Impact of Present-on-admission Indicators on Risk-adjusted Hospital Mortality Measurement

Jarrod E. Dalton, Ph.D.,* Laurent G. Glance, M.D.,† Edward J. Mascha, Ph.D.,‡ John Ehrlinger, Ph.D.,§ Nassib Chamoun, M.S.,|| Daniel I. Sessler, M.D.#

ABSTRACT

Background: Benchmarking performance across hospitals requires proper adjustment for differences in baseline patient and procedural risk. Recently, a Risk Stratification Index was developed from Medicare data, which used all diagnosis and procedure codes associated with each stay, but did not distinguish present-on-admission (POA) diagnoses from hospital-acquired diagnoses. We sought to (1) develop and validate a risk index for in-hospital mortality using only POA diagnoses, principal procedures, and secondary procedures occurring before the date of the principal procedure (POARisk) and (2) compare hospital performance metrics obtained using the POARisk model with those obtained using a similarly derived model which ignored the timing of diagnoses and procedures (AllCodeRisk).

Methods: We used the 2004–2009 California State Inpatient Database to develop, calibrate, and prospectively test our models (n = 24 million). Elastic net logistic regression was used to estimate the two risk indices. Agreement in hospital performance under the two respective risk models was assessed by comparing observed-to-expected mortality ratios; acceptable agreement was predefined as the AllCodeRisk-based observed-to-expected ratio within ±20% of the POARisk-based observed-to-expected ratio for more than 95% of hospitals.

Results: After recalibration, goodness of fit (*i.e.,* model calibration) within the 2009 data was excellent for both models. C-statistics were 0.958 and 0.981, respectively, for the POARisk and AllCodeRisk models. The AllCodeRisk-based observed-to-expected ratio was within ±20% of the POARisk-based observed-to-expected ratio for 89% of hospitals, which was slightly lower than the predefined limit of agreement.

Conclusion: Consideration of POA coding meaningfully improved hospital performance measurement. The POARisk model should be used for risk adjustment when POA data are available.

What We Already Know about This Topic

- Comparing hospitals' performance is important for payers, patients, and healthcare organizations
- The authors previously developed a Risk Stratification Index to quantify major variables for comparing performance, individual patient heterogeneity, and procedural risks

What This Article Tells Us That Is New

- In this study, the Risk Stratification Index was modified to incorporate the timing of diagnoses and procedures
- Considering the timing of diagnoses and procedures improved risk adjustment

HOSPITALS increasingly depend on performance measures, as these measurements have come to influence payment for services, patients' selection of providers, and internal quality improvement evaluations.[1,2] Fair comparisons of health outcomes across hospitals require proper adjustment for heterogeneity among providers in terms of patient and procedural risk. Consequently, risk-adjustment models abound,[3–8]—although there is evidence that many of these models are inconsistent in terms of their characterizations of risk,[9] leading to the potential for misclassification of high- and low-quality hospitals.[10,11]

A recently proposed risk adjustment methodology called the Risk Stratification Indices (RSIs) was based on

◆ This article is accompanied by an Editorial View. Please see: Collins G, Le Manach Y: Multivariable risk prediction models: It's all about the performance. ANESTHESIOLOGY 2013; 118:1252–3.

* Senior Biostatistician, Department of Quantitative Health Sciences and Department of Outcomes Research, Cleveland Clinic, Cleveland, Ohio, and Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, Ohio. † Professor, Department of Anesthesiology and Community and Preventive Medicine, University of Rochester School of Medicine, Rochester, New York. ‡ Associate Staff, Department of Quantitative Health Sciences and Department of Outcomes Research, Cleveland Clinic. § Assistant Staff, Department of Quantitative Health Sciences, Cleveland Clinic. || Adjunct Staff, Department of Outcomes Research, Cleveland Clinic. # Michael Cudahy Professor and Chair, Department of Outcomes Research, Cleveland Clinic.

Received from the Departments of Outcomes Research and Quantitative Health Sciences, Cleveland Clinic, Cleveland, Ohio. Submitted for publication August 22, 2012. Accepted for publication February 4, 2013. Support was provided solely from institutional and/or departmental sources.

Address correspondence to Dr. Sessler: Department of Outcomes Research, Cleveland Clinic, 9500 Euclid Avenue, Mail Code P-77, Cleveland, Ohio 44195. ds@or.org. On the world wide web: www.or.org.. Information on purchasing reprints may be found at www.anesthesiology.org or on the masthead page at the beginning of this issue. ANESTHESIOLOGY's articles are made freely accessible to all readers, for personal use only, 6 months from the cover date of the issue.

 Anesthesiology 2013; 118:1298-306

International Classification of Disease, 9th Revision, Clinical Modification (ICD-9-CM) codes. These risk scores were estimated and validated using the nationally representative 2001–2006 Medicare Provider Analysis and Review database. The RSIs are well validated and highly predictive for duration of hospitalization, in-hospital mortality, and 30-day mortality.[3] However, the Medicare Provider Analysis and Review database on which RSI was developed does not distinguish preexisting conditions that were present-on-admission (POA) from complications that occurred during hospitalization. This is potentially a serious limitation because risk for certain patients might be inflated by codes that result from hospital-acquired complications. Hospitals might thus appear to have a higher risk population, whereas in fact at least some fraction of the risk resulted from hospital-acquired complications rather than baseline risk *per se*.[12–14]

Risk prediction models that include POA indicators could improve adjustment of hospital outcomes by eliminating diagnoses representing hospital-acquired complications from risk-adjustment algorithms.[15] With the increasing ubiquity of administrative datasets which incorporate POA information, it appears that the establishment of such models is timely.

