Application of the Risk Stratification Index to Multilevel Models of All-condition 30-Day Mortality in Hospitalized Populations Over the Age of 65

Vikas Saini, MD and Valérie Gopinath, MS

Background: The Risk Stratification Index (RSI) is superior to Hierarchical Conditions Categories (HCC) in patient-level regressions but has not been applied to assess hospital effects.

Objective: The objective of this study was to measure the accuracy of RSI in modeling 30-day hospital mortality across all conditions using multilevel logistic regression.

Subjects and Data Sources: A 100% sample of Medicare inpatient stays from 2009 to 2014, restricted to patients greater than 65 years of age in general hospitals, resulting in 64 million stays at 3504 hospitals.

Research Design: We calculated RSI and HCC scores for patient stays using multilevel logistic regression in 3 populations: all inpatients, surgical, and nonsurgical. Correlations of risk-standardized mortality rates with rates of specific case types assessed case-mix balance. Patient stay volume was included to assess smaller hospitals.

Results: We found a negligible correlation of all-conditions risk-standardized mortality rates with hospitals' proportions of orthopedic, cardiac, or pneumonia cases. RSI outperformed HCC in multilevel regressions containing both patient and hospital-level effects. *C*-statistics using RSI were 0.87 for the all-inpatients group, 0.87 for surgical, and 0.86 for nonsurgical stays. With HCC they were 0.82, 0.82, and 0.81. Akaike Information Criteria and Bayesian Information Criteria values were higher with HCC. RSI shifted 41% of hospitals' rankings by >1 decile. Hospitals with smaller volumes had higher 30-day observed and standardized mortality: 11.2% in the lowest volume quintile versus 8.5% in the highest volume quintile.

Conclusion: RSI has superior accuracy and results in a significant shift in rankings compared with HCC in multilevel models of 30-day hospital mortality across all conditions.

Key Words: multilevel logistic regression, hospital performance, administrative data, risk adjustment, case-mix

(*Med Care* 2021;59: 836–842)

To ensure fair comparisons of hospital performance, adequate risk adjustment remains an important goal. The Risk Stratification Index (RSI), a machine algorithm that exploits the structure of the International Classification of Diseases (ICD) coding system to enhance the signal and reduce noise, substantially captures available information and yields robust, stable results at the patient level with accuracy superior to Hierarchical Condition Categories (HCC) methods and neural network models.[1–5]

RSI has not been assessed for hospital-level performance, where multilevel techniques are necessary to avoid confounding of within- and between-hospital effects.[6,7] Most hospital evaluations, including CMS Hospital Compare, apply multilevel regressions using clinical groupings within specific conditions[8–11] and variably include hospital characteristics such as volume and staffing.[5,12,13] While condition-specific evaluations can support improvement efforts, the use of all-conditions populations with adequately overlapping case-mix[14] would theoretically provide greater insights into global hospital treatment effects.

The incorporation of diagnosis codes as risk adjusters in hospital ranking may improve performance.[15–17] Because RSI efficiently incorporates large numbers of such codes, our 2 specific objectives were: (1) to assess its utility for measuring a hospital global effect by applying it to all-conditions populations; and (2) to compare its overall performance with HCC using identical methods. To do this we applied multilevel models with both patient-level risk as well as fixed and random hospital-level effects to broad classes of inpatients and investigated any residual correlation of the risk-standardized mortality (RSMR) rates with hospital rates of specific case types.

METHODS

Under a Data Use Agreement with the Center for Medicare and Medicaid Services, >106 million patient stay records from the 2009 to 2014 Medicare Analysis Provider and Review (MEDPAR) files (100% sample) served as the universe of our modeling datasets. The MEDPAR file contains fee-for-service claims data from Medicare-certified inpatient hospitals and skilled nursing facilities. Patient-level data include age, sex, race, provider information, and up to 25 fields for the International

From the Lown Institute, Brookline, MA.

V.S. owns equity but has no other role in Health Data Analytics Institute, a privately held company working on software tools for the health care sector. V.G. declares no conflict of interest.

Correspondence to: Vikas Saini, MD, Lown Institute, 21 Longwood Avenue, Fl 2, Brookline, MA 02446. E-mail: vsaini@lowninstitute.org.

Supplemental Digital Content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's website, www.lww-medicalcare.com.

ISSN: 0025-7079/21/5909-0836

TABLE 1. Data Summary

Total number of MEDPAR Patient stays 2009–2014: 106.3 million
↓
After applying hospital selection criteria* and age ≥65: 65.4 million
↓
Minimum 50 stays annually: 65.4 million patient stays at 3706 hospitals
↓
Restrict to AHA Primary Service Code 10 "General medical and surgical" and exclude hospitals defined by MedPAC as specialty
64.0 million patient stays at 3504 hospitals

Patient stay category	# of hospitals	# of inpatient stays	50% random sample
All patients	3504	64.0 million (100%)	32.0 million
Surgical	3478	37.8 million (59.0%)	18.9 million
Nonsurgical	3502	26.2 million (41.0%)	13.1 million

*Acute care and critical access hospitals in 50 US States.

