Anesthesiology 2010; 113:1026–37

Broadly Applicable Risk Stratification System for Predicting Duration of Hospitalization and Mortality

Daniel I. Sessler, M.D.,* Jeffrey C. Sigl, Ph.D.,† Paul J. Manberg, Ph.D.,‡ Scott D. Kelley, M.D.,§ Armin Schubert, M.D., M.B.A.,‖ Nassib G. Chamoun, M.S.#

CME This article has been selected for the ANESTHESIOLOGY CME Program. Learning objectives and disclosure and ordering information can be found in the CME section at the front of this issue.

ABSTRACT

Background: Hospitals are increasingly required to publicly report outcomes, yet performance is best interpreted in the context of population and procedural risk. We sought to develop a risk-adjustment method using administrative claims data to assess both national-level and hospital-specific performance.

Methods: A total of 35,179,507 patient stay records from 2001–2006 Medicare Provider Analysis and Review (MEDPAR) files were randomly divided into development and validation sets. Risk stratification indices (RSIs) for length of stay and mortality endpoints were derived from aggregate risk associated with individual diagnostic and procedure codes. Performance of RSIs were tested prospectively on the validation database, as well as a single institution registry of 103,324 adult surgical patients, and compared with the Charlson comorbidity index, which was designed to predict 1-yr mortality. The primary outcome was the C statistic indicating the discriminatory power of alternative risk-adjustment methods for prediction of outcome measures.

Results: A single risk-stratification model predicted 30-day and 1-yr postdischarge mortality; separate risk-stratification models predicted length of stay and in-hospital mortality. The RSIs performed well on the national dataset (C statistics for median length of stay and 30-day mortality were 0.86 and 0.84). They performed significantly better than the Charlson comorbidity index on the Cleveland Clinic registry for all outcomes. The C statistics for the RSIs and Charlson comorbidity index were 0.89 *versus* 0.60 for median length of stay, 0.98 *versus* 0.65 for in-hospital mortality, 0.85 *versus* 0.76 for

* Professor and Chair, Department of Outcomes Research, Cleveland Clinic, Cleveland, Ohio. † Senior Director, Analytical Research, Covidien, Inc., Dublin, Ireland. ‡ Vice President, Clinical Research and Regulatory Strategy, Covidien, Inc. § Vice President, Medical Affairs, Covidien, Inc. ‖ Professor and Chair, Department of Anesthesiology, Ochsner Clinic, New Orleans, Louisiana. # Vice President of Technology, Research and Clinical Development, Covidien, Inc., Dublin, Ireland. Current position: Chairman, Lown Cardiovascular Research Foundation, Brookline, Massachusetts; Adjunct Staff, Department of Outcomes Research, Cleveland Clinic.

Received from the Department of Outcomes Research, Cleveland Clinic, Cleveland, Ohio; Covidien, Inc., Dublin, Ireland; and the Department of Anesthesiology, Ochsner Clinic, New Orleans, Louisiana. Submitted for publication February 16, 2010. Accepted for publication July 30, 2010. Supported by Aspect Medical Systems (Norwood, Massachusetts); Aspect was recently acquired by Covidien, Inc. (Dublin, Ireland). The study was designed and conducted collaboratively by investigators from both organizations. Covidien, Inc., employees have a financial interest in their company, but none of the other authors has a personal financial interest in this research. The risk stratification system presented, along with anticipated future updates, will be placed in the public domain for unrestricted use by all interested parties. Drs. Sessler, Sigl, Manberg, Kelley, and Mr. Chamoun conceived the study and participated in study design. Drs. Sigl and Schubert acquired the study data. Dr. Sigl and Mr. Chamoun developed and implemented the statistical method. Dr. Schubert developed the Cleveland Clinic Perioperative Health Documentation System. Drs. Sigl and Sessler and Mr. Chamoun analyzed the data. All authors participated in interpretation of results. Drs. Manberg and Kelley drafted the manuscript, and all authors contributed to the critical review and revision of the manuscript. All authors have seen the final version of the manuscript.

Address correspondence to Dr. Sessler: Department of Outcomes Research, The Cleveland Clinic—P77, Cleveland, Ohio 44195. ds@or.org. On the World Wide Web at www.clevelandclinic.org/DOR and www.or.org.

◇ This article is featured in "This Month in Anesthesiology." Please see this issue of ANESTHESIOLOGY, page 9A.

◆ This article is accompanied by two Editorial Views. Please see: Orkin FK: Risk stratification, risk adjustment, and other risks. ANESTHESIOLOGY 2010; 113:1001–3; Cohen NA, Hannenberg AA: Risk stratification index: An important advance in comparing health care apples to oranges. ANESTHESIOLOGY 2010; 113:1004–6.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are available in both the HTML and PDF versions of this article. Links to the digital files are provided in the HTML text of this article on the Journal's Web site (www.anesthesiology.org).

30-day mortality, and 0.83 *versus* 0.77 for 1-yr mortality. Addition of demographic information only slightly improved performance of the RSI.

Conclusion: RSI is a broadly applicable and robust system for assessing hospital length of stay and mortality for groups of surgical patients based solely on administrative data.

What We Already Know about This Topic

- Hospitals are increasingly required to publicly report outcomes, yet performance is best interpreted in the context of population and procedural risk.
- Good predictive systems that are based on readily accessible data are not currently available.

What This Article Tells Us That Is New

- The authors developed broadly applicable and robust risk-stratification systems for assessing hospital length of stay and mortality for surgical patients based solely on administrative data.

A CENTRAL tenet of national health care quality improvement, as described by the Hospital Quality Alliance,** is public reporting of hospital-level outcome statistics. A critical assumption behind public outcome reporting is that patients (or their insurers) are rational consumers who will choose hospitals reporting superior results and that this, in turn, will serve as an economic incentive for quality improvement in health care. It is thus likely that reported results will have considerable financial impact throughout the healthcare system.

The difficulty with unadjusted outcomes is that baseline patient risk varies considerably. Even for a given procedure performed by the same surgeon in the same hospital, mortality may vary considerably because of preexisting patient demographics, comorbidities, and disease stage.[1,2] Furthermore, some medical and surgical procedures are either substantially less effective or more dangerous than others, resulting in unforeseen complications that have an adverse impact on recovery and survival. Outcomes can thus only be reasonably interpreted in light of baseline population risk, the cumulative impact of procedures, and complications associated with hospital care.

Risk stratification schemes have been developed for selected procedures. In general, they use critical laboratory test results and other disease-specific clinical findings that correlate with outcome. An alternative approach is base stratification on administrative claims databases containing diagnostic and procedure descriptors, with or without readily available patient demographic characteristics.[1,3] One shortcoming of existing risk-stratification methods is inadequate validation across diverse populations, facilities, geographical regions, and various medical procedures. For example, the widely-used Charlson comorbidity index (CCI) was developed in 1987 from fewer than 600 patients.[4] Development of a broadly validated risk-stratification method would permit relevant outcomes, such as duration of hospitalization and mortality, to be fairly compared across healthcare institutions. Availability of an open-source, reproducible method would also foster a more consistent and transparent outcome comparison process. Our goal was to develop risk-adjustment models from a national administrative database from the Centers for Medicare and Medicaid Services (CMS), and to validate performance of the resulting models in a large single-center electronic registry of surgical patients.

