment and Monitoring: An Illustrated Analysis. Ann Arbor, MI: Health Administration Press, 1985.
4. Donabedian A. The end results of health care: Ernest Codman's contribution to quality assessment and beyond. Milbank Q 1989;67: 233–56.
5. Ellwood P. Shattuck lecture—outcomes management: a technology of patient experience. N Engl J Med 1988;318:1549–56.
6. Relman AS. Assessment and accountability: the third revolution in medical care. N Engl J Med 1988;318:1220–22.
7. Iezzoni LI. Risk and outcomes. In: Iezzoni LI, editor. Risk Adjustment for Measuring Healthcare Outcomes. Chicago: Health Administration Press, 1997:1–41.
8. Iezzoni LI. The risks of risk adjustment. JAMA 1997;278:1600–7.
9. Hannan EL, Kilburn H, Racz M, Shields E, Chassin MR. Improving the outcomes of coronary artery bypass surgery in New York State. JAMA 1994;271:761–6.
10. Hannan EL, Kilburn H Jr, O'Donnell JF. Adult open heart surgery in New York State: an analysis of risk factors and hospital mortality rates. JAMA 1990;264:2768–74.
11. Chassin MR, Hannan EL, BeBuono BA. Benefits and hazards of reporting medical outcomes publicly. N Engl J Med 1996;334: 394–8.
12. Localio AR, Hamory BH, Fisher AC, TenHave TR. The public release of hospital and physician mortality data in Pennsylvania: a case study. Med Care 1997;35:272–86.
13. Bentley JM, Nash DB. How Pennsylvania hospitals have responded to publicly released reports on coronary artery bypass graft surgery. Joint Commission J Qual Improvement 1998;24:40–9.
14. Grover FL, Hammermeister KE, Burchfiel C. Initial report of the Veterans Administration preoperative risk assessment study for cardiac surgery. Ann Thorac Surg 1990;50:12–26.
15. Marshall G, Shroyer LW, Grover FL, Hammermeister KE. Time series monitors of outcomes: a new dimension for measuring quality of care. Med Care 1998;36:348–56.
16. O'Connor GT, Plume SK, Olmstead EM, et al. A regional intervention to improve the hospital mortality associated with coronary artery bypass graft surgery. JAMA 1996;275:841–6.
17. Tu JV, Naylor CD, Steering Committee Provincial Adult Cardiac Care Network Ontario. Coronary artery bypass mortality rates in Ontario: a Canadian approach to quality assurance in cardiac surgery. Circulation 1996;94:2429–33.
18. Shroyer LW, Edwards FH, Grover FL. Updates to the data quality review program: the Society of Thoracic Surgeons adult cardiac national database. Ann Thorac Surg 1998;65:1494–7.
19. Leape LL, Hilborne LH, Schwartz JS, et al. The appropriateness of coronary artery bypass graft surgery in academic medical centers. Ann Intern Med 1996;125:8–18.
20. Holman WL, Athanasuleas CL, Allman RM, Sherrill RG, for the Alabama Quality Assurance Foundation CABG Cooperative Project. Alabama CABG Cooperative Project: baseline data. Ann Thorac Surg. In Press.
21. Parsonnet V, Dean D, Bernstein AD. A method of uniform stratification of risk for evaluating the results of surgery in acquired adult heart disease. Circulation 1989;79 Suppl I:I-3–I-12.
22. O'Connor GT, Plume SK, Olmstead EM, et al. Multivariate prediction of in-hospital mortality associated with coronary artery bypass graft surgery. Circulation 1992;85:2110–8.

23. Higgins TL, Estafanous FG, Loop FD, Beck GJ, Blum JM, Paranandi L. Stratification of morbidity and mortality outcome by preoperative risk factors in coronary artery bypass patients. JAMA 1992;267:2344–8.
24. Nashef SAM, Carey F, Silcock MM, Oomen PK, Levy RD, Jones MT. Risk stratification for open heart surgery: trial of the Parsonnet system in a British hospital. BMJ 1992;305:1066–7.
25. Junod FL, Harlan BJ, Payne J, et al. Preoperative risk assessment in cardiac surgery: comparison of predicted and observed results. Ann Thorac Surg 1987;43:59–64.
26. Orr RK, Maini BS, Sottile FD, Dumas EM, O'Mara P. A comparison of four severity-adjusted models to predict mortality after coronary artery bypass graft surgery. Arch Surg 1995;130:301–6.
27. Landon B, Iezzoni LI, Ash AS, et al. Judging hospitals by severity-adjusted mortality rates: the case of CABG surgery. Inquiry 1996;33: 155–66.
28. Iezzoni LI, Ash AS, Schwartz M, Landon B, Mackiernan YD. Predicting in-hospital deaths from coronary artery bypass graft surgery: do different severity measures give different predictions? Med Care 1998;36:28–39.
29. Kassirer JP. The use and abuse of practice profiles. N Engl J Med 1994;330:634–6.
30. Schneider EC, Epstein AM. Influence of cardiac-surgery performance reports on referral practices and access to care. N Engl J Med 1996;335:251–6.
31. Green J, Wintfeld N. Report cards on cardiac surgeons: assessing New York State's approach. N Engl J Med 1995;332:1229–32.
32. Jencks SF, Wilensky GR. The health care quality improvement initiative: a new approach to quality assurance in medicare. JAMA 1992;268:900–3.
33. Vogel RA. HCFA's cooperative cardiovascular project: a nationwide quality assessment of acute myocardial infarction. Clin Cardiol 1994; 17:354–6.
34. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1983;143: 29–36.
35. Harrell FE Jr, Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic prediction. Stat Med 1984;3:142–52.
36. Lemeshow S, Hosmer DW Jr. A review of goodness of fit statistics for use in the development of logistic regression models. Am J Epidemiol 1982;115:92–106.
37. DeLong ER, Peterson ED, DeLong DM, Muhlbaier LH, Hackett S, Mark DB. Comparing risk-adjustment methods for provider profiling. Stat Med 1997;16:2645–64.
38. Efron B, Morris C. Stein's paradox in statistics. Sci Am 1977;236: 119–27.
39. Topol EJ, Califf RM. Scorecard cardiovascular medicine: its impact and future directions. Ann Intern Med 1994;120:65–70.
40. Cooley DA. Building shelters: safeguards in public disclosure of outcomes data. Circulation 1996;93:1–3.
41. Califf RM, Jollis JG, Peterson ED. Operator specific outcomes: a call for professional responsibility. Circulation 1996;93:403–6.
42. Salem-Schatz S, Moore G, Rucker M, Pearson SD. The case for case-mix adjustment in practice profiling: when good apples look bad. JAMA 1994;272:871–4.
43. Lorence D. Benchmarking quality under US health care reform: the next generation. Qual Prog 1994;27:103–7.
44. Kiefe C, Wooley TW, Allison JJ, Box JB, Craig AS. Determining benchmarks: a data-driven search for the best achievable performance. Clin Performance Qual Health Care 1994;2:190–4.
45. Peterson ED, DeLong ER, Jollis JG, Muhlbaier LH, Mark DB. The effects of New York's bypass surgery provider profiling on access to care and patient outcomes in the elderly. J Am Coll Cardiol 1998;32: 993–9.