ANESTHESIOLOGY

Extended-age Out-of-sample Validation of Risk Stratification Index 3.0 Models Using Commercial All-payer Claims

Scott Greenwald, Ph.D., George F. Chamoun, B.S., Nassib G. Chamoun, M.S., David Clain, B.S., Zhenyu Hong, M.S., Richard Jordan, Ph.D., Paul J. Manberg, Ph.D., Kamal Maheshwari, M.D., Daniel I. Sessler, M.D.

ANESTHESIOLOGY 2023; 138:264–73

EDITOR'S PERSPECTIVE

What We Already Know about This Topic

- Risk Stratification Index 3.0 predictive analytical models provide risk profiles at hospital admission from individual administrative claims histories. These models were generated and validated in a population that was mostly more than 65 yr old.

What This Article Tells Us That Is New

- In two different statewide databases, Risk Stratification Index 3.0 models worked well in younger and healthier adults.

ABSTRACT

Background: The authors previously reported a broad suite of individualized Risk Stratification Index 3.0 (Health Data Analytics Institute, Inc., USA) models for various meaningful outcomes in patients admitted to a hospital for medical or surgical reasons. The models used International Classification of Diseases, Tenth Revision, trajectories and were restricted to information available at hospital admission, including coding history in the previous year. The models were developed and validated in Medicare patients, mostly age 65 yr or older. The authors sought to determine how well their models predict utilization outcomes and adverse events in younger and healthier populations.

Methods: The authors' analysis was based on All Payer Claims for surgical and medical hospital admissions from Utah and Oregon. Endpoints included unplanned hospital admissions, in-hospital mortality, acute kidney injury, sepsis, pneumonia, respiratory failure, and a composite of major cardiac complications. They prospectively applied previously developed Risk Stratification Index 3.0 models to the younger and healthier 2017 Utah and Oregon state populations and compared the results to their previous out-of-sample Medicare validation analysis.

Results: In the Utah dataset, there were 55,109 All Payer Claims admissions across 40,710 patients. In the Oregon dataset, there were 21,213 admissions from 16,951 patients. Model performance on the two state datasets was similar or better than in Medicare patients, with an average area under the curve of 0.83 (0.71 to 0.91). Model calibration was reasonable with an R^2 of 0.93 (0.84 to 0.97) for Utah and 0.85 (0.71 to 0.91) for Oregon. The mean sensitivity for the highest 5% risk population was 28% (17 to 44) for Utah and 37% (20 to 56) for Oregon.

Conclusions: Predictive analytical modeling based on administrative claims history provides individualized risk profiles at hospital admission that may help guide patient management. Similar predictive performance in Medicare and in younger and healthier populations indicates that Risk Stratification Index 3.0 models are valid across a broad range of adult hospital admissions.

(ANESTHESIOLOGY 2023; 138:264–73)

The Risk Stratification Index 3.0 (Health Data Analytics Institute, Inc., USA) suite of predictive models is a broadly applicable set of risk adjustment measures that use administrative claims data to predict health outcomes including mortality, prolonged hospitalization, and adverse events during hospitalization and after discharge.[1] This well-validated and calibrated broad suite of predictive algorithms uses diagnostic, procedural, and demographic

This article is featured in "This Month in Anesthesiology," page A1. 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). This article has a visual abstract available in the online version.

Submitted for publication July 25, 2022. Accepted for publication December 12, 2022. Published online first on December 20, 2022.

Scott Greenwald, Ph.D.: Health Data Analytics Institute, Dedham, Massachusetts.

George F. Chamoun, B.S.: Health Data Analytics Institute, Dedham, Massachusetts.

Nassib G. Chamoun, M.S.: Health Data Analytics Institute, Dedham, Massachusetts.

David Clain, B.S.: Health Data Analytics Institute, Dedham, Massachusetts.

Zhenyu Hong, M.S.: Health Data Analytics Institute, Dedham, Massachusetts.

Richard Jordan, Ph.D.: Health Data Analytics Institute, Dedham, Massachusetts.

Paul J. Manberg, Ph.D.: Health Data Analytics Institute, Dedham, Massachusetts.

Kamal Maheshwari, M.D.: Department of Outcomes Research, and Department of General Anesthesiology, Cleveland Clinic, Cleveland, Ohio.

Daniel I. Sessler, M.D.: Department of Outcomes Research, Cleveland Clinic, Cleveland, Ohio.

 Anesthesiology 2023; 138:264–73. DOI: 10.1097/ALN.0000000000004477

information over time to anticipate evolution of health conditions based solely on administrative claims and demographic data available at the time of admission.

A limitation of Risk Stratification Index 3.0 models is that they were developed and out-of-sample validated in Medicare fee-for-service patients who are mostly age 65 yr or older. An obvious question is whether Risk Stratification Index 3.0 models are comparably predictive in other populations, especially those that are younger and healthier. Our goal was to determine how well seven Risk Stratification Index 3.0 models developed in Medicare patients perform when applied to out-of-sample younger and healthier adult populations.