Our goals were therefore to (1) develop and prospectively validate a baseline risk index for in-hospital mortality (which we denote as "POARisk") using only POA diagnoses, principal procedures, and secondary procedures occurring exclusively before admission for the principal procedure and (2) assess the degree by which risk-adjusted hospital performance measures vary under a similarly derived risk index that ignores the POA status of diagnoses and timing of procedures (thus including all diagnoses and procedures associated with the stay).

Materials and Methods

Under authorization by the US Agency for Healthcare Research and Quality, we obtained data on 24 million inpatient discharges from the 2004–2009 California State Inpatient Database.** This registry represents a census of discharges occurring within the state. POA indicators are captured for all ICD-9-CM diagnosis codes; likewise, date of procedure (relative to the admission date) is captured for all ICD-9-CM procedure codes.

We used data from 2004 to 2008 to derive our models, whereas the data from 2009 were used to prospectively validate the models (see section Model Performance and Reliability). The 2004–2008 data were further split randomly to facilitate a two-step modeling procedure. Eighty percent of those discharges were used for initial model development and the remaining 20% were used to perform an initial calibration or bias-correction of risk estimates produced by the initial model (we say an initial calibration because we feel that calibration should constantly be addressed whenever the model is applied in external populations; see the "Model Development" and "Calibration" subsections for details).

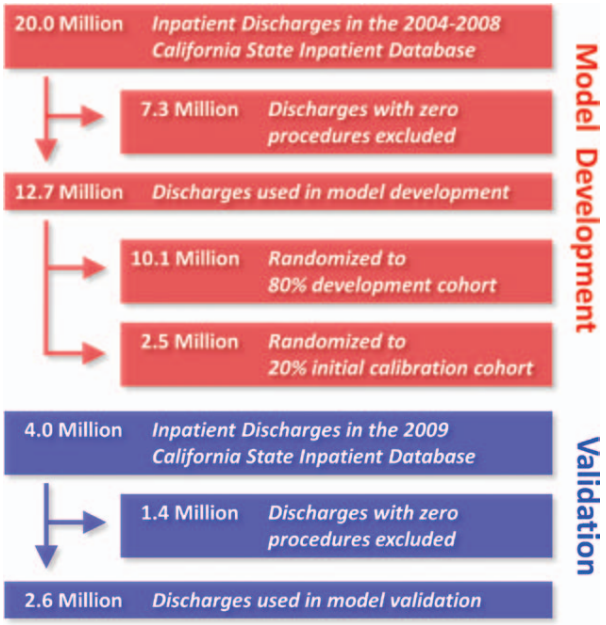

Fig. 1. Study flow diagram.

The only exclusion we made was for patients who did not undergo a procedure; thus, our models sought to characterize in-hospital mortality risk for all inpatients undergoing at least one procedure. A summary of discharges included and excluded, as well as how the included discharges were partitioned for the purposes of our study, is provided in figure 1.

Model Development

To develop the initial POARisk model, we used logistic regression with in-hospital mortality as the dependent variable and a collection of variables derived from the ICD-9-CM diagnosis and procedure codes as predictors. Considered as inputs to our model were diagnosis codes that were POA, the principal procedure code, and any secondary procedure codes for which the date of procedure was prior (but not equal) to the date of the principal procedure. (We felt that in the absence of data on which procedures were planned, these represented a reasonable approximation; see Discussion.) We also used patient age and gender in our model. Age was represented by two predictors—one which estimated risk for infants less than 1 yr old and another linear term for the rest of the patients.

The ICD-9-CM codes are hierarchical in nature. For example, acute myocardial infarctions are coded with diagnosis code 410.XX; the fourth digit further classifies these diagnoses based on the location (*e.g.*, 410.2X refers to the

** HCUP Databases: Healthcare Cost and Utilization Project (HCUP). November 2011. Agency for Healthcare Research and Quality, Rockville, MD. Available at: http://www.hcup-us.ahrq.gov/databases.jsp. Accessed July 23, 2012.

inferolateral wall), and the fifth digit specifies the episode of care. As such, many of the five-digit codes lacked sufficient representation for inclusion in our logistic model: an aggregation routine was used to ensure that predictors have adequate cell sizes. As in development of the original RSI,[3] we aggregated these sparsely represented diagnoses by truncating the fifth digit off of the corresponding ICD-9-CM diagnosis code. Codes with fewer than 1,000 discharges per year on average in the 80% model development cohort were truncated to four digits (for this average calculation, we excluded the year 2004 as there were a number of new codes introduced the next year). The process was repeated, truncating sparsely represented four-digit codes to three digits. Three-digit codes represented by fewer than 1,000 discharges per year were not considered in the model development. A comparable aggregation algorithm was implemented for the procedure codes, although we note that procedure codes are represented by a maximum of four digits and the base codes are only two digits; thus we aggregated procedure codes from four to three to two digits based on the 1,000 discharges per year criterion.

