ANESTHESIOLOGY

Trusted Evidence: Discovery to Practice®

2022
December

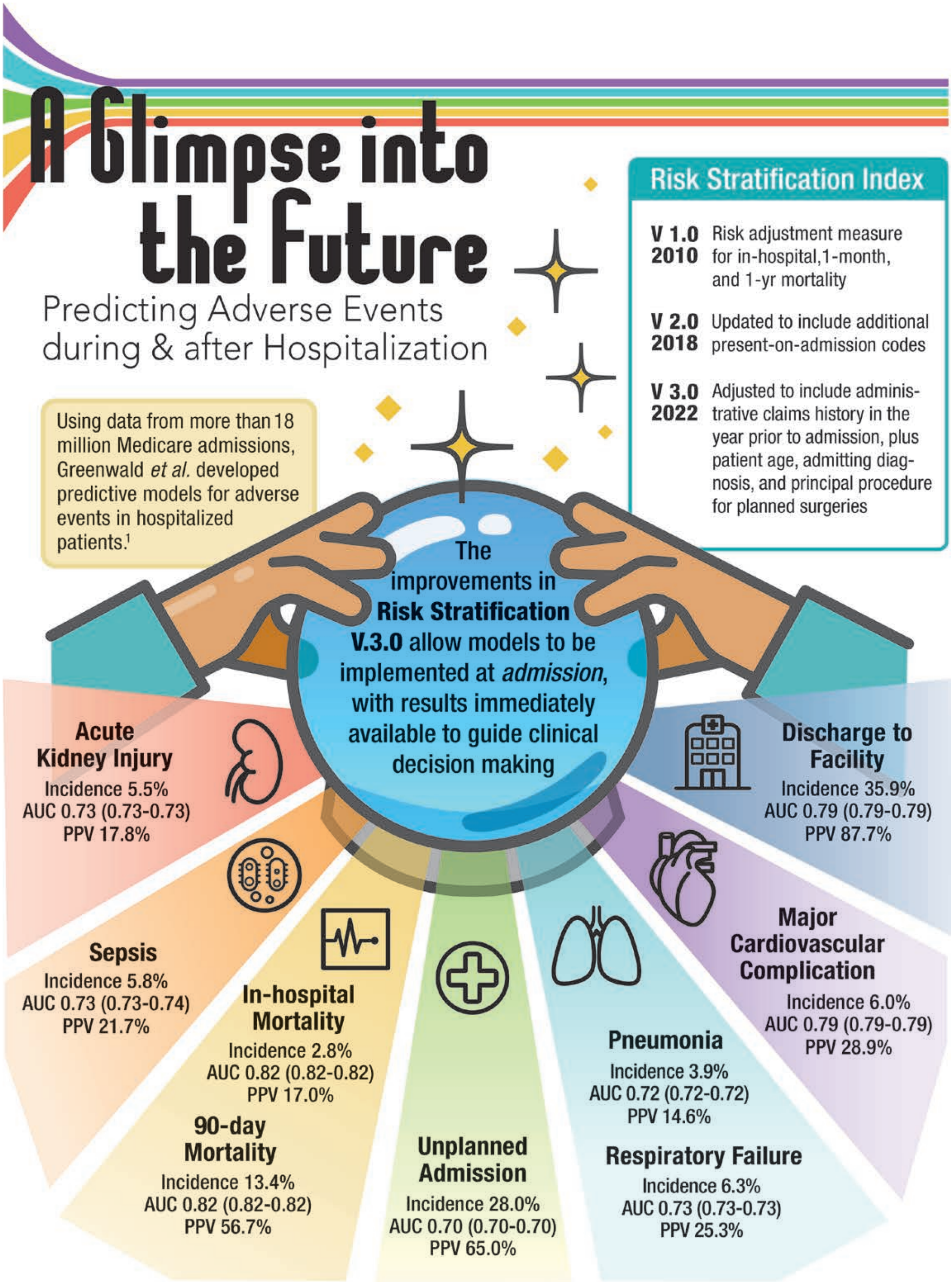
Models for Predicting Adverse Events during and after Hospital Admission