Materials and Methods

A dataset was constructed from the 2001–2006 Medicare Provider Analysis and Review (MEDPAR) database (CMS dataset; n = 79,741,480). The MEDPAR file is a national stay-based dataset derived from claims made for payment to CMS under the Medicare program. Each record in the file represents a single patient stay. Data fields include demographic data (age, gender), up to 10 diagnosis codes and 6 procedure codes (coded according to the *International Classification of Diseases, Ninth Revision,*

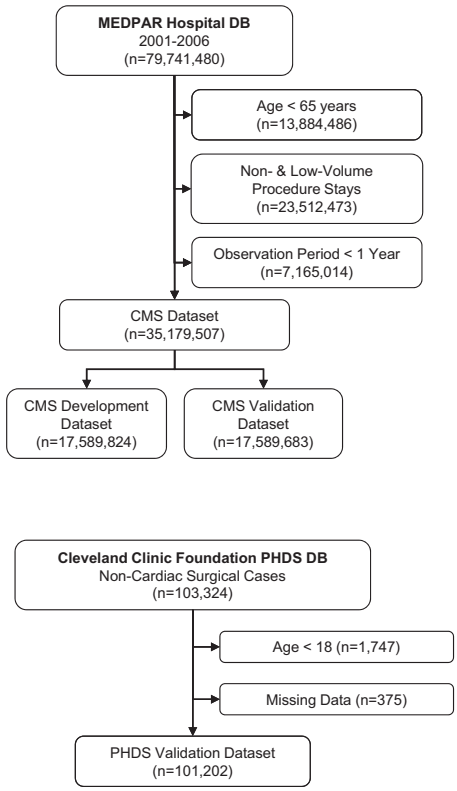


Fig. 1. (*Top*) Trial diagram for MEDPAR dataset analysis (post-discharge mortality and length of stay). (*Bottom*) Trial diagram for validation on the Cleveland Clinic PHDS registry dataset. CMS = Centers for Medicare and Medicaid Services; DB = Database; MEDPAR = Medicare Provider Analysis and Review; PHDS = Perioperative Health Documentation System.

** Hospital Quality Alliance: Quality measures. Available at http://www.hospitalqualityalliance.org/hospitalqualityalliance/qualitymeasures/qualitymeasures.html. Accessed May 11, 2010.

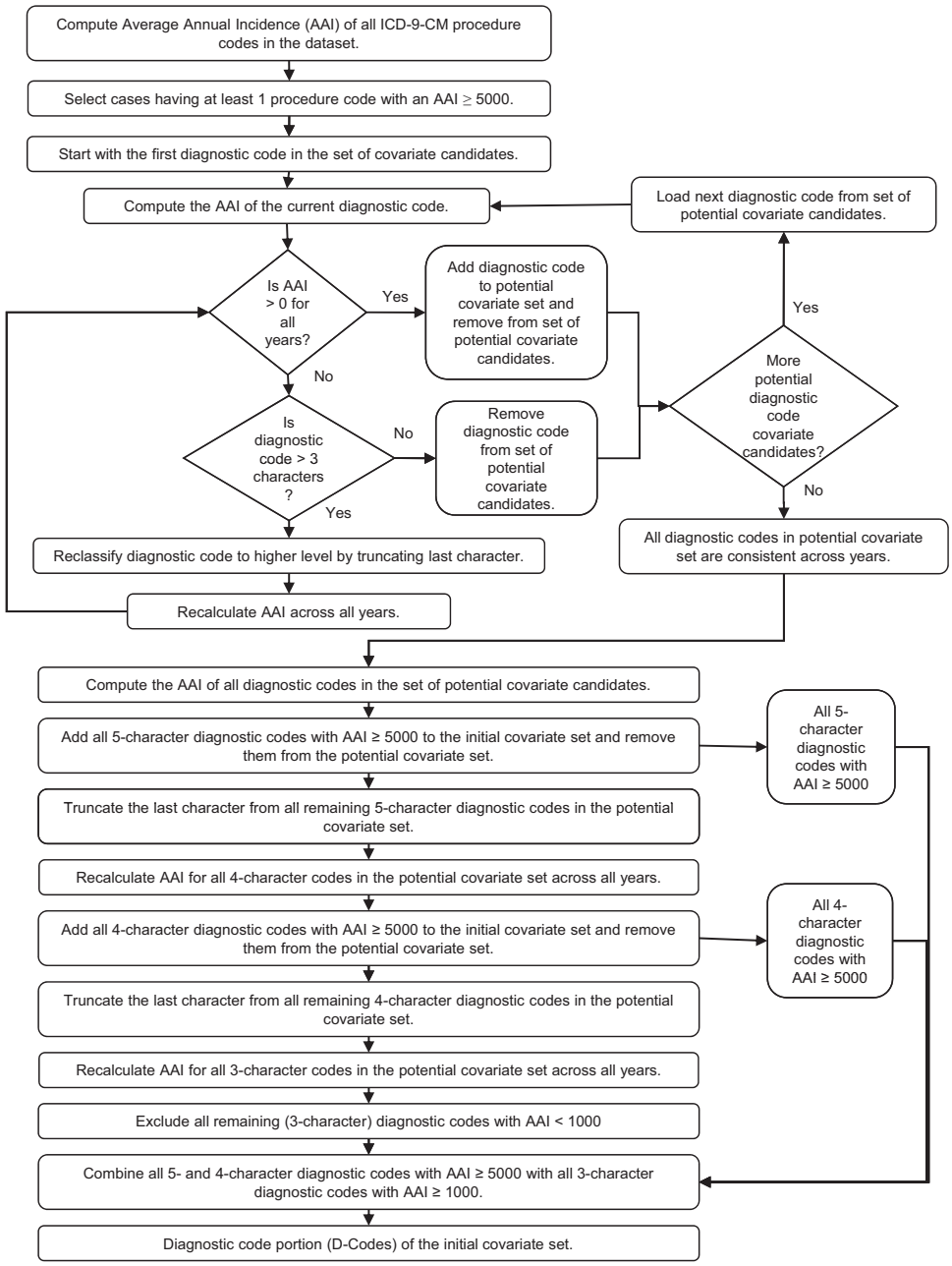


Fig. 2. Selection process for candidate diagnostic codes (D-code). Covariate candidates are made consistent across all years in the dataset and then coded into covariates based on Average Annual Incidence (AAI) criteria. ICD-9-CM = International Classification of Diseases, Ninth Revision, Clinical Modification.

Clinical Modification [ICD-9-CM]), length of stay (LOS), and days from admission to death.

