Risk-Adjusting Hospital Inpatient Mortality Using Automated Inpatient, Outpatient, and Laboratory Databases

Gabriel J. Escobar, MD,† John D. Greene, MA,* Peter Scheirer, MA,*§ Marla N. Gardner, BA,* David Draper, PhD,‡ and Patricia Kipnis, PhD*§*

Objectives: To develop a risk-adjustment methodology that maximizes the use of automated physiology and diagnosis data from the time period preceding hospitalization.

Design: Retrospective cohort study using split-validation and logistic regression.

Setting: Seventeen hospitals in a large integrated health care delivery system.

Subjects: Patients (n = 259,699) hospitalized between January 2002 and June 2005.

Main Outcome Measures: Inpatient and 30-day mortality.

Results: Inpatient mortality was 3.50%; 30-day mortality was 4.06%. We tested logistic regression models in a randomly chosen derivation dataset consisting of 50% of the records and applied their coefficients to the validation dataset. The final model included sex, age, admission type, admission diagnosis, a *L*aboratory-based *A*cute *P*hysiology *S*core (LAPS), and a *CO*morbidity *P*oint *S*core (COPS). The LAPS integrates information from 14 laboratory tests obtained in the 24 hours preceding hospitalization into a single continuous variable. Using Diagnostic Cost Groups software, we categorized patients as having up to 40 different comorbidities based on outpatient and inpatient data from the 12 months preceding hospitalization. The COPS integrates information regarding these 41 comorbidities into a single continuous variable. Our best model for inpatient mortality had a c statistic of 0.88 in the validation dataset, whereas the c statistic for 30-day mortality was 0.86; both models had excellent calibration. Physiologic data accounted for a substantial proportion of the model's predictive ability.

Conclusion: Efforts to support improvement of hospital outcomes can take advantage of risk-adjustment methods based on automated physiology and diagnosis data that are not confounded by information obtained after hospital admission.

Key Words: mortality, inpatient, risk-adjustment, severity of illness, Diagnostic Cost Groups

(*Med Care* 2008;46: 232–239)

Measurement of risk-adjusted mortality among hospitalized patients is exceedingly important in an era in which government, private purchasers, and the general public have intense interest in health care quality and safety. Because of the large sample size required to show mortality differences that are both clinically as well as statistically significant,[1,2] caution must be exercised in making inferences about quality when comparing hospitals using any methodology. Nonetheless, risk-adjusted comparisons can be helpful as a screening tool that can permit groups of hospitals to begin examining practice differences. Given the high cost of patient interviews or data extracted manually from chart review, many risk-adjustment models currently in use are based on 2 types of widely available data: (1) administrative data such as International Classification of Diseases (ICD) diagnosis codes and hospital admission and discharge times,[3] and (2) indicators of factors possibly related to the process of care that can be found in publicly available databases (eg, the number of doctors or hospital beds available in a given area).[4] One important limitation of models based solely on administrative data is that they include information from events that occurred during the hospitalization. Glance et al have shown that this can give hospitals with poor performance "credit" for their complications and can lead to such hospitals being misclassified as providing better care than they actually do.[5]

Significant advances have been made in the incorporation of physiologic data into statistical models for predicting and risk-adjusting outcomes among ill newborns,[6] children,[7,8] and adults.[9,10] Moreover, a recent study from the Veterans Administration hospital system in the United States has demonstrated the feasibility of including physiologic data from automated databases in severity of illness adjustment among adults admitted to multiple intensive care units.[11] Pine et al recently dem-

From the *Kaiser Permanente Division of Research, Systems Research Initiative and Perinatal Research Unit, Oakland, California; †Kaiser Permanente Medical Center, Department of Inpatient Pediatrics, Walnut Creek, California; ‡Department of Applied Mathematics and Statistics, Baskin School of Engineering, University of California, Santa Cruz, California; and §Kaiser Foundation Health Plan Management, Information and Analysis, Oakland, California.

Supported by The Permanente Medical Group, Inc., Kaiser Foundation Hospitals, Inc., and the Sidney Garfield Memorial Fund.

None of the funders were involved in the decision to submit this manuscript or in its preparation.

Reprints: Gabriel J. Escobar, MD, Kaiser Permanente Division of Research, Systems Research Initiative and Perinatal Research Unit, 2000 Broadway, 2nd floor, 021R10, Oakland, CA 94612. E-mail: gabriel.escobar@kp.org.

ISSN: 0025-7079/08/4603-0232

onstrated the value of adding present on admission (POA) codes and laboratory test results from the first 24 hours in the hospital for comparing inpatient mortality for a limited number of conditions.[12] In organizations with integrated care delivery and information systems, it is highly desirable to incorporate these methods to take maximal advantage of available automated patient data for internal benchmarking, quality assurance, quality improvement, and research. As increasingly sophisticated automated medical record systems start going online, it is also important to begin developing methods that can simulate what could be done with such systems.