Materials and Methods

As described elsewhere,[1] Risk Stratification Index 3.0 models were developed on the Centers for Medicare & Medicaid Services (Baltimore, Maryland) Research Identifiable File data on a remote server using the SAS Enterprise Guide (version 7.15; SAS Institute Inc., USA) under a Centers for Medicare & Medicaid Services data use agreement (No. 51870). The models are extensions of previously published Risk Stratification Index versions 1.0 and 2.0.[2,3]

Subject Selection

Our reference study design was an extension of the previously described out-of-sample validation in all 2017 to 2019 hospitalized Medicare fee-for-service and dual-eligible (Medicaid and Medicare) beneficiaries.[1] Briefly, in that study, admissions were excluded if patient age on admission was either younger than 18 or older than 99 yr, records had missing or inconsistent data (*e.g.*, missing sex or birthdate information, or had different sex, birth dates, or mortality dates [if applicable] reported in source files), or patients had either discontinuous Part A or Part B Medicare coverage or had Part C coverage in the year before admission (fig. 1A). Claims data during the year before the admission were used to characterize the patient history. Admissions were considered "planned" when designated elective, and were otherwise considered "unplanned." Claims data during the 90 days after admission characterized outcomes. Model performance results from the 2019 Medicare out-of-sample validation cohort were used as the baseline performance comparator for the current study.

Some U.S. states consolidate medical and pharmacy claims data submitted voluntarily by healthcare insurance carriers in what are commonly called All Payer Claims Databases.[4] These registries contain medical and pharmacy claims along with insurance enrollment and health provider data for a large fraction of each state's population. The Utah Office of Health Care Statistics (Salt Lake City, Utah) and the Oregon Office of Health Analytics (Portland, Oregon) have each established progressive and well-defined All Payer Claims data external access programs and were thus selected for analysis. We used the available claims files from 2017, with cases selected as shown in figure 1, B and C.

Our analysis plan was approved by the Utah Department of Health & Human Services Institutional Review Board (Salt Lake City, Utah; No. 544) with a waiver of informed consent requirements. State data were housed on a local server using R software (3.6.0; available at https://cran.r-project.org/src/base, accessed January 6, 2023) under separate data use agreements with each party. Data were handled consistent with our data use agreements, which required suppression of metrics in downloaded tables for populations smaller than 11 individuals.

This report follows the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis reporting guideline.[5]

Outcomes Selection

For our Risk Stratification Index 3.0 validation analysis, we included a suite of 10 models that predict excess length of stay and adverse events, selected to demonstrate performance of predictors for clinically and economically meaningful outcomes spanning a broad range of incidences. Cardiac complications, kidney injury, sepsis, pneumonia, and respiratory failure were defined using International Classification of Diseases, Tenth Revision (ICD-10), diagnosis and procedure codes[6] along with information about their associated claim, such as the setting and revenue center. Additionally, we considered whether codes were primary or secondary.

As reported previously, endpoint definitions were derived using published methods for classifying events using administrative data.[1] Events were identified between admission and discharge (for in-hospital endpoints) and/or between admission and 90 days thereafter (for 90-day endpoints). In-hospital mortality was defined by any-cause death between admission and discharge.

The state datasets did not include sufficient information to determine length of stay, discharge location, or vital status after discharge (for example, date of death). We were therefore unable to compare these three outcomes to the Medicare analysis. The state datasets also did not include race of included subjects, precluding description of the population's racial characteristics.

Model Development

As previously presented, medical history was represented by a set of variables indicating the presence or absence of individual and categories of ICD-10 diagnostic and procedure codes. We used a custom procedure to reduce 69,000 potential ICD-10 diagnostic codes to a representative subset of 4,426 codes by collapsing rare codes into their parent codes to avoid overfitting. ICD-10 diagnostic codes were additionally represented by their corresponding default Clinical Classifications Software Refined category.[7] Similarly, ICD-10 procedure codes were represented by their corresponding default Clinical Classifications Software

category.[8] Temporal information relative to a prediction date was encoded using two sets of these variables representing the presence or absence of relevant codes in the past 90 or 365 days.

Outcomes were indexed to the date of inpatient admission, and claims within the preceding 365 days were included in our models. The only information used from