Volume 137
Number 6
anesthesiology.org

The Official Journal of the American Society of Anesthesiologists

INFOGRAPHICS IN ANESTHESIOLOGY

Complex Information for Anesthesiologists Presented Quickly and Clearly



A Glimpse into the Future

Predicting Adverse Events during & after Hospitalization

Using data from more than 18 million Medicare admissions, Greenwald *et al.* developed predictive models for adverse events in hospitalized patients.[1]

Risk Stratification Index

V 1.0 2010 Risk adjustment measure for in-hospital,1-month, and 1-yr mortality

V 2.0 2018 Updated to include additional present-on-admission codes

V 3.0 2022 Adjusted to include administrative claims history in the year prior to admission, plus patient age, admitting diagnosis, and principal procedure for planned surgeries

The improvements in **Risk Stratification V.3.0** allow models to be implemented at *admission*, with results immediately available to guide clinical decision making

Acute Kidney Injury
Incidence 5.5%
AUC 0.73 (0.73-0.73)
PPV 17.8%

Sepsis
Incidence 5.8%
AUC 0.73 (0.73-0.74)
PPV 21.7%

In-hospital Mortality
Incidence 2.8%
AUC 0.82 (0.82-0.82)
PPV 17.0%

90-day Mortality
Incidence 13.4%
AUC 0.82 (0.82-0.82)
PPV 56.7%

Unplanned Admission
Incidence 28.0%
AUC 0.70 (0.70-0.70)
PPV 65.0%

Pneumonia
Incidence 3.9%
AUC 0.72 (0.72-0.72)
PPV 14.6%

Respiratory Failure
Incidence 6.3%
AUC 0.73 (0.73-0.73)
PPV 25.3%

Major Cardiovascular Complication
Incidence 6.0%
AUC 0.79 (0.79-0.79)
PPV 28.9%

Discharge to Facility
Incidence 35.9%
AUC 0.79 (0.79-0.79)
PPV 87.7%

AUC, area under the receiver operating curve; PPV, positive predictive value; V, version.

Positive predictive value for all models was calculated for the top 5% risk groups for each adverse event.

Infographic created by Holly B. Ende, Vanderbilt University Medical Center; James P. Rathmell, Brigham and Women's Health Care/Harvard Medical School; and Jonathan P. Wanderer, Vanderbilt University Medical Center. Illustration by Annemarie Johnson, Vivo Visuals Studio. Address correspondence to Dr. Ende: holly.ende@vumc.org.

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

ANESTHESIOLOGY

Risk Stratification Index 3.0, a Broad Set of Models for Predicting Adverse Events during and after Hospital Admission

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 2022; 137:673–86

EDITOR'S PERSPECTIVE

What We Already Know about This Topic

- Uses for risk stratification tools include setting baselines for health service evaluations, identifying patients who may need higher levels of care, and allocating hospital resources

What This Article Tells Us That Is New

- From a dataset of more than 9 million patients, a risk score based on administrative claims history was developed to provide individualized risk profiles at hospital admission that may help guide patient management

ABSTRACT

Background: Risk stratification helps guide appropriate clinical care. Our goal was to develop and validate a broad suite of predictive tools based on International Classification of Diseases, Tenth Revision, diagnostic and procedural codes for predicting adverse events and care utilization outcomes for hospitalized patients.

Methods: Endpoints included unplanned hospital admissions, discharge status, excess length of stay, in-hospital and 90-day mortality, acute kidney injury, sepsis, pneumonia, respiratory failure, and a composite of major cardiac complications. Patient demographic and coding history in the year before admission provided features used to predict utilization and adverse events through 90 days after admission. Models were trained and refined on 2017 to 2018 Medicare admissions data using an 80 to 20 learn to test split sample. Models were then prospectively tested on 2019 out-of-sample Medicare admissions. Predictions based on logistic regression were compared with those from five commonly used machine learning methods using a limited dataset.