We excluded patients younger than 65 yr, those having no procedure or procedures with an annual average occurrence of less than 5,000, or a patient stay with less than 1 yr of follow-up. The final dataset was randomly divided into development (n = 17,589,824) and validation (n = 17,589,683) datasets (fig. 1). CCI was computed for each patient stay.[4,5]

Our approach was to derive a measure of the risk posed by each patient's comorbidities, jointly with the risk associated with each procedure. Diagnosis and procedure codes (ICD-9-CM) were used to generate the optimum covariate set for modeling each endpoint (LOS, in-patient mortality, and 30-day and 1-yr postdischarge mortality). The ICD-9-CM codes are hierarchical; therefore, it was possible to truncate the codes to a higher level to ensure consistency of the covariates across time to account for new codes and changes in code use (fig. 2). In successive iterations, covariates were selected in a step-wise manner based on the statistical significance of the covariates in a multivariable model (Stepwise Hierarchical Selection). Cox proportional hazards modeling was used to model time to postdischarge death and time to discharge. Because the timing of the diagnostic and procedure codes during

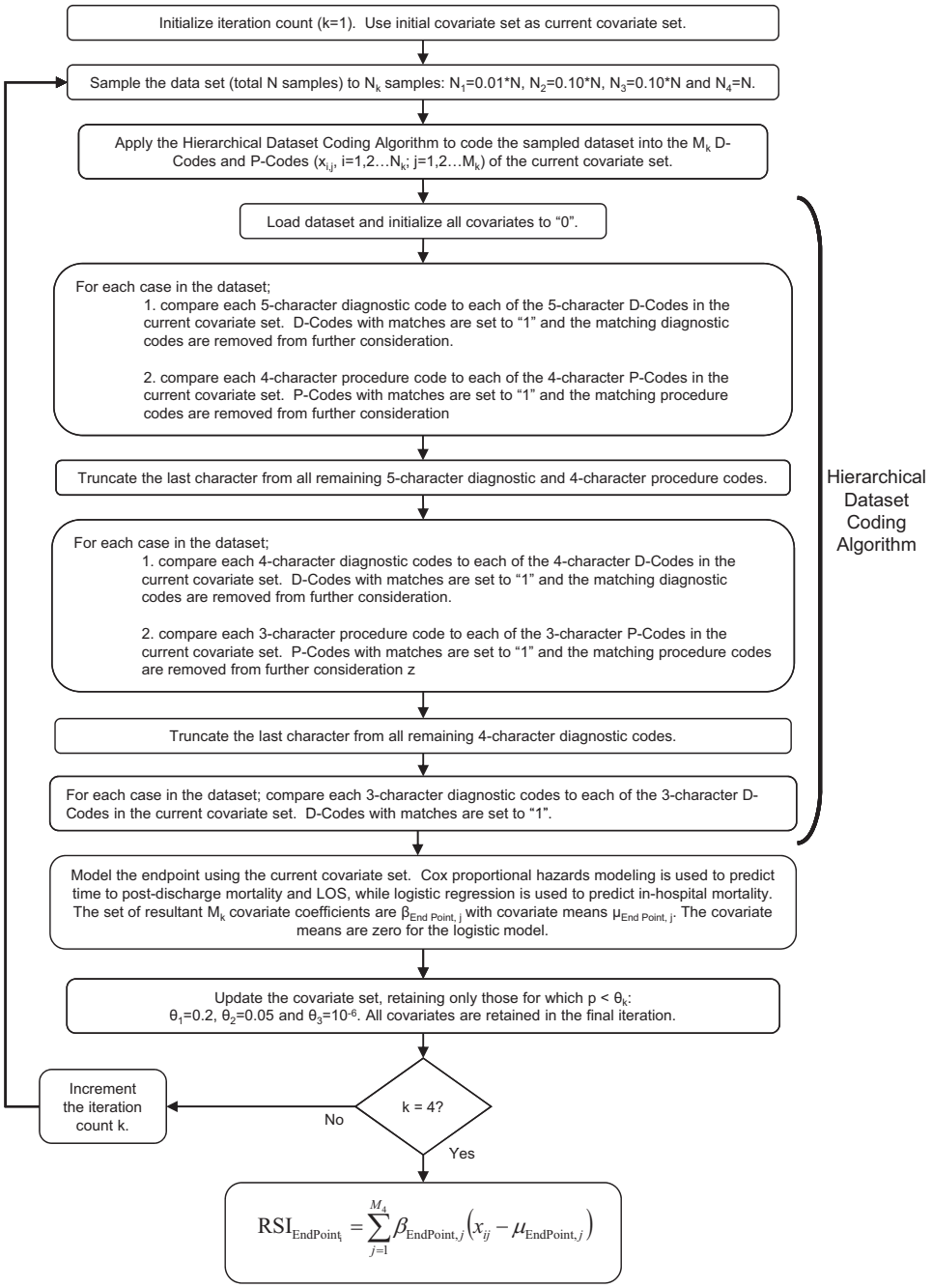


Fig. 3. Stepwise hierarchical selection algorithm: iterative endpoint modeling, using the hierarchical dataset coding algorithm. D-Codes = ICD-9 Diagnostic Codes; LOS = length of stay; P-Codes = ICD-9 Procedure Codes; RSI = risk stratification index.

the hospitalization was unknown, logistic regression was used to model in-hospital mortality.

During early iterations of the Stepwise Hierarchical Selection algorithm (fig. 3), the development dataset was randomly sampled to permit a reasonable execution time. The initial models were developed from a 1% sample of the development dataset (n = 175,898). The second and third iterations, with smaller covariate sets, were based on a 10% sample. The entire development dataset was used for the final iteration. The criteria for including covariates in the first two iterations ($P < 0.2$ and $P < 0.05$) of the algorithm were selected to allow the largest number of likely variables to be identified. The criteria for the third 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. A Hierarchical Dataset Coding algorithm was used to translate the diagnosis and procedure codes from the development dataset into the final covariate set. The Hierarchical Dataset Coding algorithm selectively collapsed diagnosis and procedure codes into binary covariates (0 or 1), as determined by the covariate set.

A Cox or logistic model was used to estimate the hazard associated with each covariate. The initial covariate set included 1,951 variables used for the initial model of each endpoint. The limit of statistical significance applied to the model covariates was P less than 0.2 in the first iteration, P less than 0.05 after the second, and P less than 10^{-6} after the third. The fourth iteration was used to recalculate the final hazard ratios. The final model for each endpoint resulted in a different number of variables: in-hospital mortality,184; 30-day mortality, 240; 1-yr mortality, 503; and LOS, 1,096.

A risk stratification index (RSI) for each of the endpoints of interest was then developed, with RSI_{1YR}, RSI_{30days}, RSI_{INHOSP}, and RSI_{LOS} denoting predictors of 1-yr, 30-day, and in-hospital mortality, and time to discharge within 30 days, respectively . The RSI value for each patient stay was calculated by adding the covariate coefficients associated with the patient's procedure and diagnostic codes linked to the patient stay. The coefficient (β_j) of each covariate calculated by the Cox modeling process was the natural log of the hazard associated with that covariate (or the natural log of the odds ratio change for the logistic model; $\beta_j = \ln(\text{hazard ratio}_j)$. The total hazard arising from a particular patient's diagnostic and procedure codes can be calculated as the exponential sum of the covariate coefficients associated with those codes. Total hazard has a non-Gaussian distribution; it is preferable, therefore, to use RSI as a risk-adjustment factor rather than the total hazard itself.

Prospective validation was initially conducted using the CMS validation dataset. The predictive power of the RSIs were evaluated on both the development and validation datasets using the C statistic (area under the receiver operating characteristic curve) for mortality and median LOS (coded as a binary covariate with values corresponding to either $>$ or $<$ the median) and Harrell's C index (a measure of relative predictive performance) for time to discharge within 30 days.[6] For each dataset, validation was conducted using all hospital stays, and separately for stays including principal procedures likely to require full anesthetic management (for details, see Supplemental Digital Content 1, a table showing likely surgical procedures, http://links.lww.com/ALN/A642). The effect of sample size on the predictive accuracy of RSI was assessed by repeatedly randomly sampling the CMS validation dataset to obtain sets from 100 to 50,000 patient stays. Confidence intervals for the C index were obtained by bootstrapping techniques. Statistical significance was defined as P less than 0.05 for comparisons of C statistics and C indices.