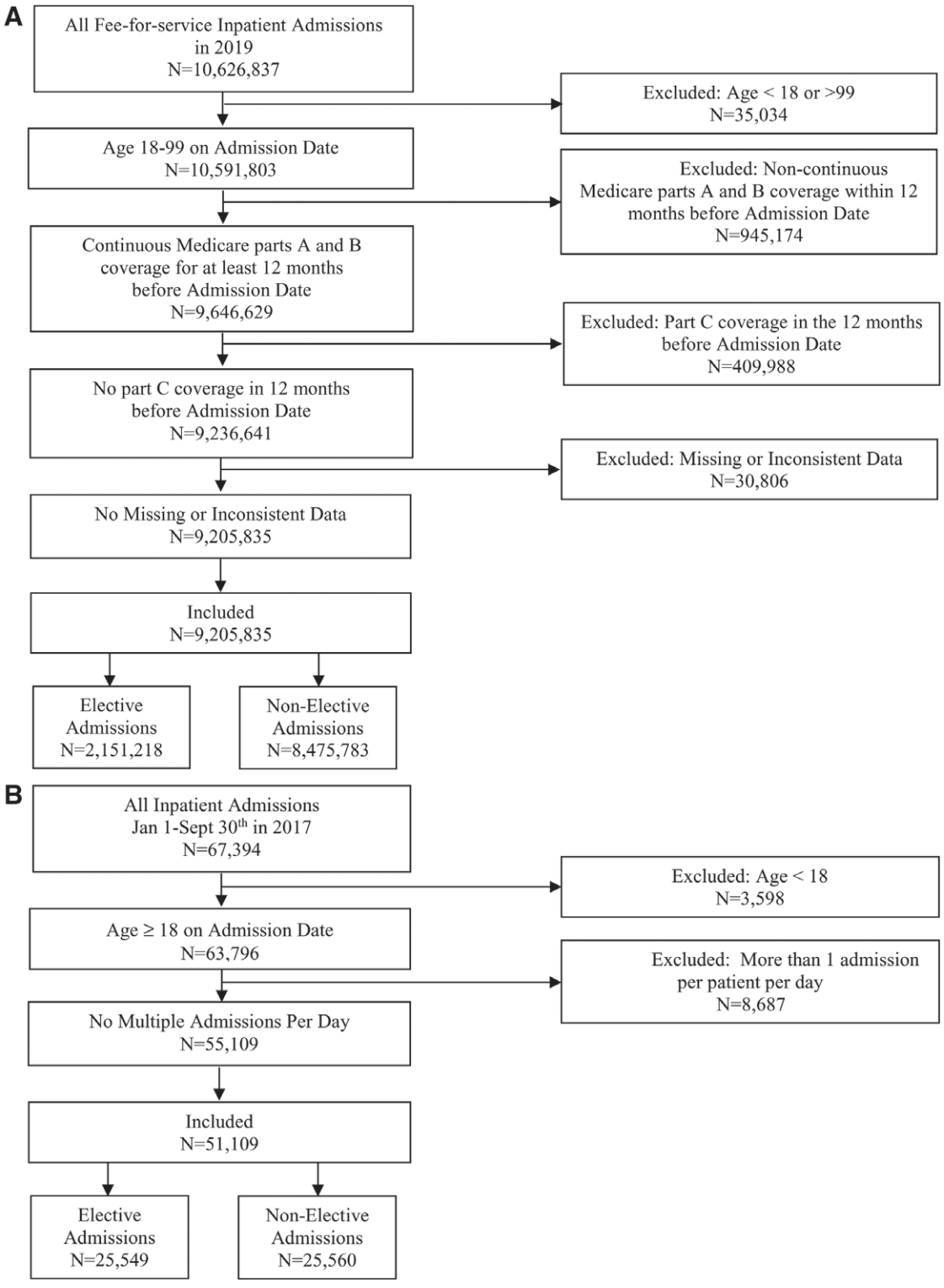

Fig. 1. Cohorts selection diagrams for (*A*) 2019 Medicare dataset; (*B*) Utah All Payer Claims Database dataset; *(continued)*

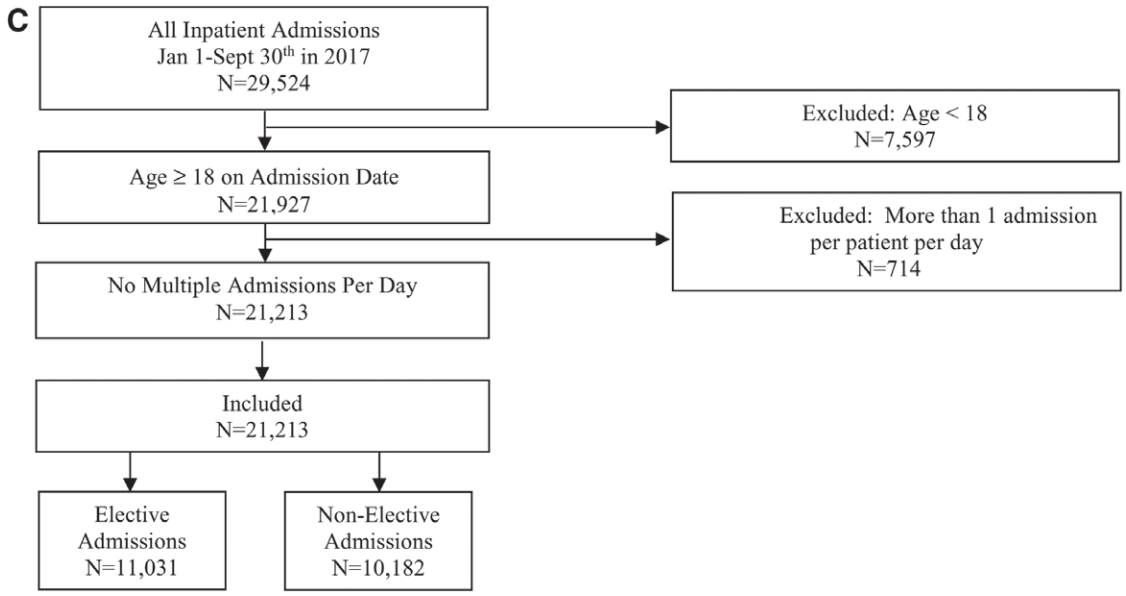

Fig.1. (*continued*) (*C*) Oregon All Payer Claims Database dataset.

the day of admission was the admitting diagnosis (not the principal diagnosis) along with the principal procedure for planned admissions. We also included age at the time of admission. We did not include present-on-admission codes because these are usually coded during or after discharge and are thus not actually available upon admission.

Logistic regression models were trained with the SAS HPLOGISTIC procedure using log–log linkage and backwards fast selection of covariates, keeping those with a $P < 0.01$ significance level. The HPLOGISTIC CODE statement option was used to generate SAS code for subsequent use in a SAS DATA step to apply the model to score new data. We used the asymmetric log–log link function because such models handle skewed extreme value distributions associated with rare events better than symmetrical link functions.[9] There were nonlinear interactions by sex and admission type between ICD-10 codes and various outcomes that precluded using a single logistic model for each outcome. We therefore constructed an overall model for each outcome, designated an ensemble model, that was based on coefficients from four models depending on sex and admission status.