In this manuscript, we describe the development and validation of a statistical model that estimates the probability of inpatient and 30-day mortality in an integrated health care delivery system in the United States, the Northern California Kaiser Permanente Medical Care Program (KPMCP). The model we describe uses automated data from the time period preceding an individual patient's hospitalization. Beginning in 2007, internal interhospital mortality comparisons and benchmarking within the KPMCP are being performed using the methods described in this report. These methods are also being used for the performance of stratified random sampling of hospital charts for quality improvement efforts and for multivariable template matching[13] for internal case-control studies.

MATERIALS AND METHODS

This project was approved by the Northern California KPMCP Institutional Review Board for the Protection of Human Subjects, which has jurisdiction over all study hospitals.

The Northern California KPMCP serves a total population of approximately 3.3 million members. Under a mutual exclusivity arrangement, physicians of The Permanente Medical Group, Inc., care for Kaiser Foundation Health Plan, Inc. members at facilities owned by Kaiser Foundation Hospitals, Inc. All 17 Northern California KPMCP hospitals and 44 outpatient clinics use the same information systems with a common medical record number and can track care covered by the plan but delivered elsewhere.

Our setting consisted of 17 hospitals whose characteristics are summarized in Table 1. Our study population consisted of all patients admitted to these hospitals who met these entry criteria: (1) hospitalization began from January 1, 2002 through July 31, 2003 and from October 1, 2003 through June 30, 2005; (2) initial hospitalization occurred at a Northern California KPMCP hospital (ie, if a hospitalization involved interhospital transfer, the first hospital stay occurred within the Northern California KPMCP); (3) age ≥15 years at the time of admission; and (4) hospitalization was not for childbirth. Comorbidity data were unavailable for 25,150 linked hospitalizations that occurred between August 1, 2003 and September 30, 2003. These 25,150 hospitalizations were similar to the rest of the cohort [the Appendix (available on the *Medical Care* website, www.lww-medicalcare.com) provides a formal comparison].

We defined a "linked hospitalization" as the time period that began with a patient's admission to the hospital and ended with the patient's discharge (home, to a nursing home or similar setting, or death). A linked hospitalization can thus involve more than 1 hospital stay and could include a patient transfer from one hospital to another before definitive discharge. For linked hospitalizations, mortality was attributed to the admitting KPMCP hospital (ie, if a patient was admitted to hospital A, transferred to B, and died at hospital B, mortality was attributed to hospital A).

TABLE 1. Characteristics of the 17 Study Hospitals*

	No. Medical/Surgical Beds	
	<200	**≥200**
Number	9	8
Total no. acute care beds	1186	2104
Average no. overnight hospitalizations per year per hospital	4655	8138
Total no. overnight hospitalization per year for group	41,895	65,106
Average no. total hospital admissions per year per hospital	6259	18,687
Total no. hospital admissions per year	56,328	93,434
Total no. ICU beds for group	127	186
Average no. ICU admits per year per hospital	753	1168
Total no. ICU admits per year for group	6777	9344

*Numbers are based on totals for the time period 2000–2004.

The principal dependent variable for our primary analyses was in-hospital mortality, which could be ascertained for the entire cohort. In addition, we also used 30-day mortality as the dependent variable for hospitalizations that began before December 1, 2004. For these hospitalizations, we could ascertain 30-day mortality based on the combination of KPMCP patient demographic databases, publicly available lists of deceased patients provided by the Social Security Administration, and linkage to the State of California Death Certificate tapes.

We used the following independent variables as predictors in our models: sex, age at the time of admission, admission type (emergency surgical and nonsurgical, elective surgical and nonsurgical), admission diagnosis, a *L*aboratory-based *A*cute *P*hysiology *S*core (LAPS), and a *CO*morbidity *P*oint *S*core (COPS).

Upon admission to a KPMCP hospital, the admitting physician must enter an admission diagnosis based on clinical judgment at the time the order to hospitalize a patient is issued. Professional medical record coders assign an ICD code to this diagnosis, which may differ from the diagnosis assigned by the physician at the time of final discharge. We grouped all possible 16,090 ICD admission diagnosis codes into 44 broad diagnostic categories, which we refer to as *Primary Conditions*. The ICD codes comprising these 44 Primary Conditions were grouped based on biologic plausibility (ie, relative similarity from a disease standpoint) and on the observed mortality rate because, for modeling purposes, it was desirable to have patient groupings with at least 40 deaths in the derivation dataset.

Using methods described elsewhere, we downloaded laboratory test results and linked them to each hospitalization