AHA indicates American Hospital Association; MedPAC, Medicare Payment Advisory Commission; MEDPAR, Medicare Analysis Provider and Review.

Classification of Diseases, Ninth Revision (ICD-9) diagnosis codes, and up to 25 procedure codes, each with an associated present on admission (POA) code indicator.[18,19] Table 1 shows the summary of MEDPAR records we used for this analysis. Patients under 65 years of age were excluded. We randomly split the data into 50% for model development and 50% holdout and within each identified 3 populations of interest: all-inpatient stays, surgical stays, and nonsurgical stays.

Hospital Selection

Excluding skilled nursing facilities, we included hospitals within the 50 US states and the District of Columbia classified as Short-Term General hospitals or Critical Access by CMS (see Fig., Supplemental Digital Content 1, which shows the flow chart of our hospital selection process, http://links.lww.com/MLR/C249).

Defining General Hospitals and Addressing Case-mix Variation

We sought to study general rather than specialty hospitals, given the substantial differences in case-mix and a priori patient risk between those 2 types.

In preliminary work, we found specialty orthopedic hospitals had lower observed and expected mortality rates (data not shown). While specialty hospitals may have lower mortality compared with community hospitals,[20] we were concerned that this could reflect case selection for elective procedures, and unfairly overestimate their expected mortality from a national regression of all inpatients. This would result in superior risk-adjusted mortality at the hospital level even after patient-level adjustment, effectively importing the ecologic fallacy into rankings.[21,22]

To reduce variation in case-mix, we restricted our analysis to General Medical and Surgical Hospitals in the American Hospital Association (AHA) database. We then applied previously published Medicare Payment Advisory Commission (MedPAC) criteria[20,23] to identify any hospitals with a specialty case-mix despite AHA classification and excluded them. We developed a data reduction tool for case-mix adjustment using factor analysis[24,25] on "Base-DRGs" (see Text, Supplemental Digital Content 2, which details the case-mix factor analysis, http://links.lww.com/MLR/C250) to assess the impact on hospital performance. Using the 50% model development sample, we performed separate multilevel logistic regressions estimating 3 models, one for each population (all-inpatient, surgical, nonsurgical) against the outcome of interest, 30-day mortality.

Multilevel Regression Modeling

Model Development

The RSI method calculates the risk posed by comorbidities jointly with the risk of any associated procedure. Diagnosis and procedure codes [International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM)] generate an optimum covariate set for modeling endpoints. The codes are hierarchical, permitting a roll-up algorithm to assign them to a higher level when the sample size is insufficient. The algorithm collapses codes into binary indicators (0 or 1) to create a candidate covariate set. This process is detailed in previous RSI publications[1–3] (see Fig., Supplemental Digital Content 3, which shows the flow chart, http://links.lww.com/MLR/C251). We restricted our analysis to POA codes, excluding those added during hospitalization to avoid contamination that might reflect hospital performance.

Adjustments for Hospital Volume

Estimates of mortality rates in low-volume hospitals are subject to greater sampling error, reducing the confidence of estimates. As a result, estimates of the performance of low-volume hospitals are "shrunken" toward the national mean, making it difficult to determine both high-performing and low-performing low-volume hospitals. Others have shown that including volume improves mortality estimates for low-volume hospitals,[5] and we, therefore, included it as a covariate.[5] To ensure sufficient representation of small volume hospitals we examined hospitals that performed surgeries during 2009–2014 ($n = 4231$). From their distribution of patient stay volume, we identified a minimum of 50 admissions per year would include 87.6% of hospitals and 99.4% of patient stays. We selected this cutoff point (see Fig., Supplemental Digital Content 4, for distribution of patient hospital stay volume in hospitals, http://links.lww.com/MLR/C252). When we used a minimum of 250 admissions per year only 2362 hospitals (55.8% of hospitals, 93.2% of patient stays) were included.

We created a 2-level logistic regression model for 30-day mortality. Level 1 effects included patient conditions, age, and sex. Level 2 effects included hospital volume quintile and hospital random intercept. The candidate covariate set included 1827 diagnosis codes derived from the truncation and roll-up algorithm. Before estimating the multilevel model, we used a forward stepwise selection process with relaxed inclusion criteria ($P < 0.001$ for covariate addition and $P < 0.0005$ to avoid covariate removal). These criteria were selected to allow the largest number of likely variables to be identified. We then estimated a multilevel model with covariates further selected based on their statistical significance. The criterion for the final iteration of the algorithm ($P < 10^{-6}$) was selected after examining the output and identifying a threshold below which the highly significant variables were clustered. In the final model, comorbidity variables not meeting a significance threshold of P-value $<10^{-6}$ were eliminated.