Model Application

Our general approach was to apply our final Medicare-derived models prospectively on each of the two out-of-population state datasets separately to document performance on each of the prospective age validation datasets. Model predictions for each state database were generated in two steps. First, a data file similar in format to that used for model development was created to house the patient demographic and medical history for each admission. Second, the SAS code generated by the SAS HPLOGISTIC procedure during model development was applied to the medical history data file in a SAS DATA step to compute model predictions.

Performance Metrics

Overall discrimination performance was evaluated using the mean and 95% CI for area under the receiver operating characteristics curve (AUC). To compare model detection performance consistently across various endpoints, we compared sensitivity for each model at an alert threshold corresponding to the highest 5% risk fraction of the population.

Calibration performance for each endpoint was evaluated using observed and predicted incidences of subpopulations in bins along the full continuum of risk. We computed R^2 goodness-of-fit values between observed and predicted incidences using all bins having more than 100 subjects. We similarly computed the slope and intercept of the best-fit line and the overall observed-to-predicted ratio. In the primary calibration analysis, we assessed calibration performance using subpopulations in steps of 1% resolution of risk. This approach used assessments of variable-sized populations at equal increments of risk. In a secondary analysis, we assessed calibration performance using decile subpopulations based on risk. This second approach used assessments of similarly sized populations at variable increments of risk.

We *a priori* applied the same minimum acceptable performance criteria as were used in our previous validation study using two metrics to reject clinically nonviable models.

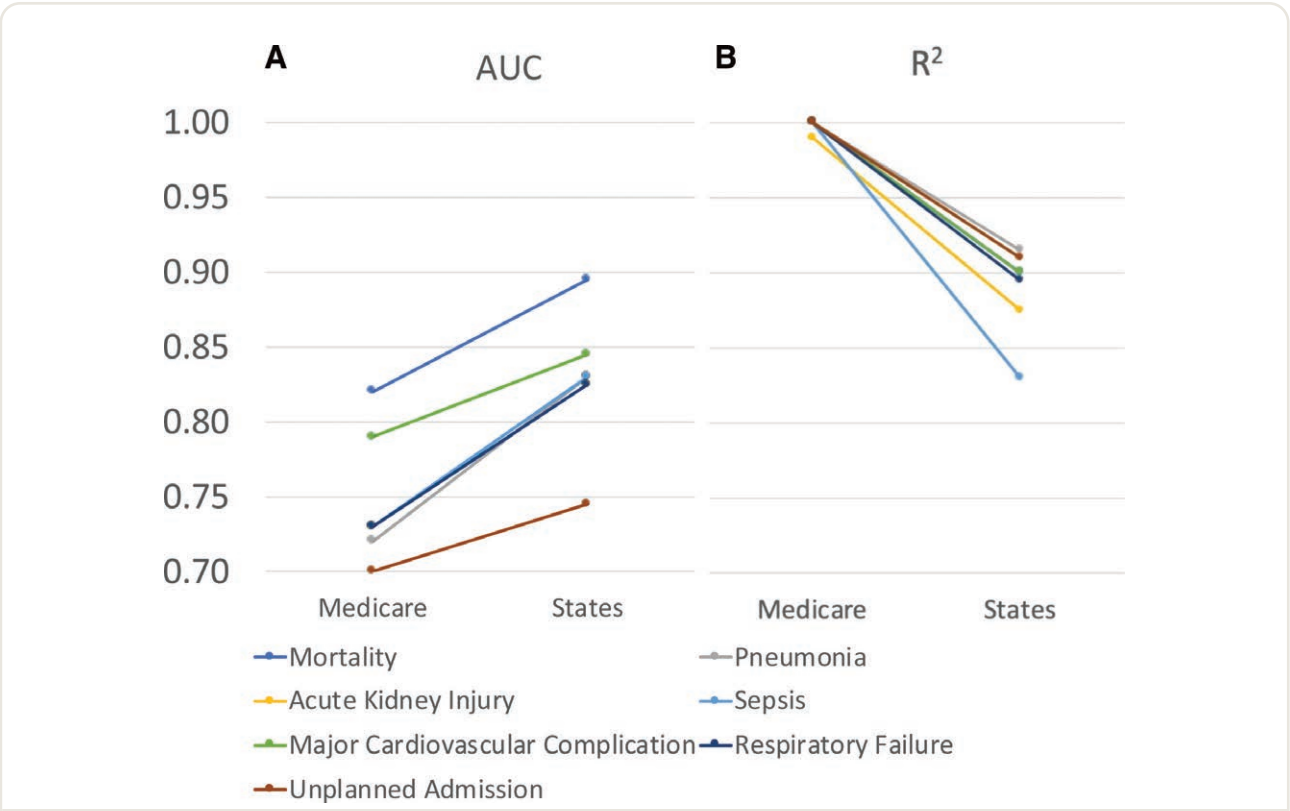

Fig. 2. Performance characteristics on Medicare and combined state admissions. (*A*) Area under the receiver operating characteristics curve (AUC). (*B*) R^2.

Model acceptance required (1) a reasonably accurate overall classification performance defined by an AUC 0.70 or greater; and (2) relatively accurate prediction defined by an observed-to-expected ratio near 1 over the full risk continuum (*i.e.*, calibration R^2 greater than 0.80). The conservative 0.7 minimum acceptance threshold for AUC was based on consultation with clinical advisors and a literature review indicating the acceptability of numerous perioperative machine-learning models with c-statistics in the 0.7 to 0.8 range.[10,11] Because no *a priori* hypotheses were tested, we did not estimate required sample size, but instead used all eligible cases available in the two state files for the selected years.