Results: The 2017 to 2018 development set included 9,085,968 patients who had 18,899,224 inpatient admissions, and there were 5,336,265 patients who had 9,205,835 inpatient admissions in the 2019 validation dataset. Model performance on the validation set had an average area under the curve of 0.76 (range, 0.70 to 0.82). Model calibration was strong with an average R^2 for the 99% of patients at lowest risk of 1.00. Excess length of stay had a root-mean-square error of 0.19 and R^2 of 0.99. The mean sensitivity for the highest 5% risk population was 19.2% (range, 11.6 to 30.1); for positive predictive value, it was 37.2% (14.6 to 87.7); and for lift (enrichment ratio), it was 3.8 (2.3 to 6.1). Predictive accuracies from regression and machine learning techniques were generally similar.

Conclusions: Predictive analytical modeling based on administrative claims history can provide individualized risk profiles at hospital admission that may help guide patient management. Similar results from six different modeling approaches suggest that we have identified both the value and ceiling for predictive information derived from medical claims history.

(Anesthesiology 2022; 137:673–86)

This article is featured in "This Month in Anesthesiology," page A1. This article is accompanied by an editorial on p. 661. This article has a related Infographic on p. A17. 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 an audio podcast. This article has a visual abstract available in the online version. The work presented in this article has been presented at the American Society of Anesthesiologists Annual Meeting in New Orleans, Louisiana, October 22, 2022.

Submitted for publication March 24, 2022. Accepted for publication August 31, 2022. Published online first on September 21, 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.

<zdoi;. DOI: 10.1097/ALN.0000000000004380>

Risk stratification tools are useful in at least four distinct situations. The first is health services research. Specifically, comparisons among various facilities or treatment groups can only fairly be evaluated after adjusting for baseline risk of the outcomes of interest. For example, health services evaluations comparing mortality among various hospitals must adjust for baseline mortality risk and procedural complexity across the relevant populations.[1] Accurate risk stratification similarly contributes to observational research by providing an accurate basis for propensity matching and multivariable regression. The second role for risk stratification is to identify enriched populations for care pathways and clinical trials[2]—that is, selecting patients most likely to experience an adverse event and benefit from specific interventions. The third role for risk stratification is to guide clinical care, including decisions about which surgical or alternative treatment options are most likely to prove helpful.[3] Finally, reliable predictions of expected duration of hospitalization and discharge disposition can help guide hospital resource management and planning for follow-up support services.[4]

The Risk Stratification Index version 1.0, first introduced in 2010, was a broadly applicable risk adjustment measure for predicting mortality in-hospital, at 1 month, and at 1 yr.[5] The original models were derived from more than 35 million Medicare hospitalizations between 2001 and 2006, and were thereafter validated in a wider age range of California inpatients[6] and in two single-center studies.[5,7,8] The index also performed well in an independent set of 39 million Medicare admissions from 2008 to 2012.[9] Version 2.0 of the Risk Stratification Index,[10] introduced in 2018, used the expanded set of International Classification of Diseases, Ninth Revision, codes and information that Medicare now allows for each admission, including up to 25 diagnostic codes, 25 procedure codes, and flags indicating conditions that were present on admission. Both Risk Stratification Index versions provided better discrimination than the Charlson Comorbidity Index and other publicly available stratification systems based on administrative data.[5,10] Version 2.0 was also well calibrated.[9] While these versions have been used for academic research,[11–16] we are not aware that they are being used for clinical care.

Previous versions of the Risk Stratification Index were based exclusively on diagnosis and procedure codes from the index hospitalization, and relied heavily on present-on-admission codes. The difficulties with this approach are that billing codes and present-on-admission flags are usually generated by specialized coders after patients are discharged. Consequently, key information necessary for accurate risk stratification is generally unavailable at the time of hospital admission—when stratification may be especially useful. An additional consequence of basing stratification on a single hospitalization is that temporally restricted information fails to capture individuals' preadmission illness trajectories, which might improve predictions. Another limitation of previous versions of the Risk Stratification Index is that they were based on International Classification of Diseases, Ninth Revision, codes, rather than International Classification of Diseases, Tenth Revision, codes, which are now universally used. Previous versions were also restricted to surgical admissions rather than also considering medical admissions. Finally, previous Risk Stratification Index models were restricted to in-hospital, 30-day, and 1-yr mortality, along with hospital length of stay.