Models were estimated from the 50% model development sample. We used samples of 1.5 million for all-inpatient and nonsurgical populations and 2 million for surgical. An RSI score for each patient was then developed from the estimated regression coefficients representing the log odds for 30-day mortality. Similarly, an HCC score was derived from the regression coefficients for HCC. Our final models included a range of 120–300 variables from the selected diagnosis and procedure codes, age, sex, hospital volume quintile, and hospital intercept.

To assess the effects of patient-level and hospital-level case-mix on model performance and rankings, we ran an identical set of models adding case-mix adjustment using the factor analysis (see Text, Supplemental Digital Content 2, showing the case-mix factor analysis, http://links.lww.com/MLR/C250). Patient-level case-mix did not contribute to model performance in early iterations. We then focused on case-mix as a hospital-level fixed effect by calculating the percent of cases in each factor and examined the impact on model performance and resulting RSMRs.

To assess residual case-mix bias, we examined the correlations of RSMRs with hospital-specific rates of orthopedic, cardiac, and pneumonia cases both with and without adjustment by the case-mix factors.

Application of Risk Stratification Index and Hierarchical Condition Categories to Hospital Performance Evaluation

After calibration, the predicted mortality scores for each patient stay were aggregated to create predicted (P) and expected (E) mortality rates for both RSI and HCC models. P was the predicted mortality for the patients of a given hospital obtained from the multilevel model based on patients' conditions, age, sex, hospital random intercept, hospital volume quintile, and in the second set of models, case-mix factors; E was the expected mortality for the patients of that hospital obtained from ordinary logistic regression using only patient conditions, age, sex, and excluding both hospital volume quintile and hospital case-mix factors. We calculated each hospital's RSMR by multiplying its P/E ratio by the national observed mortality rate across all hospitals in our data.[8,9,26]

To ensure comparability, we included the same variables in both RSI and HCC multilevel models. The only difference was the use of Hierarchical Condition Categories for HCC and diagnosis codes for RSI. In models adjusting for case-mix, we added the proportion of cases in each of the case-mix factors to the models.

We compared model performance using the receiver operating characteristic curve C-statistic for discrimination as well as the information criteria of both Akaike Information Criteria and Bayesian Information Criteria (AIC and BIC).[27,28] To assess the impact of models on ranking, we calculated the mean shift in ranking using bootstrapping. We generated 500 bootstrapped datasets and then calculated the mean change in rank position as well as the 95% confidence interval. We also compared the percent of hospitals whose rankings changed across deciles along with a 95% confidence interval.

Software

Statistical procedures were performed using SAS, version 9.4 (SAS Institute, Cary, NC) and R, version 3.3.2.

RESULTS

Using the MEDPAR files from 2009 through 2014, we found over 20,000 facilities with 106.3 million patient stays. After applying beneficiary age as well as hospital type and volume criteria, our initial dataset consisted of 65.4 million patient stays at 3706 hospitals.

Defining General Hospitals

From this list, 3525 hospitals corresponded to the AHA General Medical and Surgical category. Of these, we found 21 hospitals that met the MedPAC criteria for cardiac, orthopedic, or surgery specialty hospitals. After their removal, 3504 hospitals were our final hospital population; 81.7% were Short-Term General Hospitals and 18.3% Critical Access Hospitals. The patient population sizes are shown in Table 1. The 32.0 million patient stays of the all-inpatient group represented a 50% random sample (50% for model development and 50% holdout) captured in the CMS Chronic Conditions Data Warehouse (CCW).

Characteristics of the Study Populations by Hospitals and by Diagnostic Profiles

For the all-inpatient population, the patient stays had an average age of 78.0, were 56.4% female, and had a 30-day mortality of 8.6% (see Table, Supplemental Digital Content 5, which shows the patient characteristics, http://links.lww.com/MLR/C253). For that group, the most frequent diagnoses were septicemia not otherwise specified (3.7%), pneumonia (3.7%), obstructive bronchitis, and urinary tract infections (2.4% each). The surgical and nonsurgical populations had similar diagnoses in differing order (see Table, Supplemental Digital Content 6, which shows the 10 most frequent diagnoses, http://links.lww.com/MLR/C254).

Model Results

The Effect of Case-mix on Model Performance

RSI models had excellent discrimination. Inclusion of diagnosis-related group (DRG)-based case-mix factors to RSI models did not improve model discrimination (C-statistic of 0.87 for both, Table 2). As an additional assessment, an RSI model using a simple case adjuster (% orthopedic patients) had an identical C-statistic of 0.87 (data not shown).