Results

There were a total of 9,205,835 admissions eligible from 5,336,265 Medicare beneficiaries for analysis in 2019 used previously for model validation. For the Utah dataset, there were 55,109 admissions from 40,710 subjects, and for Oregon, there were 21,213 admissions from 16,951 subjects. The fraction of surgical admissions was 26% in our Medicare population, 26% in Utah's and 24% in Oregon's. As expected, the state populations were younger than the Medicare population, with the mean Medicare age being 74 yr *versus* 58 yr for Utah and 44 yr for Oregon (table 1). As might thus be expected, the incidence of various morbid outcomes was less, confirming that the state populations were also healthier. A consequence of the populations being healthier is that there were fewer diagnostic claims per subject in each of the state datasets.

Prospective classification and primary calibration performance characteristics for binary event predictors in the state datasets are summarized in tables 2 and 3. Secondary calibration performance characteristics are summarized in Supplemental Table 1 (http://links.lww.com/ALN/C1000). Calibration results reported in the main manuscript body are solely from the primary analysis. The observed incidence of endpoints ranged from 0.7% for pneumonia to 10.8% for unplanned readmissions within 90 days and were all considerably less than comparable incidences observed in the Medicare population.

The mean and range of AUCs across all seven outcomes were 0.83 (0.71 to 0.91). The mean and range of the calibration goodness-of-fit (R^2) were 0.89 (0.71 to 0.97). The receiver operator characteristics and calibration curves for each endpoint model for each of the state datasets are provided in the Supplemental Digital Content (http://links.lww.com/ALN/C1000). For the highest 5% risk population, the mean and range of sensitivity were 32% (17 to 56%).

Prediction of short-term (*e.g.*, in-hospital) events frequently outperforms prediction of longer-term (*e.g.*, after discharge) events. As expected, the inpatient mortality

Table 1. Mean and Interquartile of Age, and Percentage of Men, Percentage of Surgical Cases, Percentage of Unplanned Admissions, Average Number of Previous Claims per Admission, and Percentage of Admissions with No Previous Claims within 1 yr of Admission Date for the 2019 Medicare and 2017 Utah and Oregon Datasets.

Characteristic	Medicare Dataset	Utah Dataset	Oregon Dataset
No. of admissions	9,205,835	55,109	21,213
Surgical cases, % (No.)	26 (2,428,690)	26 (16,679)	24 (5,306)
No. of Individuals	5,336,265	40,710	16,951
Age, mean [interquartile range]	74 [68–83]	58 [35–76]	44 [31–58]
Age > 65 yr, % (No.)	80.9 (7,444,716)	43.1 (23,722)	0.5 (103)
Men, % (No.)	46 (4,267,427)	36 (20,060)	33 (7,092)
Unplanned admissions, % (No.)	80 (1,846,670)	46 (25,560)	48 (10,182)
Number of claims within 1 yr before admission, mean [interquartile range]	77 [32–101]	40.1 [16–48]	43.9 [17–55]
Admissions with no claims within 1 yr before admission, % (No.)	0 (10,648)	0 (0)	0 (0)

Neither state provided race information. The percentage of surgical cases was derived by first establishing a pairing between each principal diagnosis code and its most common admission type as defined by the Diagnostic Related Group (*i.e.*, medical or procedural/surgical case) using the Medicare Dataset, and then applying the mapping to identify the surgical cases in the three datasets.

predictive model showed the best performance across the Medicare (AUC, 0.82), Utah (AUC, 0.89) and Oregon (AUC, 0.90) datasets, with especially high sensitivity at the top 5% threshold (Medicare, 30.1%; Utah, 44.1%; Oregon, 56.5%). The calibration metrics, however, were lower in the state datasets because of endpoint detection issues as explained in the Discussion section (Utah R^2, 0.93; Oregon R^2, 0.85) than in the Medicare population (R^2, 1.00). Calibration results were generally similar whether assessed using subpopulations at percent increments of risk in the primary analysis (table 3) or using decile subpopulations based on risk in the secondary analysis (Supplemental Table 1, http://links.lww.com/ALN/C1000).

Five 90-day adverse event models (pneumonia, acute kidney injury, sepsis, major cardiovascular complications, and respiratory failure) all showed a similar pattern of somewhat higher state AUC (0.80 to 0.86 *vs.* 0.72 to 0.79) and sensitivity (20 to 47% *vs.* 16 to 24%) performance results relative to the Medicare population (fig. 2). The model for 90-day unplanned admissions showed the lowest, but still acceptable, performance across all datasets (AUC: 0.70 to 0.78; sensitivity: 11.6% to 19.9%).

Discussion

We present the performance of Risk Stratification Index 3.0 predictive models for seven endpoints when applied prospectively to two large out-of-sample state datasets comprised of younger and healthier subjects who are more representative of the overall adult U.S. population requiring hospital admission. For all the endpoints, performance measures including AUC, R^2, and sensitivity at the top 5% risk all were similar or better than the performance obtained on just the Medicare population. These results support the conclusion that our models are robust and widely applicable to the broad U.S. adult population including relatively young and health subjects. One factor contributing to Risk Stratification Index 3.0 models being so robust is that they were trained on more than 18 million hospital admissions and validated on more than 9 million out-of-sample Medicare admissions.