We used an elastic net approach to fit logistic models based on the aggregated predictors.[16] The elastic net is a "shrinkage" methodology devised to ensure protection against overfitting a model to the development cohort. The term "shrinkage" comes from the fact that regression coefficients are purposely biased toward zero; this action has been shown to improve prediction accuracy in external cohorts (specifically, the elastic net encourages highly correlated predictors to be averaged, whereas at the same time encouraging irrelevant predictors to be removed from the model altogether).[17,18] Removing variables in this manner has been shown to have favorable statistical properties over traditional methods such as stepwise variable selection or the use of significance criteria for entry into the model.[18] To fit these models, we used the R statistical software package "glmnet" developed by Friedman *et al.*[19] (on R version 2.13.0 for 64-bit Linux, The R Project for Statistical Computing, Vienna, Austria). The overall model shrinkage parameter (parameter λ in the glmnet software) was chosen using five-fold cross-validation[20] (specifically, we used the largest value of λ within one cross-validated standard error of the minimum in the model development cohort), and we used an elastic net mixing parameter (parameter α in the glmnet software) of 0.15, which encouraged averaging of correlated predictors a certain degree more than removing irrelevant predictors. Sensitivity analysis (not reported) revealed little change in predictive accuracy for values of α anywhere between 0.05 and 1.00.

Calibration

With pay-for-performance pressures, physicians may sometimes avoid high-risk patients for fear of the inability of the underlying risk-adjustment model to adequately adjust for their particular patient case mix.[21] In other words, there is either a perceived or real lack of agreement between predicted probability of an outcome produced by the model in question and the actual probability of the outcome in a new set of patients, *i.e.*, a lack of model calibration. Model calibration is sometimes overlooked in risk adjustment modeling[22]; even when calibration is considered, it is often as a model diagnostic[23] instead of a prescription for adjusting the model estimates to remove any biases introduced by the lack of calibration. We used a recently developed recalibration technique to adjust our model estimates.[24] Methodological details are briefly reviewed in the appendix. For our particular modeling application, we initially calibrated our model using the randomly reserved 20% calibration cohort, with the intention that—as with any risk-adjustment model—calibration should be assessed and, if necessary, corrected whenever applied to new data (such as, for instance, when we used the 2009 data to compare this model with a second model that ignored POA status of diagnoses and timing of procedures; see following section).

Comparator Models

In addition to developing an accurate baseline risk model, we sought to evaluate whether or not the absence of POA indicators precludes accurate and unbiased estimation of patients' baseline risk of mortality. To study this hypothesis, we developed a second model (which we denote "AllCodeRisk"). For this model, we used the same strategy as in the primary POARisk model (including the initial calibration step). The only difference between the two models was the inclusion of all diagnosis codes within the AllCodeRisk model regardless of whether or not they were POA, and all procedure codes regardless of when they were performed during the index hospitalization.

We also compared the POARisk and AllCodeRisk models with a modification of the original RSIs in which we used the original model coefficients; specifically, we used only the POA diagnoses, primary procedure, and secondary procedures occurring before the date of the primary procedure in the calculation of the modified RSI, whereas the original model used all codes. Although we acknowledge that it would be more appropriate to entirely redevelop the RSI model based on the subset of POA codes (due to technical issues regarding the effects of modifying correlated predictors in a regression model), this simple approach could enable practical application of the RSI model so long as hospital performance is not substantially different from that described by the POARisk model. The three models under comparison are summarized in table 1.

Model Performance and Reliability

We used the 2009 California State Inpatient Database to prospectively evaluate the performances of the POARisk, AllCodeRisk, and modified RSI models. Each risk score was recalibrated specifically with the 2009 data, using the same calibration methodology described above and in the

Table 1. Summary of the Three Prediction Models for In-hospital Mortality Being Compared

Model	Primary or Comparator Model	Development Cohort	Predictors	Number of Coefficients in Final Model	C-Statistic*
POARisk	Primary	2005–2008 California inpatients (n = 12.7 million)	Diagnoses: POA only Procedures: primary; secondary only if exclusively before date of primary Other: age, gender	1,976	0.958
AllCodeRisk	Comparator	2005–2008 California inpatients (n = 12.7 million)	Diagnoses: all Procedures: all Other: age, gender	2,091	0.981
Modified RSI†	Comparator	2001–2006 Medicare patients (n = 17.6 million)	Diagnoses: POA only Procedures: primary; secondary only if exclusively before date of primary Other: age, gender	185	0.936

Diagnosis- and procedure-based predictors are derived from International Classification of Diseases, Version 9, Clinical Modification (ICD-9-CM) codes.

* The C-statistic is a measure of discrimination between 0.5 and 1.0, where a value of 0.5 represents random guessing and 1.0 represents perfect separation of outcomes. The C-statistic was estimated within the 2009 California inpatient dataset after recalibration.
† The original RSI was developed using all diagnosis and procedure codes. We retained the original model coefficients but only applied these coefficients to the subset of diagnosis and procedure codes described in the table.

POA = present-on-admission; RSI = Risk Stratification Index.

appendix. Discriminative ability of the corrected scores among the 2009 data was evaluated using the C-statistic,[25] and models were compared on C-statistics using two-sample *z*-tests for proportions (the Bonferroni correction for three simultaneous pairwise comparisons was applied for these tests).

Hospital performance for 2009 was characterized using the ratio of observed-to-expected mortality ratios (O/E ratios), with the expected mortality rate that defined the denominator differing based on the three risk models. For a given model, the expected number of mortalities was calculated as the sum of individual patients' predicted probabilities of mortality. We excluded hospitals for which there were data on fewer than 500 inpatient stays during the year 2009.