At the patient level, case-mix factor indicators also did not significantly change RSI ranking results: only 0.4% of the hospitals shifted by >1 decile (see Table, Supplemental Digital Content 7, which shows the ranking shift for the patient-level variable, http://links.lww.com/MLR/C255). With case-mix factors as hospital-fixed effects, 32.4% (30.8–33.4) of hospitals shifted 1 or more deciles and 3.2% (2.6–3.8) of hospitals shifted 2 or more deciles (see Table, Supplemental Digital Content 8, which shows the ranking shift for the

TABLE 2. General Hospitals 30-Day Mortality: RSI Versus HCC Performance

	General Hospitals 30-Day Mortality					
	RSI vs. HCC					
	C-statistic					
	RSI			HCC		
	Development Set		Holdout Set	Development Set		Holdout Set
Inpatient Population	Without Case-mix Factors	With Case-mix Factors	With Case-mix Factors	Without Case-mix Factors	With Case-mix Factors	With Case-mix Factors
All inpatients (3504 hospitals)	0.87	0.87	0.86	0.82	0.82	0.82
Surgical stays (3478 hospitals)	0.87	0.87	0.87	0.82	0.82	0.82
Nonsurgical (3502 hospitals)	0.86	0.86	0.86	0.81	0.81	0.81

Model performance characteristics across 3 inpatient populations.
HCC indicates Hierarchical Condition Categories; RSI, Risk Stratification Index.

hospital-fixed effect, http://links.lww.com/MLR/C256) with a modest mean change in the ranking of 129.5 (confidence interval: 5.2–300.6).

Case-mix adjustment ranked the 20 hospitals with the highest % of orthopedic surgery cases lower by an average of 205 positions (median: 218) (see Fig., Supplemental Digital Content 9, which compares the performance of RSI and RSI with case-mix, http://links.lww.com/MLR/C257).

When all-conditions RSMRs were plotted against rates of 3 specific case types (Fig. 1), the correlation was minimal (R^2 values of 0.045, for orthopedic, 0.107 for cardiac, and 0.049 for pneumonia), suggesting the case-mix of this general hospital population was relatively balanced. The addition of case-mix factors into the models as hospital-fixed effects did not reduce the correlation coefficients further, except for a small drop in the inverse correlation of mortality with % orthopedic case rates (from 0.045 to 0.028)

Comparative Performance of Risk Stratification Index and Hierarchical Condition Categories Risk Adjustment

In all 3 patient groups, RSI discrimination outperformed HCC in the development set. The all-inpatient model had a higher *C*-statistic of 0.87 compared with the 0.82 of the HCC model, indicating better classification. RSI models also had lower AIC and BIC values (see Table, Supplemental Digital Content 10, which compares RSI vs. HCC performance, http://links.lww.com/MLR/C258). The inclusion of case-mix factors in RSI models did not change *C*-statistics, AIC, and BIC values, and both versions outperformed HCC models. The other 2 populations, surgical and nonsurgical, showed similar results (data not shown). The *C*-statistics of 0.86, 0.87, and 0.86 for each population in the holdout samples were essentially identical (Table 2, Fig., Supplemental Digital Content 11, which shows the associated receiver operating characteristic curves, http://links.lww.com/MLR/C259). Models using only derived RSI and HCC patient scores in the 3 populations from the holdout samples had comparable results of 0.85, 0.86, and 0.84 (see Table, Supplemental Digital Content 12, which shows comparative model performance, http://links.lww.com/MLR/C260).

The difference in accuracy between RSI and HCC resulted in a mean change in the ranking of 391.4 (19.6–845.2), with 73.9% (72.4–75.3) of hospitals shifting 1 or more deciles and 41.0% (39.4–42.7) of hospitals shifting 2 or more deciles (Table 3, see Table, Supplemental Digital Content 13, which shows shifts by decile, http://links.lww.com/MLR/C261).

Risk-standardized Mortality Rates: The Effect of Hospital Volume

The calculated P/E ratios by quintile of hospitals sorted by volume are shown in Figure 2. For each population, the highest volume hospitals had the lowest mortality. This was true for observed mortality rates (8.3% for the highest volume to 9.6% for the lowest in the all-inpatient group) as well as after risk adjustment (8.5% for the highest to 11.2% for the lowest). The results for the other population groups were similar (see Table, Supplemental Digital Content 14, which shows mortality by hospital volume, http://links.lww.com/MLR/C262). For surgical, the observed mortality range was 8.4%–11.2%, and the RSMR 9.0%–11.7%. For nonsurgical, the observed mortality range was 8.0%–9.5% and RSMR ranged from 8.5%-10.4%.

Using higher thresholds (250 all-inpatient annual stays, 150 surgical, and 100 for nonsurgical) the RSMRs were 8.7, 9.3, and 7.7, respectively and were unchanged across volume quintiles (data not shown).

DISCUSSION

RSI multilevel models applied to a group of 3504 general hospitals consistently outperformed HCC and the all-conditions RSMRs resulted in large shifts in hospital rankings. Correlations with specific case types were negligible; hospital-level, but not patient-level, case-mix factors improved these slightly and shifted rankings modestly.