An intriguing observation is that models built on an older Medicare population generally performed as well or better when applied to younger and healthier subjects. One explanation may be that older individuals often have concomitant conditions that may contribute to risk in complex ways, whereas younger patient usually have fewer chronic conditions and thus a statistically cleaner risk profile linkage to the primary admission diagnosis. For example, we observed that the average Medicare beneficiary had an average of 77 medical claims recorded in the year before admission, whereas patients in Utah had only 40, and Oregon patients had only 44.

We also note that the incidence of the seven adverse outcomes described in this study was much lower in the younger and healthier state populations. The Oregon dataset showed the lowest event rates, consistent with the lower average age of that population. This dataset also showed slightly better performance metrics, thereby suggesting a complex interaction between lower age, fewer events, and better model performance metrics. However, this pattern is insufficiently consistent to conclude with certainty that these predictive models work best on young, otherwise healthy populations.

The state datasets we used differed from the Medicare database in a number of important ways. For example, the Oregon dataset did not (based on policy) include any Medicare-eligible subjects, and both datasets contained a sizeable number of pregnancy-related admissions. Furthermore, neither state database is curated as well as the Medicare registry. We confirmed the presence of several data quality issues first identified in a 2017 report by The Agency for Healthcare Research and Quality,[12] most notably the presence of duplicate records and missing data fields. We also identified coding differences, including an

Table 2. Classification Performance of Risk Stratification Index Models on Out-of-sample Medicare and State Admissions

		Incidence, % (No.)			AUC (95% CI)			Sensitivity at Top 5% Threshold (95% CI)		
Period	**Endpoint**	**Medicare**	**Utah**	**Oregon**	**Medicare**	**Utah**	**Oregon**	**Medicare**	**Utah**	**Oregon**
In-hospital	Mortality	2.8 (259,144)	1.2 (634)	0.8 (170)	0.82 (0.82–0.82)	0.89 (0.88–0.90)	0.90 (0.88–0.93)	30.1 (29.9–30.3)	44.1 (40.2–48.0)	56.5 (49.0–64.0)
	Discharge to facility	35.9 (3,308,992)	N/A	N/A	0.79 (0.79–0.79)	N/A	N/A	12.2 (12.2–12.2)	N/A	N/A
90 days after admission	Pneumonia	3.9 (363,078)	2.6 (1,427)	0.7 (143)	0.72 (0.72–0.72)	0.83 (0.82–0.84)	0.83 (0.80–0.87)	18.5 (18.4–18.6)	30.1 (27.7–32.5)	46.9 (38.7–55.1)
	Acute kidney injury	5.5 (503,672)	5.8 (3,217)	1.6 (335)	0.73 (0.73–0.73)	0.80 (0.80–0.81)	0.85 (0.84–0.87)	16.3 (16.2–16.4)	20.1 (18.7–21.5)	34.0 (28.9–39.1)
	Sepsis	5.8 (534,402)	3.2 (1,786)	1.6 (333)	0.73 (0.73–0.73)	0.81 (0.80–0.82)	0.85 (0.83–0.87)	18.7 (18.6–18.8)	27.9 (25.8–30.0)	36.6 (31.4–41.8)
	Major cardiovascular complication	6.0 (551,203)	2.9 (1,591)	1.4 (300)	0.79 (0.79–0.79)	0.83 (0.82–0.84)	0.86 (0.85–0.88)	24.1 (24.0–24.2)	30.2 (27.9–32.5)	41.3 (35.7–46.9)
	Respiratory failure	6.3 (582,587)	4.4 (2,447)	1.7 (360)	0.73 (0.73–0.73)	0.81 (0.80–0.82)	0.84 (0.82–0.86)	20.0 (19.9–20.1)	26.7 (24.9–28.5)	36.4 (31.4–41.4)
	Unplanned admission	28.0 (2,578,303)	10.8 (5,931)	9.7 (2,051)	0.70 (0.70–0.70)	0.71 (0.70–0.72)	0.78 (0.77–0.79)	11.6 (11.6–11.6)	16.8 (15.8–17.8)	19.9 (18.2–21.6)
Overall mean (95% CI)	Including discharge to facility	11.8 (3.5–20.1)	N/A	N/A	0.75 (0.72–0.78)	N/A	N/A	18.9 (15.0,22.9)	N/A	N/A
	Excluding discharge to facility	8.3 (1.8–14.8)	4.4 (2.1–6.8)	2.5 (0.1–4.9)	0.75 (0.71–0.78)	0.81 (0.77–0.85)	0.84 (0.82–0.87)	19.9 (15.4–24.3)	28.0 (21.5–34.5)	36.8 (30.4–47.2)

Metrics include endpoint, observation window (either in-hospital or within 90 days after admission), incidence, area under the receiver operating characteristics curve (AUC), and sensitivity. Endpoints include acute kidney injury, sepsis, respiratory failure, unplanned hospitalization, discharge to facility status, major cardiovascular event, mortality, and pneumonia. Endpoints in the table are ordered by increasing incidence within their corresponding event period (*i.e.*, in-hospital or 90 days after admission). AUC, area under the receiver operating characteristics curve. N/A, not available.