record.[14–16] We used the following test results to calculate the LAPS: serum albumin; anion gap; arterial pH, $PaCO_2$, and PaO_2; bicarbonate; total serum bilirubin; blood urea nitrogen; serum creatinine; serum glucose; serum sodium; serum troponin I; hematocrit; and total white blood cell count in thousands. Eligible laboratory test results were those obtained in the 24-hour time frame *preceding* hospitalization. We deleted out of range values. For 11 of 14 laboratory tests, we followed the usual severity of illness scoring convention for handling missing data (imputing a normal value for absent test results) and, in those cases where a patient had multiple results for the same test in the 24 hours preceding admission, we selected the most deranged value in the time frame. However, in light of work done by Afessa et al[17] and internal analyses of laboratory data from patients who experienced deterioration leading to late transfer to intensive care, we used a more complex approach for 3 laboratory tests (arterial pH, troponin I, and total white blood cell count). For these tests, we assigned point scores in 2 steps. First, we divided patients into 2 risk groups (provisional mortality risk of $<6\%$ and $\geq 6\%$) using a simple model based on age, admission type, and the 5 most commonly performed laboratory tests. For patients with a provisional mortality risk of $<6\%$, missing arterial pH, troponin I, and total white blood cell count were imputed to normal. However, for patients with a provisional mortality risk of $\geq 6\%$, missing values for these 3 tests were assigned their own risk band. Using the results from these 2 sequential regression models, we integrated these 14 laboratory test results into a single continuous variable with a minimum value of 0 and a theoretical maximum of 256. For example, a hematocrit of $<20\%$ contributes 7 points to the total LAPS, whereas a hematocrit of 40–49% contributes zero points. The Appendix provides additional details on the development and performance of the LAPS.

We used Diagnostic Cost Groups (DxCG)[18–20] software to group diagnoses occurring before the hospitalization. Every month, commercially available DxCG software[21] scans data on the entire Northern California KPMCP membership and assigns patients to up to 184 possible Hierarchical Condition Categories (HCC) based on their inpatient and outpatient diagnoses during the 12-month period preceding each scan. A given patient may have multiple HCC assignments. Based on biologic plausibility and mortality risk, we grouped 147 of these 184 HCCs into 41 comorbidity groups (we did not use 37 HCCs because either they did not have a substantial mortality differential or they applied to pediatric populations). Using logistic regression, we integrated these 41 comorbidity groups into a single continuous variable with a minimum value of 0 and a theoretical maximum of 701. For example, experiencing a head injury in the year preceding hospitalization contributes 11 points to the total COPS, whereas preexisting hypertension contributes 18 points. The Appendix provides additional details on the development and performance of the COPS.

We used split validation, with 50% of the records randomly assigned to the derivation dataset and 50% to the validation dataset. We assessed the ability of the model to correctly distinguish patients who died from survivors using the area under the receiver operator characteristic curve, or c statistic.[22,23] We assessed the ability of the model to accurately predict mortality across all ranges of risk by comparing predicted and observed mortality rates in: (1) subsets of patients based on 10% increments in predicted mortality, and (2) predicted mortality risk deciles. Given the large sample size, we felt that this was more meaningful than only calculating the Hosmer and Lemeshow χ^2 goodness-of-fit test.[24] After defining our best model using the derivation dataset, we tested its discrimination and calibration by applying the coefficient estimates obtained from the derivation model to the validation dataset.

We used logistic regression to predict inpatient mortality as a function of sex, age, admission type, Primary Condition (which in some cases was further subdivided), LAPS, and COPS. We obtained the best model performance when estimating inpatient mortality for each of the 44 Primary Conditions with a separate logistic regression and pooling these results across the entire cohort. Because none of the hospitalizations with a Primary Condition related to pregnancy complications resulted in a death, the estimated probability of mortality for those hospitalizations was set to zero. The final 43 logistic regression models used the following predictors: sex, splined age,[2] admission type, Primary Condition subgroups, LAPS, COPS, and interaction terms (age^2 $\times$ LAPS, age^2 $\times$ COPS, LAPS $\times$ COPS). We expanded age^2 to 3 nonlinear terms using a restricted cubic spline based on 4 knots chosen at the 5th, 35th, 65th, and 95th quantiles.[22]

Following Knaus et al[25] and Render et al,[11] for each of the 43 models we calculated the differences between the log likelihood of the full model and the log likelihood of a model without each of the 6 main predictors. The relative contribution of a predictor was defined as the ratio of its log likelihood difference to the sum of the 6 log likelihood differences multiplied by 100.

Because the purpose of our model was to obtain point estimates of the probability of mortality for a given hospitalization (as opposed to making inferences or calculations regarding the error associated with each estimate), we included all hospitalizations experienced by the patients in our cohort, ignoring within-patient clustering (random) effects. Point estimates are known to be insensitive to misspecification of the random effects distribution.[26] In any case, given the rich set of covariates available to us, it seems reasonable to assume that hospitalizations experienced by an individual are approximately conditionally independent given the detailed information available at the time of admission. As a sensitivity analysis, we tested the model 100 times, each time randomly selecting a single hospitalization per patient for inclusion in the analysis.

We performed these additional sensitivity analyses to assess the stability of our model under varying conditions of data quality: (1) introducing random variation (eg, changing a hospitalization with a low risk Primary Condition, such as "Chest Pain," to a high risk one such as "Pneumonia") in 5%, 10%, 15%, and 25% of the Primary Conditions; (2) replacing the ICD code in the Primary Condition with the one used to assign the principal diagnosis at the time of hospital dis-

charge; (3) introducing random variation in 5%, 10%, 15%, and 25% of the LAPS values; and (4) introducing random variation in 5%, 10%, 15%, and 25% of the COPS values.