A second prospective validation evaluated the performance of the RSIs on surgical patients from one tertiary medical center. With approval of the Cleveland Clinic Institutional Review Board (Cleveland, Ohio), this validation used the Cleveland Clinic Perioperative Health Documentation System (PHDS), an electronic medical record–based registry of noncardiac surgical patients from January 2005 to December 2009 (n = 103,324). We constructed a dataset from this registry that was structured in the same stay-based format as the MEDPAR dataset, with ICD-9-CM procedure and diagnosis codes. Patients younger than 18 yrs and those with missing data were excluded (n = 2,122). The four RSIs were computed using the Hierarchical Dataset Coding algorithm, and performance was evaluated using the C statistic and C index. The performances of the RSIs were compared with the CCI. We also evaluated whether inclusion of demographic characteristics improved prediction accuracy.

Statistical programming was implemented in SPSS (SPSS Inc., Chicago, IL) and Python (Python Software Foundation, Hampton, NH). The CMS development and validation datasets were evenly split by even or odd record numbers provided by Medicare; random selection for bootstrapping was performed using the SPSS uniform distribution function.

Results

Characteristics of the CMS development and validation datasets did not differ significantly. There were significant differences between the CMS and PHDS validation datasets (table 1). Surgical patients in the PHDS dataset were younger and had fewer comorbidities and lower CCI. PHDS patients had shorter hospital stays and lower mortality rates.

Performance statistics on the CMS development and validation databases are presented in table 2. Performance on the validation dataset was not statistically different from that on the development database, indicating that the degree to which the RSIs predict the endpoints is highly consistent. Performance was significantly better in the population that had surgical procedures that likely required full anesthetic management, possibly because ICD-9 codes are better characterized in the surgical population that usually gets a careful preoperative evaluation. The predictors associated with the highest and lowest hazard ratio for each of the three models are provided in tables 3–5. The complete set of covariates and coefficients are provided in Supplemental Digital Content 2A–C, tables showing predictors and coefficients for each model, http://links.lww.com/ALN/A643.††

Prospective evaluation of the RSIs on the PHDS dataset is presented in table 2. The performance of each RSI was significantly better than CCI, a difference that persisted after the addition of demographic characteristics to both models. Adding demographic characteristics significantly improved RSI performance only for the 1-yr mortality endpoint. RSI_{1YR} predicted mortality at 30 days as well as the independent RSI_{30days} model; the RSI_{1YR} model can thus be equally used for either 30-day or 1-yr postdischarge mortality endpoints.

†† The coefficients for each RSI model are also available in formats suitable for statistical programs at www.clevelandclinic.org/RSI. Accessed September 23, 2010.

Table 1. Dataset Characterization

	CMS Development Set (N = 17,589,824)	CMS Validation Set (N = 17,589,683)	PHDS Validation Set (N = 101,202)
Age, yr	74.1 ± 10.2	74.1 ± 10.2	56.6 ± 16.0
Female, %	54.4	54.4	51.2
White/black/other, %	82.5/12.2/5.4	82.5/12.2/5.4	82.4/12.7/4.9
Surgical, %	29.9	29.9	100
CCI	1.65 ± 1.9	1.65 ± 1.9	1.35 ± 2.0
Number of diagnostic codes	7.1 ± 2.3	7.1 ± 2.3	6.8 ± 2.9
Number of procedure codes	2.6 ± 1.7	2.6 ± 1.7	3.3 ± 2.3
Length of stay, median days [IQR]			
All	5 [3,8]	5 [3,8]	
Surgical	5 [3,8]	5 [3,8]	3 [1,6]
In-hospital mortality, %			
All	5.3	5.3	
Surgical	3.1	3.2	1.3
30-Day mortality (postdischarge), %			
All	5.0	5.0	
Surgical	2.5	2.5	0.5
1-Yr mortality (postdischarge), %			
All	19.2	19.3	
Surgical	10.6	10.7	4.4

Data are presented as mean ± SD unless noted otherwise.

CCI = Charlson comorbidity index; CMS = Centers for Medicare and Medicaid Services; IQR = interquartile range; PHDS = Perioperative Health Documentation System.

Receiver operating characteristic curves for the four endpoints indicate significantly better performance for the RSIs across the sensitivity-specificity range compared with CCI and demographics alone.

Figure 4 shows that superior performance was most evident for in-hospital mortality and median LOS and less pronounced for the remaining endpoints. The ability of CCI to predict 30-day LOS was no better than chance. RSI_{1yr} predictive accuracy appeared stable down to a sample size as small as several thousand hospital stays (fig. 5).

Discussion

Hospital performance measures and public reporting are key methods to drive quality improvement. The validity of comparisons among hospitals depends critically on accurate stratification of population and procedure risk. Furthermore, accurate and universally applicable risk-stratification methods would reduce incentives for hospitals to "cherry pick" healthier patients or perform simpler procedures that might improve their unadjusted outcomes.

Table 2. Retrospective and Prospective Validations

	Mortality (C statistic, 95% CI)			Length of Stay	
	In-Hospital	30-day	1-yr	Median LOS (C statistic, 95% CI)	30-day Discharge (C index, 95% CI)
CMS development dataset					
All cases	0.930 [0.929 0.931]	0.838 [0.834 0.842]*	0.833 [0.832 0.834]	0.865 [0.865 0.865]	0.792 [0.776 0.808]
Surgical cases	0.946 [0.945 0.948]	0.859 [0.858 0.860]*	0.851 [0.850 0.851]	0.896 [0.896 0.897]	0.827 [0.814 0.840]
CMS validation dataset					
All cases	0.930 [0.929 0.931]	0.842 [0.841 0.843]†	0.833 [0.833 0.834]	0.865 [0.865 0.865]	0.792 [0.776 0.808]
Surgical cases	0.946 [0.945 0.947]	0.862 [0.859 0.865]†	0.850 [0.848 0.852]	0.896 [0.896 0.897]	0.828 [0.816 0.841]
PHDS validation dataset					
Demographics	0.684 [0.670 0.698]	0.705 [0.681 0.730]†	0.684 [0.675 0.692]	0.568 [0.564 0.571]	0.571 [0.513 0.632]
CCI	0.654 [0.640 0.669]‡	0.761 [0.738 0.784]†‡	0.767 [0.759 0.775]‡	0.596 [0.592 0.600]‡	0.523 [0.463 0.588]
CCI + Demographics	0.711 [0.697 0.724]	0.803 [0.783 0.823]†‡	0.798 [0.790 0.805]‡§	0.610 [0.606 0.614]‡§	0.575 [0.505 0.640]
RSI	0.977 [0.975 0.980]‡§‖	0.854 [0.834 0.875]†‡§‖	0.832 [0.825 0.839]‡§‖	0.886 [0.883 0.888]‡§‖	0.765 [0.668 0.846]‡§‖
RSI + Demographics	0.979 [0.977 0.981]‡§‖	0.880 [0.863 0.897]†‡§‖	0.855 [0.849 0.861]‡§‖#	0.887 [0.885 0.889]‡§‖	0.774 [0.699 0.849]‡§‖

Demographics are age, sex, and race.

* Predicted by the 30-day mortality model. † Predicted by the 1-yr mortality model. ‡ $P < 0.05$ compared with demographics alone. § $P < 0.05$ compared with CCI. ‖ $P < 0.05$ compared with CCI + demographics. # $P < 0.05$ compared with RSI.

CCI = Charlson comorbidity index; CI = confidence interval; CMS = Centers for Medicare and Medicaid Services; LOS = length of stay; PHDS = Perioperative Health Documentation System; RSI = risk stratification index.