Table 3. Calibration Performance of Risk Stratification Models on Out-of-sample Medicare and State Admissions

		Calibration Metrics* R^2			Calibration Estimates (Intercept/Slope)		
Period	**Endpoint**	**Medicare**	**Utah**	**Oregon**	**Medicare**	**Utah**	**Oregon**
In-hospital	Mortality	1.00	0.93	0.87	0.00/0.94	–0.01/0.90	0.00/0.92
	Discharge to facility	1.00	N/A	N/A	0.00/1.00	N/A	N/A
90 days after admission	Pneumonia	1.00	0.97	0.86	0.00/0.89	–0.02/2.28	–0.01/0.84
	Acute kidney injury	0.99	0.84	0.91	0.01/0.87	0.03/1.64	–0.01/0.89
	Sepsis	1.00	0.95	0.71	0.01/0.92	–0.01/1.75	–0.01/0.90
	Major cardiovascular complication	1.00	0.95	0.85	0.01/0.93	–0.01/1.75	0.00/1.23
	Respiratory failure	1.00	0.93	0.86	0.01/0.90	–0.02/2.11	–0.01/0.77
	Unplanned admission	1.00	0.95	0.87	0.00/0.98	–0.05/1.16	–0.04/1.02
Overall mean (95% CI)	Including discharge to facility	1.00 (1.00 to 1.00)	N/A	N/A	0.01 (0.00 to 0.01)/ 0.93 (0.90 to 0.96)	N/A	N/A
	Excluding discharge to facility	1.00 (1.00 to 1.00)	0.93 (0.90 to 0.96)	0.85 (0.80 to 0.89)	0.01 (0.00 to 0.01)/ 0.92 (0.89 to 0.95)	–0.01 (–0.03 to 0.00)/ 1.66 (1.29 to 2.02)	–0.01 (–0.02 to 0.00)/ 0.94 (0.83 to 1.05)

Statistics are calculated using observations per percent risk of adverse event. Metrics include endpoint, observation window (either in-hospital or within 90 days after admission), R^2 goodness-of-fit, estimates of intercept and slope of best fit regression line. Endpoints include acute kidney injury, sepsis, respiratory failure, unplanned hospitalization, discharge to facility status, major cardiovascular event, mortality, and pneumonia. Endpoints in the table are ordered by increasing incidence within their corresponding event period (*i.e.*, in-hospital or 90 days after admission).

*Calibration metrics: R^2 goodness-of-fit between observed and expected incidences among subpopulations using bins of 0.01 resolution of predicted risk, excluding the top 1% of subjects with highest risk. Slope and intercept: estimates of the regression line coefficients (*i.e.*, best fit line between actual and expected observations).

unexplained near absence of certain codes (such as for dialysis) in the Utah state database. The absence of codes used to identify endpoints impacts their apparent incidence (*e.g.*, missing dialysis codes associated with chronic kidney disease are unavailable to exclude false detection of acute kidney injury, resulting in the inflation of the apparent incidence of acute kidney injury). This inflation in turn degrades the "actual"-to-expected performance in those datasets. Taking the example of acute kidney injury in the Utah dataset, the apparent incidence is inflated to nearly twice that observed in a similar Utah population in the Medicare database (results not shown). Consequently, the calibration plot demonstrates that the rank ordering of the prediction is tightly preserved as desired, although the slope of the calibration curve markedly differs from unity. Although database limitations might make these state registries unsuitable for developing reliable new predictive models, they nonetheless provided a rigorous external validation test set.

Our overall calibration results met our prespecified acceptance criteria; however, additional refinements in the calibration of some models may be considered to optimize performance in datasets comprised of populations with observed event rates different than those in the development Medicare set. Our results indicate that our models work reasonably well even with the sort of imperfect datasets that might be encountered with clinical implementation.

A limitation of our analysis is that we were unable to evaluate our models for discharge destination, excess length of stay, and 90-day mortality because requisite data were not included in the state registries. While we present model validations applicable to broad U.S. adult populations, we did not include children. Children differ substantially from adults in rarely having serious chronic conditions. Furthermore, they are hospitalized for different reasons. Our models should therefore be properly validated in pediatric patients and refined as necessary for this special population. Future research will be needed to document applicability to other sources of diagnostic and procedural histories, such as electronic medical records, registries, health information exchanges, or institutional data warehouses. The potential impact of COVID-19 and associated disruptions in healthcare delivery on model performance must also be assessed in future research. International Classification of Disease coding in the United States is generally reliable since it is guided by well-enforced federal regulations. Less rigorous application would degrade Risk Stratification Index 3.0 predictions.

Our model validation consisted of prospectively testing previously developed models on two out-of-sample state datasets and shows that predictions exceeded our prespecified minimum acceptable performance standards. We did not employ either a "non-inferiority" or a "superiority" design because our goal was to determine whether the models can be applied successfully to other populations—especially younger and healthier patients who better represent typical U.S. hospitalized patients. Our results indicate

that they can. Although this work externally validates the models, it is currently not clear if the application of the models will be feasible across all settings in real time or, if applied, improves either efficiency or patient outcomes. We are currently testing deployment of these models for a number of clinical applications to address these open questions.

Fortunately, there are current and impending ways to electronically acquire a patient's billing record to help near real-time implementation of the models. A number of recent rulings have driven development and adoption of tools that enable and access to claims data (https://www.federalregister.gov/documents/2020/05/01/2020-05050/medicare-and-medicaid-programs-patient-protection-and-affordable-care-act-interoperability). Existing application programming interfaces enable payer-to-patient access (*e.g.*, Blue Button technology [https://bluebutton.cms.gov/]), payer-to-provider access (*e.g.*, Beneficiary Claims Data application programming interfaces for accountable care organization access [https://bcda.cms.gov]; Data at the Point of Care or Physician access [https://dpc.cms.gov, in pilot stage]), and provider-to-provider access (*e.g.*, Epic's Care Everywhere [https://www.epic.com/interoperability/ehr-interoperability-from-anywhere]). These application programming interfaces and the electronic health records of the local institution (for the current medical history) allow users to access or construct the claims stream for our predictors and clinical support software. We anticipate that these policies will help drive access to claims data from multiple additional payer organizations and facilitate more widespread practical access to the predictive models based on administrative claims.