RESULTS

During the study period, a total of 494,504 individual hospital stays involving 282,050 patients occurred at these 17 hospitals. After concatenation of interhospital transfers, we were left with 457,793 linked hospitalizations. We removed 25,150 linked hospitalization records that occurred between August 1, 2003 and September 30, 2003, because they lacked COPS data. Of the remaining 432,643 hospitalizations, 22,994 were deleted because they began at a non-KPMCP hospital (ie, they were transfers in), and 344 records were deleted because of missing data, leaving a total of 409,305 hospital stays involving 259,699 patients. These were split randomly into a derivation and a validation dataset. The derivation dataset consisted of 204,996 hospitalizations and 7102 (3.46%) deaths, whereas the validation dataset consisted of 204,309 hospitalizations and 7224 (3.53%) deaths. Table 2 summarizes the distribution of patients in the derivation, validation, and entire cohort datasets. In both the derivation and validation datasets, 79.1% of the patients experienced a single hospitalization, 14.5% experienced 2, and 6.4% had 3 or more.

Table 3 summarizes the performance characteristics of the final risk-adjustment model to predict inpatient mortality, which had excellent discrimination, with a c statistic of 0.88 in all datasets. Figures 1 and 2 show that this model is well calibrated except among patients with an extremely high (≥60%) mortality risk, where stochastic variation is present because of small sample size.[27,28] Table 3 also shows the important role played by acute physiology, where the median relative contribution of the LAPS across the 43 Primary Condition models was 56%. The c statistic for a model based on 30-day mortality was 0.86.

Our model performed well across multiple subsets. The c statistics were all >0.80 for men, women, different age groups, and ≥0.85 for all 17 individual hospital populations. They were also ≥0.80 for 29 of the 44 Primary Conditions and between 0.71 and 0.80 for 13 of them. The c statistics for 2 Primary Conditions with small numbers were relatively

TABLE 2. Patient Characteristics

	Derivation Cohort	Validation Cohort	Entire Cohort
No. patients/No. hospitalizations	155,474/204,996	154,767/204,309	259,699/409,305
Mortality rate (%)			
In-hospital	3.5	3.5	
30-day	4.2	3.9	
% Male	45.3	45.2	45.2
Age (median, mean ± SD)	65.0, 62.6 ± 18.4	65.0, 62.6 ± 18.4	65.0, 62.6 ± 18.4
% ≥65 yr	51.1	51.3	51.2
LAPS* (median, mean ± SD)	13.0, 22.3 ± 18.2	13.0, 22.4 ± 18.3	13.0, 22.4 ± 18.3
COPS† (median, mean ± SD)	59.0, 71.8 ± 52.9	62.0, 76.5 ± 58.0	61.0, 74.1 ± 55.5
% With these Primary Conditions‡			
Pneumonia	4.7	4.6	4.6
Sepsis	1.4	1.4	1.4
Catastrophic conditions	1.4	1.4	1.4
Gastrointestinal bleeding	7.9	8.0	7.9
Hip fracture	1.5	1.6	1.6
Any malignancy	3.4	3.5	3.4
Race (% of patients)			
White	65.9	65.9	65.9
African American	9.7	9.7	9.7
Asian	8.1	8.1	8.1
Hispanic	10.0	10.0	10.0
Other	6.2	6.2	6.2

*Laboratory-based *A*cute *P*hysiology *S*core. The LAPS is based on 14 laboratory test results obtained in the 24 hours preceding hospitalization. Increasing degrees of physiologic derangement are reflected in a higher LAPS, which is a continuous variable that can range between a minimum of zero and a theoretical maximum of 256, although <0.05% of patients in our cohort had LAPS exceeding 120 and none had a LAPS >165. For example, a hematocrit of <20% contributes 7 points to the total LAPS, whereas a total white blood cell count of <2000/mL contributes 29 points.

†*CO*morbidity *P*oint *S*core. Based on outpatient and inpatient utilization in the preceding year, a patient can be categorized as having up to 41 different comorbidities. The COPS integrates information regarding these comorbidities into a single continuous variable that has a value that can range between a minimum of zero and a theoretical maximum of 701 (in our dataset, however, patients with COPS values >300 are uncommon). For example, experiencing a head injury in the year preceding hospitalization contributes 11 points to the total COPS, whereas preexisting hypertension contributes 18 points.

‡Primary Conditions are groupings of related International Classification of Diseases codes assigned at the time of admission to the hospital. Catastrophic conditions are selected diagnoses associated with high mortality (eg, ruptured aortic aneurysm); see text for additional details.

TABLE 3. Risk-Adjustment Model Performance

	Derivation Cohort	Validation Cohort	Entire Cohort
c statistic	0.88	0.88	0.88
Relative contribution of predictors*			
Admission category† (%)	7	9	7
Age (%)	16	16	17
Sex (%)	1	1	1
LAPS‡ (%)	56	55	54
COPS§ (%)	8	11	9
Diagnostic subgroup¶ (%)	9	6	6

*See text for details on how relative contribution of predictors was estimated.

†Can be 1 of the following 4: emergency medical, emergency surgical, elective medical, or elective surgical.