Our primary goal was therefore to develop and validate a broad suite of practical analytic tools based on International Classification of Diseases, Tenth Revision, diagnostic and procedural code histories for predicting hospital utilization outcomes and adverse events for both surgical and medical inpatient admissions. Specifically, we derived predictors for meaningful utilization endpoints including unplanned hospital admissions, discharge status, and excess length of stay, along with major adverse events and complications including in-hospital and 90-day mortality, acute kidney injury, sepsis, pneumonia, respiratory failure, and a composite of major cardiac complications.

As in previous versions of the Risk Stratification Index, we primarily used logistic regression because the method provides easily interpretable results including model coefficients that identify key drivers of risk and quantify their relative contributions. However, machine learning methods have become increasingly popular and have shown better predictive perioperative performance than clinical scoring systems in some but not all settings.[17–19] We therefore developed analogous models using five commonly used machine learning methods and compared model performance characteristics with each method.

Materials and Methods

Our primary analyses were conducted on the Centers for Medicare and Medicaid Services (Baltimore, Maryland) Research Identifiable File data on a remote server under a Centers for Medicare and Medicaid Services data use agreement (No. 51870). Access to the remote server was provided through VM Horizon Client (5.3; VMware, Inc, USA) and analysis conducted using SAS (9.04; SAS Institute, USA) within SAS Enterprise Guide (7.15; SAS Institute). Secondary (robustness) analyses were conducted on a 5% sample (Limited Data Set) of the same Medicare dataset provided by the Centers for Medicare and Medicaid Services housed on a local server using R software (version 4.2.0; available at https://cran.r-project.org/src/base/release 2022-04-22) under a separate data use agreement (LDSS-2017-51396). Data were handled consistent with our data use agreements, which required suppression of metrics in downloaded tables for populations smaller than 11 individuals.

Our analyses were determined to be exempt from informed consent requirements by the New England Institutional Review Board (Needham, Massachusetts). This report follows the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis reporting guideline.[20]

The bulk of our statistical analysis plan was submitted to the Centers for Medicare and Medicaid Services in response to their Artificial Intelligence Health Outcome Challenge in February 2020, before formal analysis began. The plan included primary use of logistic regression, specific endpoints, the metrics to be reported, reporting results at outcome incidence and at the top 5%, and *a priori* definitions of adequate model performance. A comparison to various machine learning methods was anticipated, although the specific methods were not prespecified. Lift (enhancement ratio) was not part of the original plan, and was added during the analysis, which was conducted from October 2021 to January 2022.

Subject Selection

We used the full Medicare fee-for-service and dual-eligible (Medicaid and Medicare) files for beneficiaries hospitalized in 2017 to 2019. Admissions were excluded if patient age on admission was 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. Claims data during the year before the admission were used to characterize the patient history. Claims data during the 90 days after admission characterized outcomes. The admission status was classified as "planned" if the reason for admission was elective, and otherwise designated "unplanned."

Outcomes Selection

We present 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, diagnosis and procedure codes[21] along with information about their associated claim, such as the setting and revenue center. Additionally, we considered whether codes were primary or secondary.

Endpoint definitions were derived using published methods for classifying events using administrative data. Event definitions and their associated references are presented in Supplemental Table 1 (http://links.lww.com/ALN/C923). Events were identified between admission and discharge (for in-hospital endpoints) and/or between admission and 90 days thereafter (for 90-day endpoints). A 90-day observation window was chosen for events and mortality because previous reports suggest that 90-day outcomes may be more reliable than 30-day outcomes for measuring hospital performance.[22–24] In-hospital mortality was defined by any-cause death between admission and discharge. Ninety-day mortality was defined by death between admission and 90 days thereafter. Excess length of stay was defined as the difference between the observed duration of hospitalization and the geometric mean length of stay associated with the default 2021 v1 Clinical Classifications Software Refined[25] category for the admitting diagnosis when the admission was unplanned, or the 2020 Clinical Classifications Software[26] category associated with the primary procedure for planned admissions. Discharge status to a facility was defined by discharge to locations other than home, with or without organized home health care.[27]