As an example, the models described in this manuscript are undergoing field testing at several major institutions, including the Cleveland Clinic (Cleveland, Ohio). While code latency could theoretically lead to underprediction of risk, real-time implementation of the models permits code feeds from multiple sources such as the Beneficiary Claims Data application programming interfaces (https://bcda.cms.gov), which are updated weekly, or directly from local EPIC sources, which can update faster. Future research is needed to define whether using Risk Stratification Index 3.0 predictive modeling at admission improves clinical efficiency and patient outcomes.

In summary, we demonstrate that a suite of predictive Risk Stratification Index 3.0 models developed using a very large population of Medicare fee-for-service beneficiaries, mostly older than 65 yr, also performs well when applied prospectively to two large out-of-sample state datasets that include younger and healthier subjects.

Acknowledgments

The authors thank Thomas DeRito, B.S. (Health Data Analytics Institute, Dedham, Massachusetts), for technical assistance in querying the state databases and conducting certain statistics.

Research Support

Funded by the Health Data Analytics Institute (Dedham, Massachusetts).

Competing Interests

G. F. Chamoun, N. G. Chamoun, D. Clain, Z. Hong, and Drs. Greenwald, Manberg, and Jordan are employees of the Health Data Analytics Institute (Dedham, Massachusetts). Dr. Sessler is a consultant and shareholder in the company. Dr. Maheshwari declares no competing interests.

Correspondence

Address correspondence to Dr. Sessler: Department of Outcomes Research, Anesthesiology Institute, Cleveland Clinic, 9500 Euclid Ave — L1-407, Cleveland, Ohio 44195. DS@OR.org. This article may be accessed for personal use at no charge through the Journal Web site, www.anesthesiology.org.

Supplemental Digital Content

Supplemental Digital Content, http://links.lww.com/ALN/C1000

Supplemental Figure 1. Performance characteristics: in-hospital mortality.

Supplemental Figure 2. Performance characteristics: 90-day pneumonia.

Supplemental Figure 3. Performance characteristics: 90-day acute kidney injury.

Supplemental Figure 4. Performance characteristics: 90-day sepsis.

Supplemental Figure 5. Performance characteristics: 90-day cardiovascular complications.

Supplemental Figure 6. Performance characteristics: 90-day respiratory failure.

Supplemental Figure 7. Performance characteristics: 90-day unplanned admission.

Supplemental Table 1. Calibration Performance of Risk Stratification Models on Out-of-sample Medicare and State Admissions (Risk Decile Groups).

References

1. Greenwald S, Chamoun GF, Chamoun NG, Clain D, Hong Z, Jordan R, Manberg PJ, Maheshwari K, Sessler DI: Risk Stratification Index 3.0, a broad set of models for predicting adverse events during and after hospital admission. Anesthesiology 2022; 137:673–86
2. 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

3. Chamoun GF, Li L, Chamoun NG, Saini V, Sessler DI: Validation and calibration of the Risk Stratification Index. Anesthesiology 2017; 126:623–30
4. Agency for Healthcare Research and Quality. All-Payer Claims Databases. Available at https://www.ahrq.gov/data/apcd/index.html. Accessed June 1, 2022.
5. Collins GS, Reitsma JB, Altman DG, Moons KG: Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): The TRIPOD statement. Ann Intern Med 2015; 162:55–63
6. Centers for Medicare & Medicaid Services. International Classification of Diseases, Tenth Revision. Available at https://www.cms.gov/medicare/coding/icd10. Accessed February 28, 2021.
7. Agency for Healthcare Research and Quality. Clinical Classification Software Refined (CCSR). Available at https://www.hcup-us.ahrq.gov/toolssoftware/ccsr/ccs_refined.jsp. Accessed August 21, 2021.
8. Agency for Healthcare Research and Quality. Clinical Classification Software ICD-10-PCS (beta version). Available at https://www.hcup-us.ahrq.gov/toolssoftware/ccs10/ccs10.jsp. Accessed August 21, 2021.
9. Van der Paal B: A Comparison of Different Methods for Modelling Rare Events Data, Department of Applied Mathematics, Computer Science and Statistics. Ghent, University of Ghent, 2014, pp 70
10. Centers for Disease Control and Prevention. ICD-10-CM Official Coding and Reporting Guidelines April 1, 2020 through September 30, 2020. Available at https://www.cdc.gov/nchs/data/icd/COVID-19-guidelines-final.pdf. Accessed February 16, 2021.
11. Bellini V, Valente M, Bertorelli G, Pifferi B, Craca M, Mordonini M, Lombardo G, Bottani E, Del Rio P, Bignami E: Machine learning in perioperative medicine: A systematic review. J Anesth Analg Crit Care 2022; 2:2
12. Bardach N LG, Wade E, Dean M, Shultz E, McDonald K, Dudley RA: Final Report: All-Payer Claims Databases Measurement of Care: Systematic Review and Environmental Scan of Current Practices and Evidence (report AHRQ publication No. 17-0022-2-EF). Available at https://www.ahrq.gov/sites/default/files/publications/files/envscanlitrev.pdf. Accessed June 1, 2022.