‡*L*aboratory-based *A*cute *P*hysiology *S*core. The LAPS is based on 14 laboratory test results obtained in the 24 hours preceding hospitalization. Increasing degrees of physiologic derangement are reflected in a higher LAPS, which is a continuous variable that can range between a minimum of zero and a theoretical maximum of 256, although <0.05% of patients in our cohort had LAPS exceeding 120 and none had a LAPS >165. For example, a hematocrit of <20% contributes 7 points to the total LAPS, whereas a total white blood cell count of <2000/mL contributes 29 points.

§*CO*morbidity *P*oint *S*core. Based on outpatient and inpatient utilization in the preceding year, a patient can be categorized as having up to 41 different comorbidities. The COPS integrates information regarding these comorbidities into a single continuous variable that has a value that can range between a minimum of zero and a theoretical maximum of 701 (in our dataset, however, patients with COPS values >300 are uncommon). For example, experiencing a head injury in the year preceding hospitalization contributes 11 points to the total COPS, whereas preexisting hypertension contributes 18 points.

¶Some Primary Conditions included subgroups within the set of diagnoses.

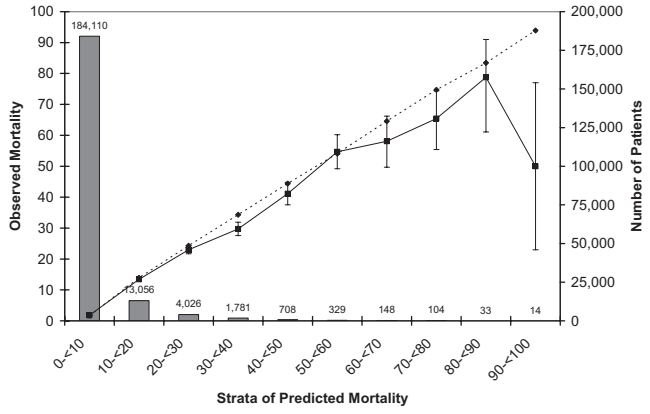

FIGURE 1. Calibration curve for risk-adjustment model in the validation dataset. The horizontal axis shows 10 predicted mortality ranges. The total number of patients in these ranges is shown by bar graphs with its vertical axis scale on the right, whereas the actual mortality, with its associated 95% confidence interval, is shown by dark squares, with its vertical axis scale on the left. The dashed line shows what would be seen if the model were perfectly calibrated. Note that the greatest deviation between predicted and observed mortality occurs among hospitalizations with predicted mortality rates ≥60%, whose numbers are very small (299 hospitalizations, <0.2% of the cohort).

low: pericarditis and valvular heart disease (c = 0.63, N = 562 in validation dataset) and miscellaneous cardiac conditions and congenital heart disease (c = 0.66, N = 987). The Appendix provides additional details, including Hosmer and Lemeshow statistics.

Our final model showed substantial stability in sensitivity analyses. Both the mean and median c statistics for the 100 analyses restricted to a single hospitalization per patient were 0.90. Introducing random variation in the Primary Conditions changed the c statistic to 0.87 when 5% of the records had a change in Primary Condition and to 0.86 when 25% of the records were so altered. Replacing the ICD code used to assign patients' Primary Conditions from admission diagnosis to final (principal discharge) diagnosis led to a slightly higher c statistic (0.89). Random changes in LAPS and COPS produced virtually no change in the c statistic, even in models where we introduced fluctuation in proportions as high as 25% of the records. Calibration curves from these alternative models were similar to those shown in Figures 1 and 2.

DISCUSSION

Perfect measures for quality of care do not exist. Although they are commonly used in the literature and in policy discourse, inpatient and 30-day mortality have important statistical and sample size limitations, which have been discussed elsewhere.[1,2] However, there is value in using mortality rates as part of a quality improvement process that motivates collaboration between hospitals and as a starting point for further investigation of differences between hospitals. In this process, risk-adjustment that has credibility with clinicians is an essential component.

We found that it is possible to perform rigorous risk-adjustment for inpatient mortality by enhancing administrative data with preadmission automated physiology and diagnosis data. Using a very large hospitalization cohort, we developed a risk-adjustment methodology with excellent discrimination and calibration. It expands upon recent work done by Render et al,[11] Zimmerman et al,[10] Glance et al[5] and Pine et al,[12] who used similar methods to risk-adjust outcomes in a more restricted population (patients admitted to the intensive care unit or with a limited set of conditions).

Several aspects of our approach merit further comment. First, unlike other risk-adjustment approaches reported in the literature, our methodology does not use any data from the time period after a physician's order to admit a patient. For example, a patient may be admitted to the hospital with an admission diagnosis of dehydration but, upon hydration, be found to actually have pneumonia. Risk-adjustment based on the final diagnosis (pneumonia) in effect assumes that all physicians would be equally careful in reassessing patients after hydration, which may not, in fact, be the case. This consideration is important in quality of care studies that provisionally identify unusual hospital-level outcomes by comparing observed (O) and expected (E) mortality rates,[29–31] because in observational studies of this type[32] it is crucial not to permit hospital processes of care to confound the indirect estimation of quality provided by