Model Development

Medical history was represented by a set of variables indicating the presence or absence of individual and categories of International Classification of Diseases, Tenth Revision, diagnostic and procedure codes. We used a custom procedure to reduce 69,000 potential International Classification of Diseases, Tenth Revision, diagnostic codes to a representative subset of 4,426 codes by collapsing rare codes into their parent codes to avoid overfitting (Supplemental Figure 1, http://links.lww.com/ALN/C923). International Classification of Diseases, Tenth Revision, diagnostic codes were additionally represented by their corresponding default Clinical Classifications Software Refined category.[25] Similarly, International Classification of Diseases, Tenth Revision, procedure codes were represented by their corresponding default Clinical Classifications Software category.[26] 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 the day of admission was the admitting diagnosis (rather than the principal diagnosis) along with the principal procedure for planned admissions. We also included age at the time of 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. 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.[28] There were nonlinear interactions by sex and admission type between International Classification of Diseases, Tenth Revision, codes and various outcomes which preclude using a single logistic model for each outcome. We therefore constructed an overall process to apply the appropriate model coefficients from the ensemble of four models depending on sex and admission status.

Model Application

Our general approach was to (1) train a model on 80% of qualifying Medicare admissions from 2017 and 2018 from the development dataset (training set); (2) apply the resulting model to the remaining 20% of the development dataset (test set) to document modeling robustness; and (3) prospectively evaluate the resulting final model on out-of-sample admissions from 2019 to document model validation (prospective validation set).

Performance Metrics

Overall discrimination performance was evaluated using the area under the receiver operating characteristics curves (AUC). Performance at a given operating threshold was assessed by sensitivity, positive predictive value, and lift.[29] Lift is defined as the ratio of detected events using the classifier relative to not using the classifier, which is equivalent to the positive predictive value divided by the incidence. When predictive classifiers are used to identify an enriched subpopulation, lift therefore quantifies the enrichment ratio.

To compare model detection performance consistently across various endpoints, we compared sensitivity, positive predictive value, and lift for each model at an alert threshold corresponding to the highest 5% risk fraction of the population. This sort of high-risk threshold might be used clinically to identify subpopulations most likely to benefit from intervention. We similarly compared sensitivity, positive predictive value, and lift for each model at thresholds corresponding to the observed incidence for each endpoint within the population. Evaluating detector performance using endpoint-specific thresholds normalizes performance results, thereby simplifying performance comparisons across various endpoints and modeling methods. We did not try to identify thresholds that optimize positive and negative predictive values because optimizing depends upon the relative costs of false detections and missed events, which are specific to individual endpoints and use cases. Specifically, selection of the most appropriate thresholds represents a form of resource constrained ranking,[12] where selection of a particular threshold to define a "higher-risk" subgroup from a ranked population is based on a particular use case.

One hundred bins in steps of 1% resolution of risk were used to identify subpopulations along the full continuum of risk. To evaluate calibration, we computed correlation coefficients between observed and predicted incidences for each endpoint using all bins having more than 100 subjects. We similarly computed observed-to-predicted ratios. For the excess length of stay model, the only nonbinary event in the suite, we estimated the root mean squared error of the absolute predictions *versus* the observed values in groups of 1,000 individuals. We computed the mean and 95% CI for AUC.

We *a priori* set conservative minimum acceptable performance criteria using two metrics to reject clinically nonviable models. 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.[18,30] Because no *a priori* hypotheses were tested, we did not estimate required sample size, and instead used all eligible cases available in the Medicare fee-for-service files for the selected years. To evaluate the importance of endpoint-specific models, we compared incidence of various complications in patients selected for having the highest 5% risk of 90-day mortality to those with the highest 5% risk of specific complications.