Scatterplots of hospital O/E ratios (comparing either the AllCodeRisk or the modified RSI model with the definitive standard POARisk model) were made to graphically depict the nature of any changes in individual hospital performance. For each comparator model, respectively, the percent difference in O/E ratio was computed for each hospital and analyzed using a histogram. Good approximation of hospital performance was defined *a priori* as at least 95% of hospitals with an O/E ratio within ±20% of that defined by the POARisk model, or in other words a "ratio of O/E ratios" between 0.8 and 1.2.

We also characterized hospital performance according to rank-based categories (top 10%, 10–30%, 30–70%, 70–90%, and bottom 10%) under the three models and compared category assignments for the AllCodeRisk and modified RSI models to the assignments for the POARisk model, respectively. The number of hospitals for which the performance category assignment was the same and the number for which the assignment differed by only one performance category from the POARisk model were reported.

Results

Of the 20 million discharges in the 2004–2008 California State Inpatient Database, 7.3 million were associated with inpatient stays for which no procedures were performed. Removing these discharges and randomly partitioning the data, we used 10.1 million discharges (80%) for fitting the logistic models and 2.5 million (20%) for estimating the calibration curves. Aggregation of the ICD-9-CM codes based on the cell-size criterion of 1,000 patients per year (on average, for the years 2005–2008) resulted in 2,476 predictors for the POARisk model (1,807 diagnosis-related predictors, 666 procedure-related predictors, and 3 demographic-related predictors) and 2,584 predictors for the AllCodeRisk model (1,870 predictors, 711 predictors, and 3 predictors, respectively). The elastic net logistic regression modeling algorithm removed 501 of 2,476 (20.2%) and 494 of 2,584 (19.1%) irrelevant predictors, respectively.

Calibration of the raw risk scores among the randomly reserved 20% initial calibration cohort was generally poor (fig. 2A). Correcting the raw risk scores based on the calibration curves in figure 2A and applying these final risk scores to the 2009 data, calibration performance generally improved (fig. 2B) for the POARisk and AllCodeRisk models. However, risk as defined using the modified RSI was consistently higher than that observed, mainly because the