Table 3. Covariates for In-hospital Mortality with the Largest and Smallest Odds Ratios

Diagnostic Codes			Procedure Codes		
Predictor	Description	Odds Ratio [95% CI]	Predictor	Description	Odds Ratio [95% CI]
RSI_{INHOSP}					
D432	Other and unspecified intracranial hemorrhage	9.262 [8.095 10.598]*	P9960	Cardiopulmonary resuscitation, not otherwise specified	57.821 [54.308 61.562]†
D78001	Coma	8.794 [8.066 9.587]†	P3761	Implant of pulsation balloon	5.333 [5.069 5.610]†
D4410	Dissection of aorta	7.357 [6.692 8.087]†	P9604	Insertion of endotracheal tube	3.468 [3.403 3.535]†
D42741	Ventricular fibrillation	7.197 [6.802 7.616]†	P5411	Exploratory laparotomy	3.287 [3.053 3.540]*
D3481	Anoxic brain damage	6.817 [6.425 7.232]†	P370	Pericardiocentesis	3.131 [2.772 3.538]‡
D20500	Acute myeloid leukemia, without mention of having achieved remission	5.806 [5.429 6.210]†	P9390	Non-invasive mechanical ventilation	2.540 [2.462 2.621]†
D853	Other and unspecified intracranial hemorrhage following injury	5.536 [4.887 6.271]§	P9605	Other intubation of respiratory tract	2.530 [2.291 2.794]‡
D801	Fracture of base of skull	5.044 [4.439 5.731]§	P9905	Transfusion of platelets	2.015 [1.934 2.099]*
D5728	Other sequelae of chronic liver disease	4.931 [4.375 5.558]§	P9671	Continuous invasive mechanical ventilation for less than 96 consecutive hours	2.010 [1.968 2.052]†
D852	Subarachnoid, subdural, and extradural hemorrhage, following injury	4.838 [4.580 5.112]†	P3893	Venous catheterization, not elsewhere classified	1.918 [1.894 1.942]†
D4359	Unspecified transient cerebral ischemia	0.574 [0.539 0.612]‡	P0066	Percutaneous transluminal coronary angioplasty [PTCA] or coronary atherectomy	0.453 [0.422 0.487]§
D2724	Other and unspecified hyperlipidemia	0.566 [0.555 0.577]†	P3812	Endarterectomy, other vessels of head and neck	0.375 [0.349 0.402]§
D2720	Pure hypercholesterolemia	0.561 [0.547 0.575]†	P9205	Cardiovascular and hematopoietic scan and radioisotope function study	0.359 [0.323 0.400]‡
D44021	Atherosclerosis of native arteries of the extremities with intermittent claudication	0.555 [0.505 0.609]‖	P8192	Injection of therapeutic substance into joint or ligament	0.356 [0.317 0.400]‡
D7840	Headache	0.460 [0.418 0.507]‡	P0309	Other exploration and decompression of spinal canal	0.338 [0.311 0.369]§
DV5789	Care involving other specified rehabilitation procedure	0.429 [0.409 0.450]*	P6029	Other transurethral prostatectomy	0.300 [0.276 0.327]§
D7812	Abnormality of gait	0.419 [0.386 0.455]‡	P3726	Catheter based invasive electrophysiologic testing	0.294 [0.270 0.320]§
D2113	Benign neoplasm of colon	0.414 [0.390 0.439]§	P0051	Implantation of cardiac resynchronization defibrillator, total system [CRT-D]	0.189 [0.162 0.220]§
D4550	Internal hemorrhoids without mention of complication	0.396 [0.359 0.437]‡	P3794	Implantation or replacement of automatic cardioverter/defibrillator, total system [AICD]	0.161 [0.144 0.180]*
D71595	Osteoarthrosis, unspecified whether generalized or localized, pelvic region and thigh	0.352 [0.317 0.392]‡	P8154	Total knee replacement	0.131 [0.121 0.142]†

† $P < 10^{-300}$, *$P < 10^{-200}$, § $P < 10^{-100}$, ‡ $P < 10^{-50}$, ‖ $P < 10^{-25}$.
CI = confidence interval; RSI_{INHOSP} = risk stratification index, in-hospital mortality.

Risk-stratification systems developed for specific subpopulations may generalize poorly and are thus unsuitable for characterizing all outcomes within a single hospital, much less for comparing among diverse hospitals. There are also outcome prediction systems, such as the National Surgical Quality Improvement Program stratification,[7] that are broad-based but depend on clinical information that is not readily available for all hospitalizations. The National Surgical Quality Improvement Program, for example, depends on highly trained nurse reviewers who collect clinical data from a small fraction of patients at participating centers. These clinical details presumably augment prediction accuracy but are not easily available for other patients, even in National Surgical Quality Improvement Program participating centers, much less for patients in nonparticipating hospitals. Any system

Table 4. Covariates for One-year Post-discharge Mortality with the Largest and Smallest Hazard Ratio

Diagnostic Codes			Procedure Codes		
Predictor	Description	Hazard Ratio [95% CI]	Predictor	Description	Hazard Ratio [95% CI]
RSI_{1YR}					
D191	Malignant neoplasm of brain	5.315 [5.217 5.414]†	P5498	Peritoneal dialysis	1.800 [1.762 1.838]†
D20500	Acute myeloid leukemia, without mention of having achieved remission	4.256 [4.186 4.328]†	P5491	Percutaneous abdominal drainage	1.690 [1.674 1.708]†
D163	Malignant neoplasm of pleura	3.521 [3.388 3.658]†	P462	Ileostomy	1.518 [1.474 1.564]§
D155	Malignant neoplasm of liver and intrahepatic bile ducts	3.231 [3.178 3.285]†	P5198	Other percutaneous procedures on biliary tract	1.436 [1.399 1.474]§
D1579	Malignant neoplasm of pancreas, part unspecified	3.174 [3.124 3.224]†	P4613	Permanent colostomy	1.427 [1.379 1.475]‡
DV66	Convalescence and palliative care	2.970 [2.903 3.039]†	P5011	Closed (percutaneous) [needle] biopsy of liver	1.412 [1.395 1.430]†
D157	Malignant neoplasm of pancreas	2.902 [2.843 2.961]†	P3995	Hemodialysis	1.392 [1.384 1.400]†
D1570	Malignant neoplasm of head of pancreas	2.794 [2.743 2.847]†	P4610	Colostomy, not otherwise specified	1.379 [1.346 1.414]§
D1628	Malignant neoplasm of other parts of bronchus or lung	2.769 [2.730 2.808]†	P9960	Cardiopulmonary resuscitation, not otherwise specified	1.364 [1.332 1.396]§
D1629	Malignant neoplasm of bronchus and lung, unspecified	2.762 [2.742 2.782]†	P3129	Other local excision or destruction of lesion or tissue of lung	1.357 [1.325 1.389]§
D7851	Palpitations	0.602 [0.574 0.631]‡	P656	Bilateral salpingo-oophorectomy	0.265 [0.233 0.302]‡
D6256	Stress incontinence, female	0.584 [0.557 0.613]§	P062	Unilateral thyroid lobectomy	0.259 [0.234 0.287]§
D71535	Osteoarthrosis, localized, not specified whether primary or secondary, pelvic region and thigh	0.577 [0.560 0.595]*	P7051	Repair of cystocele	0.243 [0.214 0.276]§
D3861	Other and unspecified peripheral vertigo	0.569 [0.536 0.604]‡	P4701	Laparoscopic appendectomy	0.226 [0.206 0.249]*
D34690	Migraine, unspecified, without mention of intractable migraine without mention of status migrainosus	0.556 [0.527 0.586]§	P8155	Revision of knee replacement, not otherwise specified	0.212 [0.199 0.225]†
D5921	Calculus of ureter	0.545 [0.528 0.561]†	P7050	Repair of cystocele and rectocele	0.204 [0.186 0.223]*
D220	Benign neoplasm of ovary	0.527 [0.492 0.565]‡	P8363	Rotator cuff repair	0.198 [0.184 0.212]†
D2330	Carcinoma in situ of breast	0.505 [0.461 0.552]‖	P8180	Total shoulder replacement	0.169 [0.152 0.188]*
D71515	Osteoarthrosis, localized, primary, pelvic region and thigh	0.502 [0.466 0.540]‡	P8154	Total knee replacement	0.133 [0.129 0.136]†
D6180	Prolapse of vaginal walls without mention of uterine prolapse	0.412 [0.388 0.436]§	P605	Radical prostatectomy	0.068 [0.061 0.075]†