Model Comparisons

In addition to our primary models, which were developed using multivariable regression, we developed models based on five machine learning methods including random forest, boosting, rule-based, and deep learning (neural network). The Centers for Medicare and Medicaid Services computing environment, which must be used for the 100% Medicare sample, does not provide advanced machine learning tools. Consequently, our machine learning models were based on a 5% Limited Data Set sample, which we were able to host locally. The models were developed using R software (4.2.0; available at https://cran.r-project.org/src/base/ release 2022-04-22). The methods we explored were the following:

1. Ranger Random Forest (RangerRF, using R package RANGER 0.13.1; available at https://cran.r-project/src/contrib/Archive/ranger/released 2021-07-14)[31,32]
2. Extreme Gradient Boosting (XGBoost, using R package XGBOOST 0.90.02; available at https://cran.r-project/src/contrib/Archive/xgboost/released 2019-08-01)[33]
3. Combination Gradient Boosting with Random Forests (XGBoostRF, using R package XGBOOST 0.90.02)[33]
4. RuleFit (RuleFit, using R H20 package 3.36.1.2; available at https://cran.r-project/src/contrib/Archive/h20/ released 2022-05-28)[34]
5. Automated Machine Learning (autoML, using R H20 pacakge 3.36.1.2)[35]

A reference model on the local system was created using logistic regression (using R package GLMNET 3.0;

available at https://cran.r-project/src/contrib/Archive/glmnet/ released 2019-11-09).

Logistic and machine learning models were developed and evaluated on the 5% sample database using methods similar to those employed for the primary analyses, with the following exceptions:

1. The development set was restricted to 2018 admissions.
2. The set of features available for modeling were restricted to a subset of the International Classification of Diseases, Tenth Revision, variables corresponding to the 400 having the highest predictive power per endpoint as identified using Extreme Gradient Boosting (maximum depth, 4; maximum rounds, 200).
3. Separate models were developed for each endpoint, one each for unplanned and planned admissions. Sex was included among the top 400 features.

Specifics on model development for each method are detailed in the following sections.

RangerRF. RangerRF models (with minimum node size of 8 and maximum tree depth of 50) were identified as those with the highest out-of-bag AUC when optimized over a grid where the number of trees ranged over 1,000, 2,500, and 4,000, and the splitting method was either Gini or Hellinger.

XGBoost. Utilizing a 75%/25% learn/test random split of the development database, XGBoost models were identified as those with the highest test set AUC when optimizing hyperparameters over a grid where the learning rate ranged over 0.1, 0.25, and 0.4, the maximum tree depth ranged over 2, 4, and 6, the fraction of variables sampled in training each tree ranged over 0.33, 0.5, and 0.67, and the default values were used for other parameters. A maximum of 400 boosting rounds was used for each parameter combination.

XGBoostRF. Combination Gradient Boosting with Random Forest models were developed by boosting candidate random forest models. XGboost RF models were identified as those with the highest test set AUC when optimizing hyperparameters over the same range of values described in the prior section for XGboost, and additionally expanding the dimensionality of the grid to include 10, 20, and 40 trees.

RuleFit. The RuleFit algorithm creates a model in four steps. First, it fits a tree ensemble to the data. Second, it builds a rule ensemble by traversing each tree. Third, it evaluates the rules on the data to generate additional sets of features that represent interaction terms identified by the rules. Fourth, it fits a sparse Least Absolute Shrinkage and Selection Operatory regression model to the enlarged pool of features containing the original 400, augmented with the newly created rule-based features. RuleFit models were identified as the resultant set of Least Absolute Shrinkage and Selection Operatory models when using 400 rule-generation trees and restricting the candidate pool of features for regression to 1,000 (*i.e.*, the original 400 plus 600 rule-based features).

AutoML. The automated machine learning framework H20 trains a collection of models using various boosting ensembles, random forest methods, generalized linear models, and deep learning (neural networks), with grid search to find optimal parameter values. It then uses a generalized linear model to combine the individual models into an optimal metalearner. The final models were identified as those with the highest AUC on the cross-validation test set optimized over its default parameter settings.