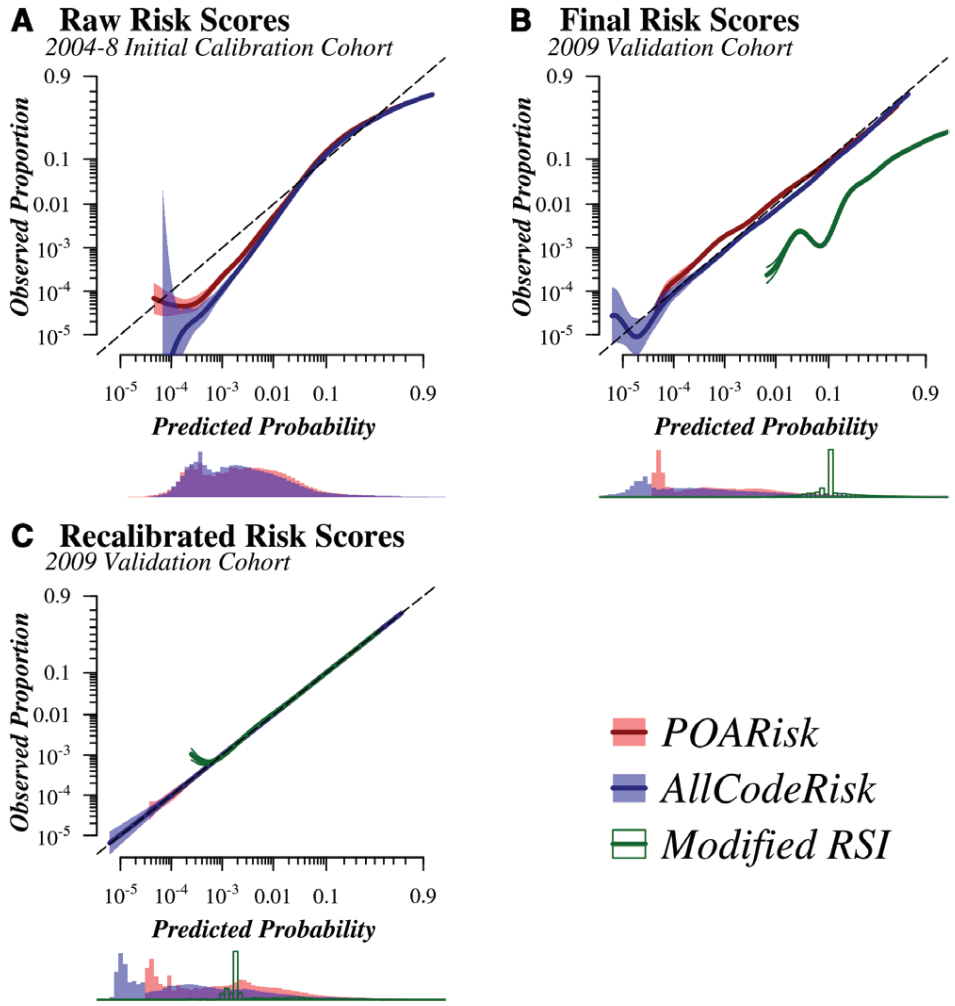

Fig. 2. Calibration curves displaying the relation between observed outcomes and model predictions. Perfect calibration is designated by the 45° line through the origin. Histograms of the risk scores underlie each panel. Calibration curves are truncated to the middle 99% of the data, or to the limits of the displayed axes. (*A*) Displays the calibration of raw POARisk and AllCodeRisk scores from the logistic model, within the random 20% calibration cohort from 2004 to 2008. Correcting POARisk and AllCodeRisk predictions based on these curves improved calibration but was insufficient to ensure complete calibration in the prospective 2009 data (*B*). The modified Risk Stratification Index (RSI) tended to overestimate risk in the general California inpatient population. However, recalibration within the 2009 data based on the curves in *B* yielded favorable calibration for all three models, as displayed in *C*. POA = present-on-admission.

RSI was developed among the higher risk Medicare population, and comparatively speaking, risk among all California patients undergoing a procedure is lower; as the histogram reveals, the modified RSI model predicted risk at approximately 10% for a large proportion of patients. Recalibration of all the three published models specifically to the 2009 data seemed to ameliorate the miscalibrations associated with year-to-year differences and differences in patient populations (fig. 2C).

Overall, the incidence of in-hospital mortality within the 2009 patients undergoing at least one procedure was 2.46%. C-statistics for the POARisk, AllCodeRisk, and modified RSI models are provided in table 1. The AllCodeRisk discriminated outcomes better than both the POARisk model (difference [Bonferroni-adjusted 95% CI] in proportions of 0.0227 [0.0224–0.0230]; $P < 0.001$, two-sample z-test for proportions) and the modified RSI model (0.0444 [0.0440–0.0448], $P < 0.001$). In addition, the POARisk model discriminated better than the modified RSI (0.0216 [0.0212–0.0221], $P < 0.001$).

In the analyses comparing hospital performance metrics under the three different risk models, we excluded 71 of 424 hospitals (16.7%) due to the fact that there were fewer than 500 inpatient discharges reported by them in 2009. A scatterplot of O/E mortality ratios based on the (recalibrated, as in fig. 2C) POARisk and AllCodeRisk models is provided in figure 3A. The median percent difference in O/E ratio—using the AllCodeRisk model *versus* using the POARisk model—was 0.0% (fig. 3B). The O/E ratio under the AllCodeRisk model was within ±20%

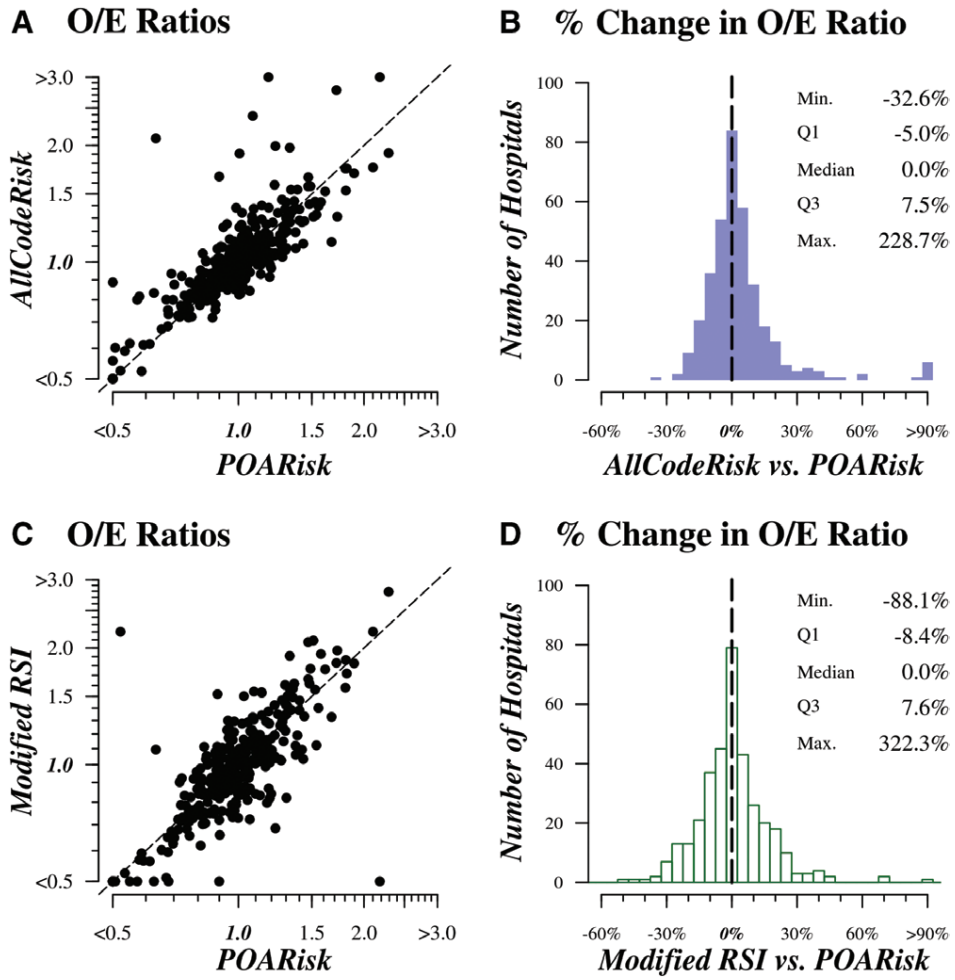

Fig. 3. Comparison of the AllCodeRisk model to the POARisk model (*A* and *B*) and of the modified Risk Stratification Index (RSI) model to the POARisk model (*C* and *D*). On the left are scatterplots of hospital observed-to-expected (O/E) ratios under each respective comparator model *vs.* O/E ratios under the POARisk model (risk scores based on the models as recalibrated to the 2009 data, as in fig. 2C). Histograms of percent difference in O/E ratios are given on the right. Hospitals with fewer than 500 discharges in 2009 were excluded from these plots. POA = present-on-admission.