† $P < 10^{-300}$, *$P < 10^{-200}$, § $P < 10^{-100}$, ‡ $P < 10^{-50}$, ‖ $P < 10^{-25}$.
CI = confidence interval; RSI_{1YR} = risk stratification index, 1-yr mortality.

with potential for broad applicability must therefore be based exclusively on readily available administrative claims data.

We developed broadly applicable empirical models for stratifying postoperative risk that are based on ICD-9 diagnostic and procedure codes and demographic characteristics, information that is standardized, objective, and available for virtually every admitted patient requiring a procedure. Unlike proprietary systems,[8] ours is publicly available and transparent and can thus be applied by any stakeholder to objectively risk-adjust hospital outcomes. Furthermore, this method can be easily updated to reflect evolving coding conventions (*i.e.*, conversion to ICD-10 or introduction of entirely new codes), and can be extended to include other populations and outcomes, such as morbidity and cost of care.

It is noteworthy that demographic characteristics only modestly improved some of our models' predictive accuracy. Including age, weight, sex, and race, for example, improves the C statistic based on ICD-9 codes alone by only ≈0.02 for 1-yr mortality but has no significant impact on the models for in-hospital mortality or LOS. It is thus apparent that risk is better characterized by diagnosis and procedure codes rather than by demographic characteristics including age—a result that is consistent with previous observations.[9]

Table 5. Covariates for Hospital Length-of-stay with the Largest and Smallest Hazard Ratios

Diagnostic Codes			Procedure Codes		
Predictor	Description	Hazard Ratio [95% CI]	Predictor	Description	Hazard Ratio [95% CI]
RSI_{LOS}*					
D29534	Paranoid type schizophrenia, chronic with acute exacerbation	4.439 [4.233 4.654]†	P9427	Other electroshock therapy	6.663 [6.551 6.777]†
D2964	Bipolar I disorder, most recent episode (or current) manic	3.727 [3.610 3.849]†	P9463	Alcohol rehabilitation and detoxification	3.492 [3.379 3.609]†
D2966	Bipolar I disorder, most recent episode (or current) mixed	2.953 [2.836 3.074]†	P9425	Other psychiatric drug therapy	3.229 [3.174 3.286]†
DV63	Unavailability of other medical facilities for care	2.876 [2.766 2.990]†	P5651	Formation of cutaneous uretero-ileostomy	3.113 [3.056 3.171]†
DV5789	Care involving other specified rehabilitation procedure	2.835 [2.822 2.848]†	P9444	Other group therapy	2.727 [2.682 2.771]†
D2957	Schizoaffective disorder	2.800 [2.692 2.913]†	P311	Temporary tracheostomy	2.487 [2.462 2.513]†
D29634	Major depressive affective disorder, recurrent episode, severe, specified as with psychotic behavior	2.768 [2.710 2.828]†	P485	Abdominoperineal resection of rectum	2.438 [2.390 2.487]†
D29633	Major depressive affective disorder, recurrent episode, severe, without mention of psychotic behavior	2.362 [2.322 2.403]†	P3129	Other local excision or destruction of lesion or tissue of lung	2.401 [2.348 2.455]†
DV571	Care involving other physical therapy	2.310 [2.285 2.335]†	P3844	Resection of vessel with replacement, aorta, abdominal	2.330 [2.305 2.356]†
D011	Pulmonary tuberculosis	2.256 [2.191 2.323]†	P8669	Other skin graft to other sites	2.264 [2.238 2.291]†
D6185	Prolapse of vaginal vault after hysterectomy	0.823 [0.814 0.832]*	P0040	Procedure on single vessel	0.807 [0.802 0.811]†
D2729	Unspecified disorder of lipoid metabolism	0.820 [0.811 0.829]*	P9227	Implantation or insertion of radioactive elements	0.801 [0.792 0.811]†
DV72	Special investigations and examinations	0.818 [0.806 0.830]§	P8183	Other repair of shoulder	0.796 [0.783 0.810]§
D7274	Ganglion and cyst of synovium, tendon, and bursa	0.808 [0.795 0.822]§	P5979	Other repair of urinary stress incontinence	0.763 [0.756 0.770]†
D44021	Atherosclerosis of native arteries of the extremities with intermittent claudication	0.764 [0.759 0.769]†	P4023	Excision of axillary lymph node	0.748 [0.737 0.759]†
D794	Nonspecific abnormal results of function studies	0.757 [0.746 0.768]*	P3607	Insertion of drug-eluting coronary artery stent(s)	0.744 [0.741 0.747]†
D2330	Carcinoma in situ of breast	0.743 [0.731 0.755]*	P8166	Percutaneous vertebral augmentation	0.731 [0.722 0.741]†
D1744	Malignant neoplasm of upper-outer quadrant of female breast	0.740 [0.729 0.751]†	P0689	Other parathyroidectomy	0.668 [0.655 0.682]†
D79439	Other nonspecific abnormal results of function study of cardiovascular system	0.665 [0.659 0.671]†	P062	Unilateral thyroid lobectomy	0.652 [0.641 0.663]†
DV5331	Fitting and adjustment of cardiac pacemaker	0.613 [0.604 0.622]†	P8363	Rotator cuff repair	0.637 [0.630 0.644]†

† $P < 10^{-300}$, *$P < 10^{-200}$, § $P < 10^{-100}$. Hazard is inverted to present risk of not being discharged within 30 days.
CI = confidence interval; RSI_{LOS} = risk stratification index, length of stay.