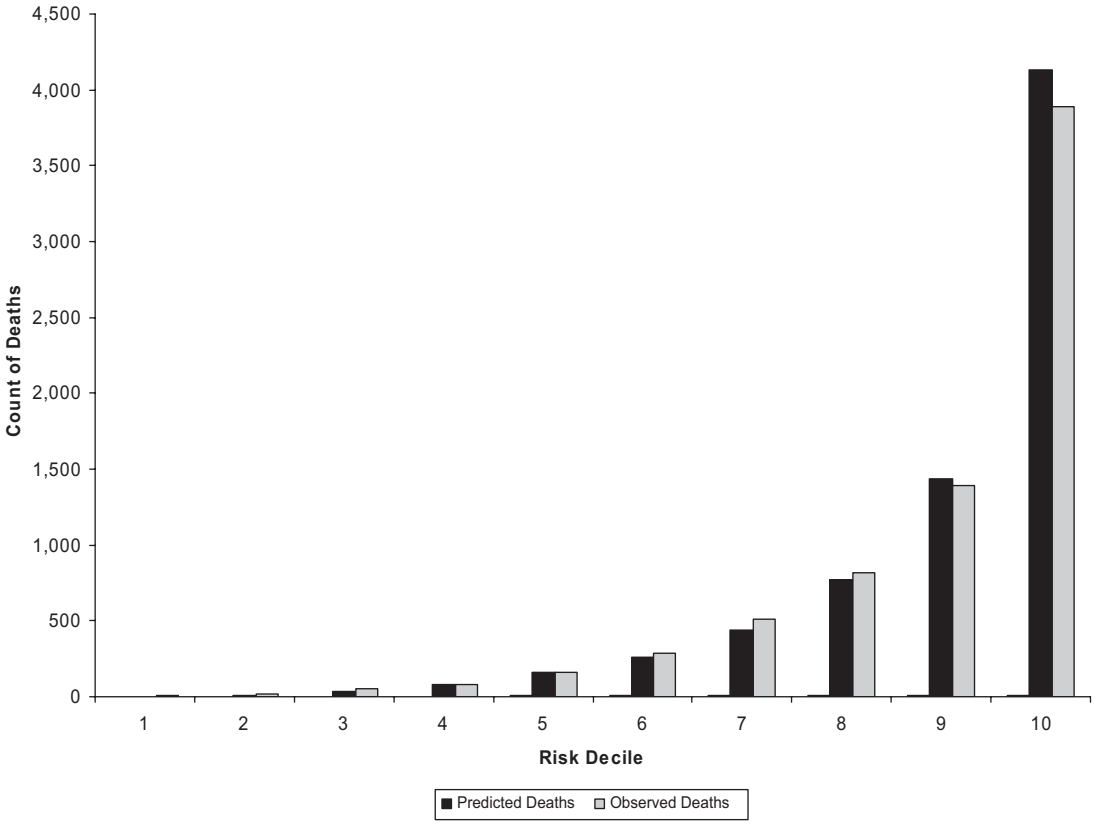

FIGURE 2. Predicted and observed deaths in the validation dataset, by risk decile. The horizontal axis shows predicted mortality by risk decile, whereas the vertical axis shows the predicted and observed number of deaths in each decile.

examining the residual R = O − E, as has been documented by Glance et al.[5]

Second, our methodology takes maximal advantage of outpatient laboratory, diagnosis, and patients' previous hospitalization data. DxCG HCCs are based on a patient's outpatient and inpatient diagnoses in the previous year,[18–20] offering an advantage over the use of coded comorbidities based on ICD codes assigned on or after the time of discharge. Using these is problematic because it is not always possible to distinguish between those comorbidities that preceded the hospitalization and those that developed during the hospitalization (ie, complications during the stay). Our results thus provide added support to the use of POA coding for comorbidities in risk-adjustment, an issue receiving increasing policy attention.[33,34]

Third, our methodology incorporates physiologic data. In addition to their statistical contribution to model performance, physiologic data are important because including them might enhance credibility with clinicians, many of whom are familiar with severity of illness scores, which, until now, have been largely confined to populations receiving intensive care.

It is important to highlight the limitations of our approach. The first is that it involves data from an organization with highly integrated information systems permitting multiple record linkages across the continuum of care. Many hospitals cannot yet take advantage of all the methods described here. However, the methods we used could be modified for use by hospital systems in countries such as Great Britain and Canada, and entities such as the Veterans Administration Hospital System in the United States. Moreover, as more and more patient data become available electronically, increasing numbers of hospitals will be able to use these data elements for purposes other than direct patient care. For example, the new automated inpatient medical record being deployed in the KPMCP includes a comorbidity section (Physician's Problem List) that is completed by the admitting physician at the time of admission. The automated record transforms the physician's text entry into ICD codes—thus, in the future, in addition to using DxCG HCCs, a score similar to the COPS could be generated from the admission history and physical.

A second important limitation of our methods, which is shared by other risk-adjustment methodologies such as that of Render et al and the APACHE IV (Acute Physiology and Chronic Health Examination), is that the modeling approach we use, which includes the use of cubic splines and interaction terms, no longer has the intuitive simplicity of severity of illness scores such as the APACHE I,[35] PRISM (Pediatric Risk of Mortality),[36] and SNAP-II (Score for Neonatal Acute Physiology, version II).[6] These scores can be assigned manually, and it is easy for a clinician to understand the degree to which physiologic derangement or presence of comorbidities contributes to an elevated mortality risk. However, we think

that most clinicians would prefer risk-adjustment approaches with superior performance to those that are more intuitive but which have inferior discrimination or calibration. Given improved model performance, clinicians are more likely to accept models that take advantage of increasing computing power and data availability.