We computed performance metrics for these alternate models just as we did for our primary multivariable regression model. Due to the exploratory nature of this robustness testing component of our work, we avoided statistical comparisons among various models.

Results

There was a total of 18,899,224 admissions in 2017 to 2018 across 9,085,968 beneficiaries in the Medicare research identifiable database who were eligible for analysis in the development set (many patients were admitted multiple times; Supplemental Figure 2, http://links.lww.com/ALN/C923). There were 9,205,835 admissions eligible from 5,336,265 beneficiaries for analysis in the 2019 prospective validation set (fig. 1). For our machine learning analysis, there were 476,593 admissions across 279,016 beneficiaries in the 5% limited database in 2018 for development, and 465,064 admissions from 272,220 beneficiaries in 2019 for the prospective out-of-sample evaluation. Population characteristics of the development and validation sets were similar, with a slight predominance of women (54%), and the average age was 74 yr. About 80% of admissions were unplanned (Supplemental Table 2, http://links.lww.com/ALN/C923).

Performance of models developed on the learn and test sets was nearly identical (not shown), confirming robustness of the modeling process and a lack of overfitting. Prospective performance characteristics of predictors of the binary events in the 2019 out-of-sample validation set are summarized in table 1. The incidence of endpoints ranged from 2.8% for in-hospital mortality to 35.9% for discharge to a care facility. The mean and range of AUCs across nine outcomes were 0.76 (0.70 to 0.82). The mean and range of the calibration goodness-of-fit measure R^2 for the 99% of patients at lowest risk were 1.00 (0.99 to 1.00).

The mean and range of the observed-to-expected ratio were 0.97 (0.90 to 1.00). For the highest 5% risk population, mean and range of sensitivity were 19.2% (11.6 to 30.1%), for positive predictive value they were 37.2% (14.6 to 87.7%), and for lift (enrichment ratio) they were 3.8 (2.3 to 6.1). At the observed incidence, the mean and range of sensitivity were 31.2% (15.5 to 63.9%), for positive predictive value they were 31.2% (15.5 to 63.9%), and for lift they