of the O/E ratio under the POARisk model for 89.0% of hospitals, which was lower than our predefined criterion for agreement (*i.e.*, at least 95% of hospitals within ±20%). For 95% of hospitals, the O/E ratio under the AllCodeRisk model was between −18.1 and +51.2% of the O/E ratio under the POARisk model. Comparing the modified RSI model with the POARisk model (fig. 3, C and D), we found similar results. The median percent difference was again 0.0%. The O/E ratio under the modified RSI model was within ±20% of the O/E ratio under the POARisk model for 81.3% of hospitals. For 95% of hospitals, the O/E ratio under the modified RSI model was between −31.3 and +35.8% of the O/E ratio under the POARisk model.

Rank-based hospital performance categorizations comparing both the AllCodeRisk and Modified RSI models to the POARisk model are given in table 2. Overall, 122 of 353 hospitals (34.6%) had a different category classification under the AllCodeRisk model than under the POARisk model. Likewise, 127 hospitals (36.0%) had a different category classification under the modified RSI model than under the POARisk model.

Discussion

Risk adjustment is fraught with difficulties. Choosing the optimal risk adjustment model for assessing hospital quality has important strategic and financial implications. To a large extent, statistical performance should guide this decision. We developed two highly accurate models for in-hospital mortality, based on differing sets of ICD-9-CM codes. The POARisk model used only codes that would resemble our best conception of baseline risk (*i.e.*, POA diagnosis codes, principal procedure codes, and secondary procedure codes occurring on dates exclusively before the date of the principal procedure), whereas the AllCodeRisk model took all

Table 2. Cross-tabulations of 2009 Rank-based Hospital Performance Categories for the AllCodeRisk and Modified RSI Models *vs.* the POARisk Model

		Performance Category (AllCodeRisk Model)				
		Top 10%	10–30%	30–70%	70–90%	Bottom 10%
Performance category (POARisk model)	Top 10%	33	1	1		
	10–30%	2	51	16		1
	30–70%		18	94	26	4
	70–90%			30	31	9
	Bottom 10%			1	13	22
		Performance Category (Modified RSI Model)				
		Top 10%	10–30%	30–70%	70–90%	Bottom 10%
Performance category (POARisk model)	Top 10%	34				1
	10–30%	4	42	22	2	
	30–70%	1	18	94	27	2
	70–90%		5	24	32	9
	Bottom 10%	1		2	9	24

Data presented as hospital counts. Hospitals with fewer than 500 discharges in 2009 were excluded.
POA = present-on-admission; RSI = Risk Stratification Index.

codes regardless of their timing. Both models were assessed for calibration and corrected for use in external datasets, although we recommend that an additional calibration step be performed whenever the model is to be applied in order to adjust model estimates specifically to the population being analyzed (as we did to the 2009 data in our validation analysis). Instructions for downloading model coefficients and calculating POARisk predicted probabilities can be found at the Cleveland Clinic POARisk Model website.††

We confirmed with these two models that if proper modeling techniques are used, highly predictive models are possible using administrative data. The AllCodeRisk model discriminated outcomes slightly better than the POARisk model in prospective test data. The fact that better discrimination of outcomes could be achieved using all codes is not surprising because mortality is predicted better at discharge than at admission. Modifying the previously published RSI model gave comparable results, although the discriminative ability was slightly lower than that of the other two models.

Independent of statistical performance, however, is the greater question of which variables to include in the risk-adjustment model and when these variables should be measured. The primary goal of risk-adjusted outcomes comparison is to "level the playing field" by accounting for differences in patient case mix and severity of disease. As such, the ideal risk adjustment model would take advantage of all available baseline information to provide the most accurate possible prediction of risk. Our results indicate, unsurprisingly, that more accurate portrayals are possible by incorporating all information that are accrued through the stay, but including hospital-acquired complications detracts from the main goal. Adjusting for hospital-acquired complications inflates expected outcome risk and makes lower-performing hospitals appear as if they are higher performing.

The term "baseline" can be open to interpretation. Some may argue that the term refers to any preexisting condition. However, in practice, preexisting conditions can only include conditions that are known at the time of admission, and may not necessarily include unknown patient conditions. For example, of the hypertensive patients presenting for surgical or medical procedures, there will be a certain proportion for which the hypertension diagnosis is not known upon admission. Newly diagnosing the hypertension would likely change the course of care so that the risk of cardiac complications from surgery (and thus risk of mortality) is reduced. Ensuring that the risk adjustment model captures only what is known upon admission appropriately credits the provider for acknowledging and adapting to the risk. Analogously, quality metrics should be sensitive to the ignorance of existent hypertension and the associated increased risk of outcomes.