Previous work has shown that at the patient level, RSI risk adjustment is more granular than HCC, resulting in superior model performance.[1,3] Since we used an identical

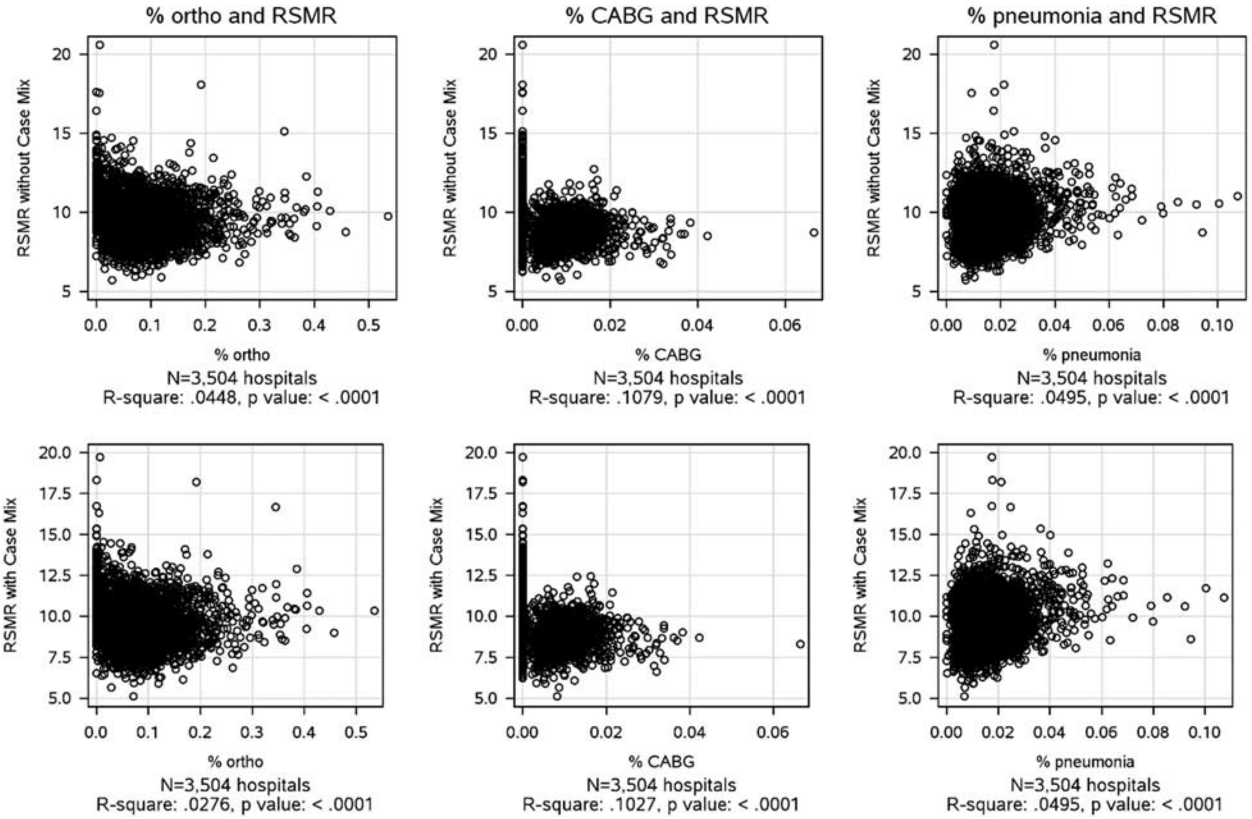

FIGURE 1. Correlations of RSMRs with hospital rates of conditions. Scatterplots of RSMRs against hospitals' percentage of 3 case types: orthopedic surgery (all-inpatient Base-DRGs from factor 8), CABG surgery, and pneumonia. Upper row: RSMRs from models without case-mix factors. Lower row: RSMRs from models in which case-mix factors were included as hospital-fixed effects. CABG indicates coronary artery bypass graft; DRG, diagnosis-related group; RSMR, risk-standardized mortality rate.

modeling approach for both RSI and HCC in the current study, it appears that the granularity of the RSIs 1827 diagnostic codes is also responsible for its superior performance here. Furthermore, RSI appears to capture sufficient information that the addition of DRG or other case-mix variables does not materially improve model fit. We believe that the negligible correlation of risk-adjusted mortality with rates of specific case types is evidence suggesting that our approach provides an estimate of a global hospital effect. To test this hypothesis, a systematic comparison of all-patient P/E ratios to condition-specific P/E ratios is required, something that is outside of the scope of the current paper.

TABLE 3. All-conditions Risk-standardized Mortality Rate Differences in Decile Ranking When Using Case-Mix Adjusted Risk Stratification Index Versus Hierarchical Condition Categories

Decile Shift	# of Hospitals	%	Cumulative Percentage
0	916	26.1	26.1
1	1151	32.9	59.0
2	726	20.7	79.7
3	403	11.5	91.2
4+	308	8.8	100.0
Total	3504		

N = 3504 general hospitals, all-inpatient model of 30-day mortality.

Performance decile is based on the rank of the hospital's risk-standardized mortality rate. Decile shift = number of deciles that a hospital's rankings differed between the 2 methods.