One of the most commonly-used stratification systems, the CCI, was designed to predict 1-yr mortality. We found that our long-term mortality RSI model comparably predicts both 30-day and 1-yr postdischarge mortality more accurately than the CCI, although unsurprisingly, the difference for 1-yr mortality was less. In contrast to longer-term mortality, distinct models were required for the most accurate prediction of LOS and in-hospital mortality. Our models for these acute outcomes were considerably better than the CCI. For example, the C statistic for in-hospital mortality is 0.977 with our RSI *versus* 0.654 with the CCI. We thus present three models that accurately predict four impor-

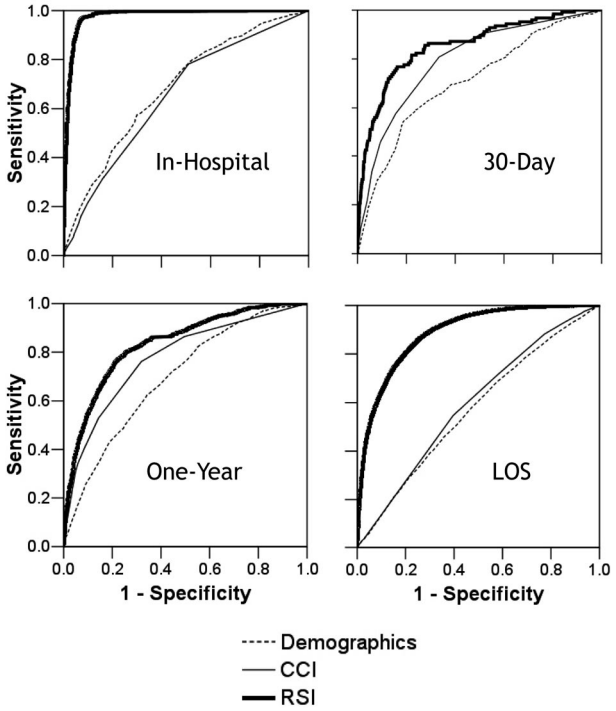


Fig. 4. Receiver operating characteristic curves for length of stay (LOS), and in-hospital, 30-day, and 1-yr mortality. Demographics = age, weight, sex, and race; CCI = Charlson comorbidity index; RSI = risk stratification index.

tant outcomes: LOS and in-hospital, 30-day, and 1-yr mortality.

The RSI models were developed from ≈17 million MEDPAR records and validated on an additional ≈17 million MEDPAR records. It is reassuring that the results were not statistically different, with C statistics typically differing by less than 0.001. However, the more important validation was to apply the RSI model developed from MEDPAR data to the Cleveland Clinic's PHDS database. This was a considerably stricter test because the

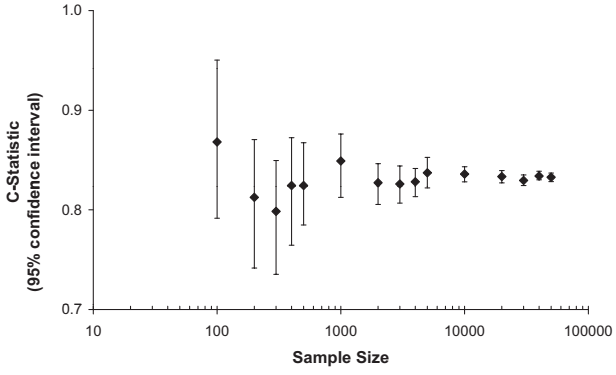


Fig. 5. Bootstrap analysis of the effect of sample size on the estimation of the C statistic for risk stratification index, 1-yr mortality, in the Centers for Medicare and Medicaid Services validation dataset. The analysis was performed by selecting, without replacement, 100 repeated random samples at each sample size. The C statistic estimate and C index was stable down to a sample size of 2,000 admissions.

populations differ in several important ways. For example, the MEDPAR dataset includes all stay-based procedures whether surgical or not, whereas PHDS is surgical. Thus, only approximately 30% of the MEDPAR cases were likely to have been surgical, whereas all the PHDS cases were. Furthermore, the average age of the MEDPAR patients was 18 yr older than in the PHDS population, and only 32% of the PHDS patients were older than 65 yr and thus eligible for Medicare. Finally, the baseline comorbidity, as measured by the CCI and the number of diagnostic codes, was lower in the PHDS patients. Nonetheless, the predictive accuracy of RSI was not statistically different between the MEDPAR and Cleveland Clinic patients, indicating that the RSI system is broadly applicable.

RSI performance appears to remain accurate in samples as small as several thousand hospital stays. This suggests that risk stratification can be used in smaller hospitals or at frequent intervals in larger hospitals.

There are more than 16,000 ICD-9 diagnostic codes and more than 4,500 procedure codes, of which ≈10,000 and ≈3,000, respectively, are in common use. All were considered in development of our risk-stratification models. ICD-9 codes are hierarchical, enabling the "collapsing" of codes to higher (more general) levels. Our method takes advantage of the possibility that marginally predictive codes may increase in predictive power when combined with other related codes because doing so increases the occurrence rate. By first retaining strongly predictive (small P value) individual codes as covariates and then collapsing the remaining codes to create composite covariates with higher occurrence rates, we have derived a highly predictive set of covariates without relying on *a priori* assumptions to create covariates. The result is a set of models that, unlike various proprietary systems, is reproducible and transparent.

Our models include between 184 and 1,096 codes. Although this might appear overly complicated, CMS billing conventions supply up to 16 ICD-9 codes for each patient record. Individual risk for each outcome can thus be determined from a look-up table and simple calculations; however, our results suggest that at least several thousand patients need to be aggregated to produce reliable predictions.

That various baseline characteristics are associated with poor outcome is consistent with clinical intuition. Among the strongest predictors of mortality, for example, were diagnostic codes associated with preexisting malignancy; intracerebral hemorrhage, organic brain syndrome, and heart failure were also strong predictors of 30-day and 1-yr mortality—all with P values less than 10^{-300}.

Less intuitive is that certain baseline characteristics were protective. For example, a diagnosis of hypercholesterolemia reduced the risk of mortality at all time points. In the MEDPAR dataset, 90.8% of patients with a diagnosis of hypercholesterolemia also have a diagnosis of cardiovascular or cerebrovascular disease, which is a strong

predictor of poor outcome. Statin therapy, the primary treatment for hypercholesterolemia, is associated with a reduction in coronary and all-cause mortality as well as major vascular events.[10] It is likely that patients with cardiovascular disease who carried an ICD-9 code for hypercholesterolemia were treated with statins and thus protected relative to patients with cardiovascular or cerebrovascular disease who did not take statins.

Certain surgical procedures were also found to be protective, especially radical prostatectomy. In the MEDPAR dataset, 99.4% of patients undergoing radical prostatectomy had cancer. Malignant neoplasms are among the highest risk factors in our model. Prostatectomy was thus apparently protective compared with patients with cancer who did not have a radical prostatectomy.

These examples show that individual codes cannot be considered in isolation because each patient's risk is determined by the totality of the codes they carry. In other words, our models are predicated on a relative relationship between covariates associated with an underlying risk and diagnoses or procedures associated with treatment that reduces that risk. Furthermore, this relative relationship is based on a MEDPAR record, which consistently includes up to 6 procedure codes and 10 diagnostic codes for each admission. Covariates, therefore, should not be used in isolation or in databases that are not consistent with the MEDPAR stay-based ICD-9-CM format. The general method we present can easily be extended to other administrative record formats and, although similar predictive performance may be achieved, the relative risk associated with specific procedures and diagnoses is likely to vary based on the coding method used.

Use of administrative claims information, including our RSIs, can suffer from regional variations in coding validity or reimbursement gaming.[11,12] But given the penalties for fraudulent coding, it seems unlikely that many hospitals consistently game the system. The contribution of miscoding to our nationally derived models should thus be minimal.

A more serious limitation of our system is that it does not distinguish between *a priori* codes related to baseline health status and planned procedures from actual procedure codes and complications accumulated during hospitalization. The reason is the MEDPAR and most of the PHDS data are derived from claims reports that do not indicate the diagnostic codes present on admission, which reflect baseline patient characteristics, or the principal planned or required procedures as opposed to diagnosis and procedure codes arising from complications during hospitalization. Our system thus assigns risk stratification based on all reported ICD-9 codes, including those that resulted from care-induced complications.