Procedural risk is equally difficult. Ideally, the risk scores would incorporate risk associated with planned procedures. Unfortunately, we only have available to us those procedures that were actually performed. We felt that the principal procedure and any secondary procedure occurring on days exclusively before the date of the principal procedure provided the best possible representation of the planned procedures.

Although the State of California, on which our models were developed, represents over 10% of the overall US population and has worked to ensure quality of POA indicator coding, results may not extend to other states' data. First, POA indicators might be less reliable in states which have only recently implemented POA coding than in the State

†† Available at: http://www.clevelandclinic.org/POARisk. Accessed July 23, 2012.

of California. Even excluding the POA issue, there is recent evidence of regional variability in coding practices that may also detract from the broader applicability of our models.[26] Finally, our analysis did not account for potential correlation among data from multiple repeated visits within a person. Thus, external validation is necessary to evaluate the suitability of our models in other states' data.

Our POARisk model might be considered for use in studies comparing two or more treatments, as an index for confounding due to baseline conditions. This approach seems reasonable so long as the treatments under comparison occur perioperatively. However, if a chronic exposure is being studied, an "all cause" adjustment might shroud important treatment effects that are indirectly caused (*i.e.*, mediated) by the chronic exposure. For example, if the goal of a study is to estimate the independent risk associated with diabetes mellitus, one would likely not wish to adjust for secondary diseases that are caused by diabetes and thus worsen outcomes.

In summary, we present a highly predictive baseline risk adjustment model for comparing in-hospital mortality among providers, based on POA diagnoses, the principal procedure, and secondary procedures occurring exclusively before the date of the principal procedure. Rank-based performance classifications based on a risk adjustment model which used all administrative codes (including those discovered in the hospital) departed from those based on a model restricted to solely baseline codes for roughly a third of hospitals. Thus, our POARisk model is preferable for risk adjustment when POA indicators are available.

References

1. Mukamel DB, Glance LG, Dick AW, Osler TM: Measuring quality for public reporting of health provider quality: Making it meaningful to patients. Am J Public Health 2010; 100:264–9
2. Conway PH, Clancy C: Transformation of health care at the front line. JAMA 2009; 301:763–5
3. Sessler DI, Sigl JC, Manberg PJ, Kelley SD, Schubert A, Chamoun NG: Broadly applicable risk stratification system for predicting duration of hospitalization and mortality. Anesthesiology 2010; 113:1026–37
4. Daley J, Khuri SF, Henderson W, Hur K, Gibbs JO, Barbour G, Demakis J, Irvin G III, Stremple JF, Grover F, McDonald G, Passaro E Jr, Fabri PJ, Spencer J, Hammermeister K, Aust JB, Oprian C: Risk adjustment of the postoperative morbidity rate for the comparative assessment of the quality of surgical care: Results of the National Veterans Affairs Surgical Risk Study. J Am Coll Surg 1997; 185:328–40
5. Charlson ME, Pompei P, Ales KL, MacKenzie CR: A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. J Chronic Dis 1987; 40:373–83
6. Khuri SF, Daley J, Henderson W, Hur K, Gibbs JO, Barbour G, Demakis J, Irvin G III, Stremple JF, Grover F, McDonald G, Passaro E Jr, Fabri PJ, Spencer J, Hammermeister K, Aust JB: Risk adjustment of the postoperative mortality rate for the comparative assessment of the quality of surgical care: Results of the National Veterans Affairs Surgical Risk Study. J Am Coll Surg 1997; 185:315–27
7. Zhao Y, Ash AS, Ellis RP, Slaughter JP: Disease burden profiles: An emerging tool for managing managed care. Health Care Manag Sci 2002; 5:211–9
8. Goldman L, Caldera DL, Nussbaum SR, Southwick FS, Krogstad D, Murray B, Burke DS, O'Malley TA, Goroll AH, Caplan CH, Nolan J, Carabello B, Slater EE: Multifactorial index of cardiac risk in noncardiac surgical procedures. N Engl J Med 1977; 297:845–50
9. Atherly A, Fink AS, Campbell DC, Mentzer RM Jr, Henderson W, Khuri S, Culler SD: Evaluating alternative risk-adjustment strategies for surgery. Am J Surg 2004; 188:566–70
10. Lezzoni LI: The risks of risk adjustment. JAMA 1997; 278:1600–7
11. Shahian DM, Wolf RE, Iezzoni LI, Kirle L, Normand SL: Variability in the measurement of hospital-wide mortality rates. N Engl J Med 2010; 363:2530–9
12. Glance LG, Dick A, Osler TM, Li Y, Mukamel DB: Impact of changing the statistical methodology on hospital and surgeon ranking: The case of the New York State cardiac surgery report card. Med Care 2006; 44:311–9
13. Glance LG, Dick AW, Osler TM, Mukamel DB: Does date stamping ICD-9-CM codes increase the value of clinical information in administrative data? Health Serv Res 2006; 41:231–51
14. Glance LG, Osler TM, Mukamel DB, Dick AW: Impact of the present-on-admission indicator on hospital quality measurement: Experience with the Agency for Healthcare Research and Quality (AHRQ) Inpatient Quality Indicators. Med Care 2008; 46:112–9
15. Lezzoni LI: Finally present on admission but needs attention. Med Care 2007; 45:280–2
16. Zou H, Hastie T: Regularization and variable selection *via* the elastic net. J R Statist Soc B 2005; 67:301–20
17. Hastie T, Tibshirani R, Friedman JH: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edition. New York, NY, Springer, 2009
18. Tibshirani R: Regression shrinkage and selection *via* the lasso. J R Statist Soc B 1996:267–88
19. Friedman J, Hastie T, Tibshirani R: Regularization paths for generalized linear models *via* coordinate descent. J Stat Softw 2010; 33:1–22
20. Pawitan Y: In All Likelihood: Statistical Modelling and Inference Using Likelihood. Oxford, Oxford University Press, 2001
21. Rosenthal MB: What works in market-oriented health policy? N Engl J Med 2009; 360:2157–60
22. Pace NL: Incomplete validation of risk stratification indices. Anesthesiology 2011; 115:214–5; author reply 215–6
23. Sigl JC, Sessler DI, Kelley SD, Chamoun NG: Incomplete validation of risk stratification indices. Anesthesiology 2011; 115:215–6
24. Dalton JE: Flexible recalibration of binary clinical prediction models. Stat Med 2013; 32:282–9
25. Harrell FE Jr, Califf RM, Pryor DB, Lee KL, Rosati RA: Evaluating the yield of medical tests. JAMA 1982; 247: 2543–6
26. Song Y, Skinner J, Bynum J, Sutherland J, Wennberg JE, Fisher ES: Regional variations in diagnostic practices. N Engl J Med 2010; 363:45–53