Compared with CMS Hospital Compare, which uses HCC in condition-specific models, the RSI multilevel models had superior discriminatory accuracy without evidence of overfitting. In the literature, reported *C*-statistics for 30-day mortality are 0.68–0.70 for heart failure,[5,9] 0.73 for acute myocardial infarction,[5] pneumonia and acute myocardial infarction,[15] and 0.864 for stroke with clinical severity adjustment.[29]

The inclusion of case-mix factors modestly shifted rankings but was more pronounced at the tail of the distribution of orthopedic case rates. We believe these modest improvements warrant their inclusion in all-conditions MLM models for comparison of hospital global effects.

We adjusted for hospital volume and added to the evidence from others that volume is an independent predictor of 30-day mortality, with lower hospital volumes associated with higher mortality.[5,12,13] Our results confirm this across all conditions and support the incorporation of volume in regression models to compare smaller hospitals fairly to each other or to larger facilities. Our results suggest there may be a threshold effect: above 250 stays per year, we found flat, essentially identical RSMRs across volume quintiles. This

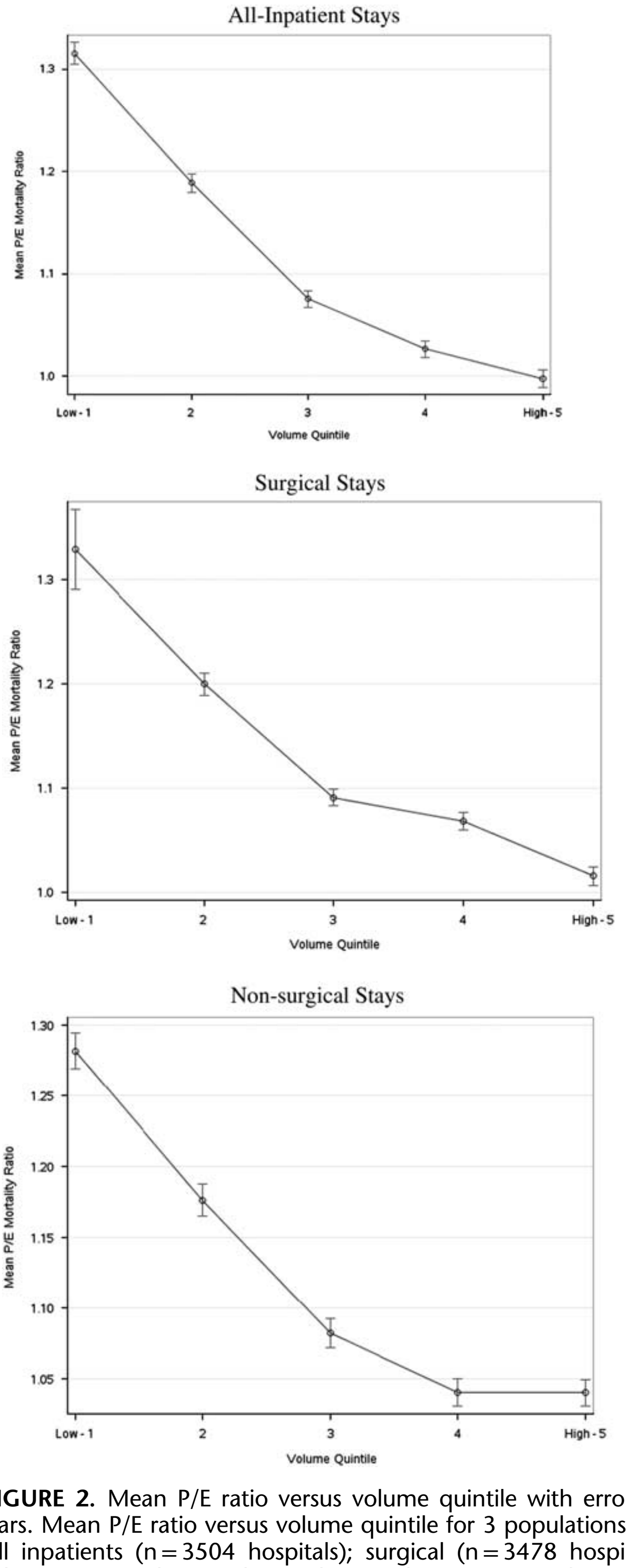

FIGURE 2. Mean P/E ratio versus volume quintile with error bars. Mean P/E ratio versus volume quintile for 3 populations. All inpatients (n = 3504 hospitals); surgical (n = 3478 hospitals); nonsurgical (3502 hospitals). E indicates expected mortality; P, predicted mortality.

threshold may be a useful minimum for comparisons when not including volume. This issue matters for consumers who might want to know the quality of the facility near them when making a decision regarding local care versus the inconvenience of traveling a significant distance. Accuracy also matters for policy, regulation, payment, and quality improvement, since low-volume hospitals are often smaller rural facilities functioning as sole providers.