Not all investigators would necessarily agree with our approach of how to collapse thousands of ICD codes into a more manageable set of Primary Conditions, and other ways of doing this are certainly possible. Our model performed well for most Primary Conditions. However, it may not perform as well with some conditions because it does not include vital signs, which contribute significantly to the performance of the APACHE IV, and because the degree of illness severity of some illnesses (eg, pericarditis, valvular heart disease) may not be reflected in derangements among the 14 laboratory tests we included. Moreover, we did not have access to data on patients' functional status, which would be expected to significantly enhance the predictive ability of any model. Some investigators would dispense with diagnoses altogether. Graham et al used a variant of the APACHE III in which, instead of using all 78 Disease Categories (Admit Diagnoses), they collapsed principal diagnoses into 3 categories—"high," "medium," and "low" risk. They reported that their simplified model had a c-statistic of 0.87.[37] Hucker et al reported on a model for predicting in-hospital mortality for patients admitted through the emergency department using only 4 measurements: age, heart rate, phosphate, and albumin; this parsimonious model had a c-statistic of 0.82.[38]

Some investigators could also object to our validation approach. Several validation approaches are possible, including validation at the hospitalization level (the method used in this article) and at the facility level. Some information about the latter is available in the Appendix, in which it can be seen that the facility-level c-statistics in our validation sample ranged from 0.85 to 0.89.

Considering these studies in the light of our findings suggests the following. First, it is very likely that other hospitals and hospital chains will be able to generate large datasets that include laboratory data. Consequently, in the coming years, the ability to generate some sort of physiology-based severity of illness data will become more widespread among hospitals in developed nations. This study also suggests that the value of administrative data sources cannot be ignored even as automated clinical data become more commonly available. In addition, hospitals with an automated medical record could define a variant of our COPS using text searches for the comorbidities listed by the admitting physician. Thus, with respect to use of these data for risk-adjustment purposes, the important question will not be "Should one employ automated physiology-based risk-adjustment?" or "Should we include comorbidity data from outpatient records in hospital risk-adjustment models?" but, rather, "Is it possible to convince different institutions to standardize data definitions so as to permit larger collaborative studies?"

Lastly, if the goal is to develop the fullest possible understanding of reasons for variation in outcomes (rather than for risk-adjusted comparison of facilities), use of our methodology need not exclude the use of hierarchical variables found in publicly available databases (eg, the number of hospital beds or nurses at a given hospital[39]). Indeed, in settings with a larger number of hospitals with automated medical records containing historical comorbidity data as well as physiologic measurements, combining our methodology with other approaches, such as quantitative assessment of process measures,[40] could provide greater insights in identifying the factors that contribute most to variation in the outcomes of hospital care.

ACKNOWLEDGMENTS

The authors thank Drs. Joseph V. Selby, Paul Feigenbaum, Philip Madvig, and Maria Massolo for their administrative support and for reviewing the manuscript, and Drs. Paul Feigenbaum and Theodore Levin for their assistance in grouping ICD codes. The study would not have been possible without the assistance and administrative support the authors received from Dr. Jed Weissberg, Mr. Edward Thomas, and Mr. Marcus Lee. The authors also thank Ms. Kimberley Harris for formatting the manuscript.

REFERENCES

1. Hofer TP, Hayward RA. Identifying poor-quality hospitals. Can hospital mortality rates detect quality problems for medical diagnoses? *Med Care*. 1996;34:737–753.
2. Dimick JB, Welch HG, Birkmeyer JD. Surgical mortality as an indicator of hospital quality: the problem with small sample size. *JAMA*. 2004; 292:847–851.
3. Iezzoni LI. Coded data from administrative sources. In: Iezzoni LI, ed. *Risk Adjustment for Measuring Health Care Outcomes*. Chicago, IL: Health Administration Press; 2003: Chapter 5, 83–138.
4. Jarman B, Gault S, Alves B, et al. Explaining differences in English hospital death rates using routinely collected data. *BMJ*. 1999;318:1515–1520.
5. Glance LG, Dick AW, Osler TM, et al. Accuracy of hospital report cards based on administrative data. *Health Serv Res*. 2006;41:1413–1437.
6. Richardson D, Corcoran J, Escobar G, et al. SNAP-II and SNAPPE-II: simplified newborn illness severity and mortality risk scores. *J Pediatr*. 2001;138:92–100.
7. Pollack MM, Patel KM, Ruttimann UE. PRISM III: an updated Pediatric Risk of Mortality Score. *Crit Care Med*. 1996;24:743–752.
8. Pediatric intensive care unit evaluations using PRISM severity score. Available at: http://www.dcchildrens.com/picues/about.aspx. Accessed March 22, 2007.
9. McMahon LF Jr, Hayward RA, Bernard AM, et al. APACHE-L: a new severity of illness adjuster for inpatient medical care. *Med Care*. 1992; 30:445–452.
10. Zimmerman JE, Kramer AA, McNair DS, et al. Acute Physiology and Chronic Health Evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients. *Crit Care Med*. 2006;34:1297–1310.
11. Render ML, Kim HM, Welsh DE, et al. Automated intensive care unit risk adjustment: results from a National Veterans Affairs study. *Crit Care Med*. 2003;31:1638–1646.
12. Pine M, Jordan HS, Elixhauser A, et al. Enhancement of claims data to improve risk adjustment of hospital mortality. *JAMA*. 2007;297:71–76.
13. Silber JH, Rosenbaum PR, Trudeau ME, et al. Multivariate matching and bias reduction in the surgical outcomes study. *Med Care*. 2001;39: 1048–1064.
14. Go AS, Hylek EM, Chang Y, et al. Anticoagulation therapy for stroke prevention in atrial fibrillation: how well do randomized trials translate into clinical practice? *JAMA*. 2003;290:2685–2692.
15. Selby JV. Linking automated databases for research in managed care settings. *Ann Intern Med*. 1997;127:719–724.
16. Escobar GJ, Greene J, Hulac P, et al. Rehospitalization after birth