Appendix: Recalibration Methodology

Suppose there is a risk score for patient i defined as follows:

$$RS_i = \log\left[\frac{\hat{p}}{(1-\hat{p})}\right].$$

That is, the risk score is defined as the *log-odds* of the estimated outcome probability for the patient. It is assumed that the risk score was derived from a cohort that is independent from the cohort used to calibrate the models.

A standard linear-logistic regression model can be used for model calibration:

$$\text{logit}(p_i) = \log\left[\frac{p_i}{(1-p_i)}\right] = \beta_0 + \beta_1 RS_i.$$

In this model, we might conclude that the risk score is well-calibrated if $\beta_0 = 0$ and $\beta_1 = 1$. However, this calibration model can be generalized by introducing an *offset term*, *i.e.*, a predictor in the regression equation for which the coefficient is fixed at 1.0:

$$\begin{aligned}\text{logit}(p_i) &= \log\left[\frac{p_i}{(1-p_i)}\right] = \beta_0 + (\beta_1 + 1)RS_i \\ &= RS_i + \beta_0 + \beta_1 RS_i.\end{aligned}$$

Subtracting RS_i from both sides and performing some further algebraic manipulation, we can represent the *observed-to-expected odds ratio* by a linear regression equation:

$$\begin{aligned}O/E\ Odds\ Ratio &= \log\left[\frac{p_i}{1-p_i}\right] - \log\left[\frac{\hat{p}}{1-\hat{p}_i}\right] \\ &= \log\left[\frac{p_i/(1-p_i)}{\hat{p}/(1-\hat{p}_i)}\right] = \beta_0 + \beta_1 RS_i.\end{aligned}$$

Rearranging the regression equation in this way allows us to generalize the parametric form of the right-hand side $(\beta_0 + \beta_1 RS_i)$. For example, we might utilize a quadratic term $\beta_2 RS_i^2$ or discrete risk strata (*i.e.*, $\beta_1 I(RS_i < -2) + \beta_2 I(-2 \le RS_i < 0) + \ldots$). Furthermore, this rearrangement allows us to correct (recalibrate) our model by simply adding to the risk score the model prediction (for instance, if the calibration model results in a regression equation of $\hat{\beta}_0 + \hat{\beta}_1 RS + \hat{\beta}_2 RS^2$, then correcting the risk score for a new patient can be achieved by evaluating $(RS_i + \hat{\beta}_0 + \hat{\beta}_1 RS + \hat{\beta}_2 RS^2)$. We used restricted cubic splines for our calibration models.

Sample R Implementation

An example R implementation of a calibration model which incorporates restricted cubic splines can be achieved using the glm() function as follows:

$$g \leftarrow glm(Mortality \sim offset(RiskScore) + bs(RiskScore, 5), \\ data = validationData, family = "binomial")$$

Here, the model is developed from the dataset named "validationData," which includes a binary indicator variable named "Mortality" and a numeric variable named "RiskScore." The risk scores are expressed on the log-odds scale as mentioned earlier. This particular implementation uses a five-degree-of-freedom restricted cubic spline by invoking the bs() function (which can be found in the splines library).

The calibration curve can then be obtained and plotted from this model by defining a sequence of risk score values "rs" (again, on the log-odds scale), and then obtaining model predictions from the glm object named "g":

$$predProbs \leftarrow seq(from = 0.001, to = 0.999, by = 0.001)$$
$$rs \leftarrow log(predProbs / (1 - predProbs)$$
$$calCurve \leftarrow predict(g, newdata = data.frame(RiskScore = rs))\$fit$$
$$plot(x = rs, y = calCurve,\ type = "l")$$

Correcting the risk scores for the individual patients in the dataset validationData according to the calibration curve amounts to simply applying the model g to the risk scores:

$$validationData\$NewRiskScore \leftarrow predict\left(g,\ newdata = data.frame(RiskScore = validationData\$RiskScore)\right)\$fit$$