The RSI method uses a large number of variables to achieve its results, raising a concern of overfitting. We do not believe that such concern is reasonable here. First, we validated our results using holdout samples of 1.5 million patient stays for all inpatients, 2 million for surgical, and 1.5 million for nonsurgical. *C*-statistics were essentially identical, as has been true in previous RSI reports. Second, the lower AIC and BIC values with RSI compared with HCC support this conclusion. Furthermore, RSI generates coefficients that are remarkably stable: application to a completely out of sample dataset of 39 million people 5 years later yielded essentially identical discrimination.[2] The similar performance of models using only patient scores supports the use of archived RSI coefficients for any given risk adjustment task without the laborious rederivation of coefficients.

Accurate assessment of hospital performance remains an important objective as efforts to improve the value of care delivery in the United States continue. As alternative payment models are refined and become more widely adopted, the ability to assess global hospital effects would enable adjustments to global payment models and allow movement beyond bundled payments for episodes of care limited to specific case types.

Limitations

Our study is limited because we chose to examine only POA codes in this initial examination of RSI in multilevel regression models. If we had incorporated a look-back period, it is possible that the capture of more codes might result in additional improvements in model fit. Our study is limited to hospitalized patients; harvesting additional diagnostic codes from outpatient and carrier claims might improve results further.

Our 3 inpatient populations have important dissimilarities and there is likely residual variation in individual procedures or diagnoses within subpopulation groups. For instance, while the case-mix variation within the surgical group may be relatively narrower, the nonsurgical group is likely more heterogenous in its risk profile. Furthermore, our study does not account for any underlying selection bias driven by provider behavior, particularly of elective cases, nor does it account for bias arising from unmeasured confounding by social risk factors, even if a case-mix adjustment were perfect. Thus, extrapolation of our results to subpopulations of patients should be done with caution.

Inclusion of our case-mix factor analysis re-ranked hospitals modestly; alternative case-mix measurement tools that capture other elements of between-hospital variation, or are more sensitive to the joint association of risk and case type might yield stronger results.

Additional efforts to address selection bias and socioeconomic status bias continue to be necessary for the fair assessment of hospital performance. Our results are only applicable to Medicare fee-for-service patients, and it is possible that our method may not be applicable to Medicare Advantage patients.

This study was limited to datasets containing ICD-9 codes. Although results may differ with ICD-10, the nested hierarchical structure is maintained and will permit the use of truncation and roll-up algorithms. Gao et al[17] have exploited this architecture using a different method with a similar effect. However, because code numbers are greater and nesting ceases at the fourth level, total numbers entering regressions will likely be substantially greater, requiring additional computation time. Preliminary results applying RSI to ICD-10 confirm this surmise but suggest similar levels of model performance.

Unsupervised machine learning or "artificial intelligence" algorithms are subject to important biases that may skew machine inferences and reinforce erroneous or unethical historical practices.[30–32] It is important to note that the RSI method is not such a machine learning/artificial intelligence program. Biases embedded in RSI will be a function of the underlying structure of the ICD-9 and 10 architectures and as such, limited by the design decisions of those systems.

CONCLUSIONS

We applied the RSI to multilevel regression models to assess global hospital performance across all conditions and found that it yielded superior discriminatory accuracy compared with the HCC method when applied to large inpatient populations over the age of 65. The greater accuracy of RSI models also resulted in a substantial re-ranking of hospitals.

These results matter because there is a public and policy interest in assessing global hospital effects as an enabling tool for global payments models. More accurate risk adjustment would improve and promote accountability and fairness in payment models. There is also a strong public interest in including the widest possible range of hospitals that Americans use, particularly low-volume hospitals in rural areas. We believe that RSI is a broadly applicable method and represents an opportunity to improve the measurement of hospital performance.