hospitalization: patterns among infants of all gestations. *Arch Dis Child*. 2005;90:125–131.
17. Afessa B, Keegan MT, Gajic O, et al. The influence of missing components of the Acute Physiology Score of APACHE III on the measurement of ICU performance. *Inten Care Med*. 2005;31:1537–1543.
18. Ellis RP, Ash A. Refinements to the Diagnostic Cost Group (DCG) model. *Inquiry*. 1995;32:418–429.
19. Ash AS, Ellis RP, Pope GC, et al. Using diagnoses to describe populations and predict costs. *Health Care Financ Rev*. 2000;21:7–28.
20. DxCG Inc. Available at: www.dxcg.com. Accessed November 17, 2006.
21. DxCG Inc. *Risk Smart™ Models and Methodologies Guide*. Boston, MA: DxCG Inc.; 2002.
22. Harrell FE Jr, Lee KL, Califf RM, et al. Regression modelling strategies for improved prognostic prediction. *Stat Med*. 1984;3:143–152.
23. Iezzoni AA, Shwartz M. Risk adjustment for measuring health care outcomes. In: Iezzoni LI, ed. *Risk Adjustment for Measuring Health Care Outcomes*. 2nd ed. Chicago, IL: Health Administration Press; 1997: Chapter 9.
24. 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.
25. Knaus WA, Wagner DP, Draper EA, et al. The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. *Chest*. 1991;100:1619–1636.
26. Butler SM. Random effects models with nonparametric priors. *Stat Med*. 1992;11:1981–2000.
27. Snedecor G, Cochran W. *Statistical Methods*. 8th ed. Ames, IA: Iowa State University Press; 1989.
28. Zhu BP, Lemeshow S, Hosmer DW, et al. Factors affecting the performance of the models in the Mortality Probability Model II system and strategies of customization: a simulation study. *Crit Care Med*. 1996; 24:57–63.
29. Jencks SF, Daley J, Draper D, et al. Interpreting hospital mortality data. The role of clinical risk adjustment. *JAMA*. 1988;260:3611–3616.
30. Normand S, Glickman M, Gatsonis C. Statistical methods for profiling providers of medical care: issues and applications. *J Am Stat Assoc*. 1997;92:803–814.
31. Zheng H, Yucel R, Ayanian JZ, et al. Profiling providers on use of adjuvant chemotherapy by combining cancer registry and medical record data. *Med Care*. 2006;44:1–7.
32. Rosenbaum P. *Observational Studies*. 2nd ed. New York, NY: Springer; 2002.
33. Iezzoni LI. Finally present on admission but needs attention. *Med Care*. 2007;45:280–282.
34. Zhan C, Elixhauser A, Friedman B, et al. Modifying DRG-PPS to include only diagnoses present on admission: financial implications and challenges. *Med Care*. 2007;45:288–291.
35. Knaus WA, Zimmerman JE, Wagner DP, et al. APACHE-acute physiology and chronic health evaluation: a physiologically based classification system. *Crit Care Med*. 1981;9:591–597.
36. Pollack MM, Ruttimann UE, Getson PR. Pediatric risk of mortality (PRISM) score. *Crit Care Med*. 1988;16:1110–1116.
37. Graham PL, Cook DA. Prediction of risk of death using 30-day outcome: a practical end point for quality auditing in intensive care. *Chest*. 2004;125:1458–1466.
38. Hucker TR, Mitchell GP, Blake LD, et al. Identifying the sick: can biochemical measurements be used to aid decision making on presentation to the accident and emergency department. *Br J Anaesth*. 2005; 94:735–741.
39. Needleman J, Buerhaus PI, Mattke S, et al. Nurse-staffing levels and the quality of care in hospitals. *N Engl J Med*. 2002;346:1715–1722.
40. Mant J, Hicks N. Detecting differences in quality of care: the sensitivity of measures of process and outcome in treating acute myocardial infarction. *BMJ*. 1995;311:793–796.