Fortunately, the Agency for Healthcare Research and Quality has published a set of codes usually associated with complications.[13] In a study of two state-wide databases, 92–94% of secondary diagnoses were present on admission, so the contribution of additional in-hospital complication codes might be expected to be limited .[14] It is thus possible to perform risk stratification with and without these "complication codes," which will provide a reasonable distinction between baseline and procedure-related risk *versus* complications associated with hospital care. The RSI covariates associated with the Agency for Healthcare Research and Quality Clinical Classification Software complications‡‡ are denoted in SDC table 2A–C. To further evaluate the effect of in-hospital complications, we backed the risk associated with the Clinical Classification Software complication codes out of the RSI models; the predictive performance of the residual models on the PHDS validation database was not statistically different from RSI, including complications. This lack of significant impact may in part result from the low complication rates encountered at the Cleveland Clinic and theoretically may be greater at other institutions.

In summary, hospitals are increasingly required to publicly report outcomes. However, outcomes can only be reasonably interpreted in the context of baseline-related and procedure-related risk. We thus present three validated RSIs that predict four major outcomes for hospitalization with procedures: LOS and in-hospital, 30-day postdischarge, and 1-yr postdischarge mortality. Our system, RSI, uses only readily available administrative claim codes. It can thus be used to perform risk-adjusted hospital outcomes wherever these claim codes are used to describe patient stays.

The authors gratefully acknowledge the contributions of Eric K. Christiansen, M.B.A. (Senior Operations Analyst, Anesthesia Operations Group, Anesthesiology Institute, Cleveland Clinic, Cleveland, Ohio), who led extraction of data from the Perioperative Health Documentation System.

References

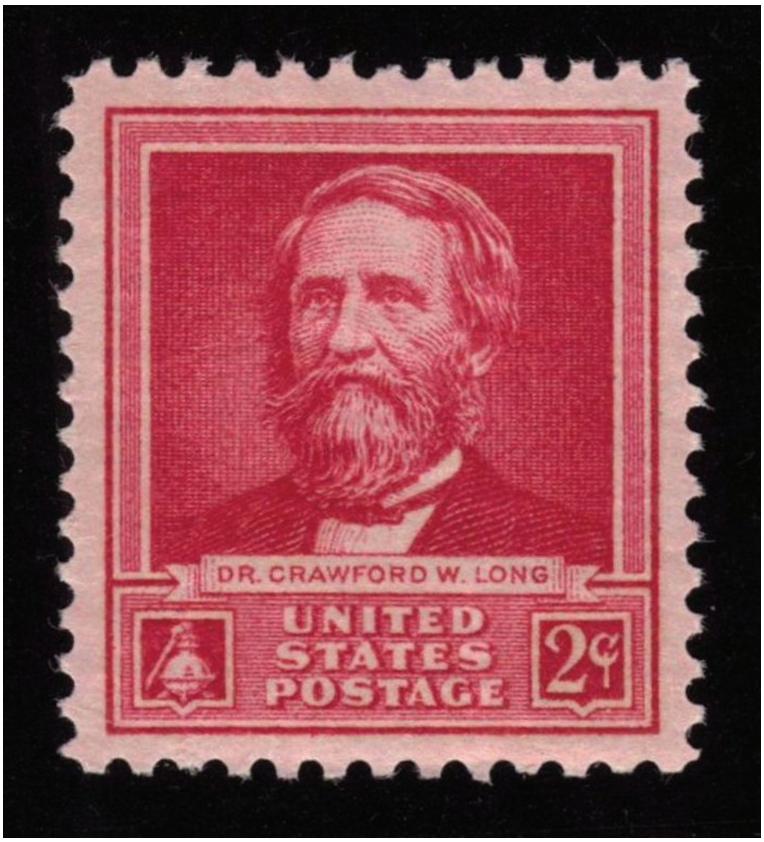
1. Elixhauser A, Steiner C, Harris DR, Coffey RM: Comorbidity measures for use with administrative data. Med Care 1998; 36:8–27
2. Ferraris VA, Ferraris SP, Singh A: Operative outcome and hospital cost. J Thorac Cardiovasc Surg 1998; 115:593–602
3. Iezzoni LI: The risks of risk adjustment. JAMA 1997; 278: 1600–7
4. 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
5. Deyo RA, Cherkin DC, Ciol MA: Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. J Clin Epidemiol 1992; 45:613–9
6. Harrell FE, Jr., Lee KL, Califf RM, Pryor DB, Rosati RA: Regression modelling strategies for improved prognostic prediction. Stat Med 1984; 3:143–52
7. Khuri SF, Daley J, Henderson WG: The comparative assessment and improvement of quality of surgical care in the Department of Veterans Affairs. Arch Surg 2002; 137:20–7
8. Hall BL, Hirbe M, Waterman B, Boslaugh S, Dunagan WC: Comparison of mortality risk adjustment using a clinical data algorithm (American College of Surgeons National Surgical Quality Improvement Program) and an administrative data algorithm (Solucient) at the case level within a single institution. J Am Coll Surg 2007; 205:767–77
9. Cohen MM, Duncan PG, Tate RB: Does anesthesia contribute to operative mortality? JAMA 1988; 260:2859–63

‡‡ Healthcare Cost and Utilization Project (HCUP): HCUP-US Tools & Software Page. Available at: http://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp. Accessed May 10, 2010.

10. Baigent C, Keech A, Kearney PM, Blackwell L, Buck G, Pollicino C, Kirby A, Sourjina T, Peto R, Collins R, Simes R, Cholesterol Treatment Trialists' (CTT) Collaborators: Efficacy and safety of cholesterol-lowering treatment: Prospective meta-analysis of data from 90,056 participants in 14 randomised trials of statins. Lancet 2005; 366:1267–78
11. Horn SD: Validity, reliability and implications of an index of inpatient severity of illness. Med Care 1981; 19:354–62
12. Lorence DP, Ibrahim IA: Benchmarking variation in coding accuracy across the United States. J Health Care Finance 2003; 29:29–42
13. HCUP Clinical Classifications Software (CCS) for ICD-9-CM, Healthcare Cost and Utilization Project (HCUP). Agency for Healthcare Research and Quality, Rockville, MD. 2000–2003
14. Houchens RL, Elixhauser A, Romano PS: How often are potential patient safety events present on admission? Jt Comm J Qual Patient Saf 2008; 34:154–63

ANESTHESIOLOGY REFLECTIONS

The 2-cent Crawford Long Postage Stamp



In Jefferson, Georgia, on what we now celebrate as "Doctors' Day," on March 30 of 1842, pharmacist-physician Crawford Williamson Long (1815–1878) etherized James Venable for removal of a neck tumor. This anesthetic occurred more than 4 yr before (but was publicized 3 yr after) Morton's public ether demonstration in Boston. The U.S. Postal Service used an 1873 photograph of Long as inspiration for engraving its postage stamp honoring the Georgian in 1942. Ironically, since national postal rates that year were 3 cents for a typical enveloped letter, one Crawford Long stamp could only fund local delivery within the town limits of Jefferson. Those seeking postal delivery beyond Jefferson were forced to stick two or more Long stamps on each of their envelopes.

George S. Bause, M.D., M.P.H., Honorary Curator, ASA's Wood Library-Museum of Anesthesiology, Park Ridge, Illinois, and Clinical Associate Professor, Case Western Reserve University, Cleveland, Ohio. UJYC@aol.com.