REFERENCES

1. Sessler DI, Sigl JC, Manberg PJ, et al. Broadly applicable risk stratification system for predicting duration of hospitalization and mortality. *Anesthesiology*. 2010;113:1026–1037.
2. Chamoun GF, Li L, Chamoun NG, et al. Validation and calibration of the Risk Stratification Index. *Anesthesiology*. 2017;126:623–630.
3. Chamoun GF, Li L, Chamoun NG, et al. Comparison of an updated Risk Stratification Index to Hierarchical Condition Categories. *Anesthesiology*. 2018;128:109–116.
4. Lee CK, Hofer I, Gabel E, et al. Development and validation of a deep neural network model for prediction of postoperative in-hospital mortality. *Anesthesiology*. 2018;129:649–662.
5. Silber JH, Rosenbaum PR, Brachet TJ, et al. The Hospital Compare mortality model and the volume-outcome relationship. *Health Serv Res*. 2010;45(pt 1):1148–1167.
6. Ash AS, Louis TA, Normand LT, et al. The COPSS-CMS White Paper Committee. Statistical issues in assessing hospital performance; 2011.
7. Normand SL. Some old and some new statistical tools for outcomes research. *Circulation*. 2008;118:872–884.
8. Krumholz HM, Wang Y, Mattera JA, et al. An administrative claims model suitable for profiling hospital performance based on 30-day mortality rates among patients with an acute myocardial infarction. *Circulation*. 2006;113:1683–1692.
9. Krumholz HM, Wang Y, Mattera JA, et al. An administrative claims model suitable for profiling hospital performance based on 30-day mortality rates among patients with heart failure. *Circulation*. 2006;113: 1693–1701.
10. Lindenauer PK, Bernheim SM, Grady JN, et al. The performance of US hospitals as reflected in risk-standardized 30-day mortality and readmission rates for medicare beneficiaries with pneumonia. *J Hosp Med*. 2010;5:E12–E18.
11. Li P, Kim MM, Doshi JA. Comparison of the performance of the CMS Hierarchical Condition Category (CMS-HCC) risk adjuster with the Charlson and Elixhauser comorbidity measures in predicting mortality. *BMC Health Serv Res*. 2010;10:245.
12. Han KT, Kim SJ, Kim W, et al. Associations of volume and other hospital characteristics on mortality within 30 days of acute myocardial infarction in South Korea. *BMJ Open*. 2015;5:e009186.
13. Ross JS, Normand SL, Wang Y, et al. Hospital volume and 30-day mortality for three common medical conditions. *N Engl J Med*. 2010;362:1110–1118.
14. Shahian DM, Normand SL. Comparison of "risk-adjusted" hospital outcomes. *Circulation*. 2008;117:1955–1963.
15. DeCenso B, Duber HC, Flaxman AD, et al. Improving hospital performance rankings using discrete patient diagnoses for risk adjustment of outcomes. *Health Serv Res*. 2018;53:974–990.
16. Krumholz HM, Coppi AC, Warner F, et al. Comparative effectiveness of new approaches to improve mortality risk models from Medicare Claims Data. *JAMA Netw Open*. 2019;2:e197314.
17. Gao J, Moran E, Almenoff PL. Case-mix for performance management: a risk algorithm based on ICD-10-CM. *Med Care*. 2018;56:537–543.
18. Pope GC, Ellis RP, Ash AS, et al. Diagnostic Cost Group Hierarchical Condition Category Models for Medicare Risk Adjustment: final report. Centers for Medicare and Medicaid Services; 2020.
19. Pope GC, Kautter J, Ellis RP, et al. Risk adjustment of Medicare capitation payments using the CMS-HCC model. *Health Care Financ Rev*. 2004;25:119–141.
20. Greenwald L, Cromwell J, Adamache W, et al. Specialty versus community hospitals: referrals, quality, and community benefits. *Health Aff*. 2006;25: 106–118.
21. Robinson WS. Ecological correlations and the behavior of individuals. *Am Sociol Rev*. 1950;15:351–357.
22. Greenland S, Robins J. Invited Commentary: ecologic studies—biases, misconceptions, and counterexamples. *Am J Epidemiol*. 1994;139:747–760.
23. Leavitt MO. Study of Physician-owned Specialty Hospitals Required in Section 507(c)(2) of the Medicare Prescription Drug, Improvement, and Modernization Act of 2003; 2005.
24. Desmet P. Buying behavior study with basket analysis: pre-clustering with a Kohonen map. *EJESS*. 2001;15:17–30.
25. Forcum L, Joo H. Using market basket analysis in management research. *J Manag*. 2013;39:1799–1824.
26. Krumholz HM, Brindis RG, Brush JE, et al. Standards for statistical models used for public reporting of health outcomes: an American Heart Association Scientific Statement from the Quality of Care and Outcomes Research Interdisciplinary Writing Group: cosponsored by the Council on Epidemiology and Prevention and the Stroke Council. Endorsed by the American College of Cardiology Foundation. *Circulation*. 2006;113:456–462.
27. Akaike H. Akaike's Information Criterion. In: Lovric M, ed. *International Encyclopedia of Statistical Science*. Berlin, Heidelberg: Springer; 2011:25.
28. Kass RE, Wasserman L. A reference Bayesian test for nested hypotheses and its relationship to the Schwarz Criterion. *J Am Stat Assoc*. 1995;90:928–934.
29. Fonarow GC, Pan W, Saver JL, et al. Comparison of 30-day mortality models for profiling hospital performance in acute ischemic stroke with vs without adjustment for stroke severity. *JAMA*. 2012;308:257–264.
30. Estava A, Kuprel B, Novoa R. Dermatologist level classification of skin cancer with deep neural networks. *Nature*. 2017;542:115.
31. Buolamwini J, Gebru T. Gender shades: intersectional accuracy disparities in commercial gender classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency; Proceedings of Machine Learning Research; 2018.
32. Zou J, Schiebinger L. AI can be sexist and racist—it's time to make it fair. *Nature*. 2018;559:324–326.