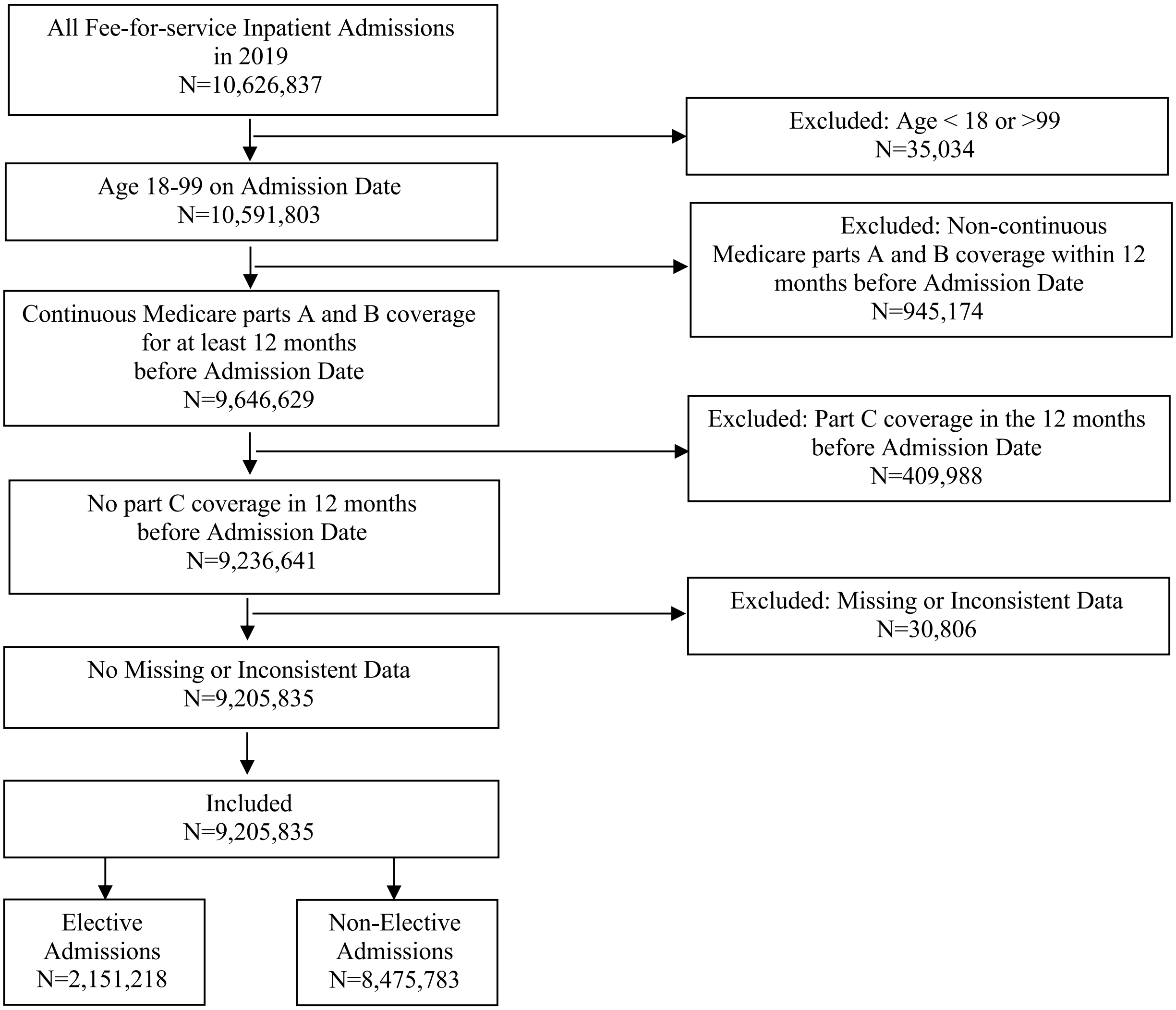

All Fee-for-service Inpatient Admissions in 2019
N=10,626,837
Excluded: Age < 18 or >99
N=35,034
Age 18-99 on Admission Date
N=10,591,803
Excluded: Non-continuous Medicare parts A and B coverage within 12 months before Admission Date
N=945,174
Continuous Medicare parts A and B coverage for at least 12 months before Admission Date
N=9,646,629
Excluded: Part C coverage in the 12 months before Admission Date
N=409,988
No part C coverage in 12 months before Admission Date
N=9,236,641
Excluded: Missing or Inconsistent Data
N=30,806
No Missing or Inconsistent Data
N=9,205,835
Included
N=9,205,835
Elective Admissions
N=2,151,218
Non-Elective Admissions
N=8,475,783

Fig. 1. Cohort selection diagram for validation dataset.

were 3.7 (1.7 to 7.4). Note that the sensitivity and positive predictive value are equal when the detector operates at a threshold resulting in positive decisions (alerts) at a rate equaling the event incidence. A sample composite plot of charts describing performance characteristics for discharge to facility is presented in fig. 2. Similar plots for all other endpoints are in Supplemental Figures 3 to 11 (http://links.lww.com/ALN/C923). Excess length of stay had a root mean square error of 0.19 and R^2 of 0.99.

Table 2 shows the prospective performance on the 5% limited dataset of logistic and various machine learning models developed on the 5% limited dataset and the performance of the logistic model developed on the 100% Research Identifiable File. The 5% sample appeared sufficient for comparative performance to the 100% sample because it lacked only one feature found in the 100% sample (F328, other depressive episodes.) Overall, AUC performance was similar for each endpoint across all model types, with only minor differences in relative performance among endpoints.

The logistic model developed on the 100% Research Identifiable File performed best on eight of the nine endpoints, probably because a larger selection of statistically significant features afforded by the larger pool of events available in the much larger database. Likewise, machine learning models developed using the larger database would presumably perform better than those developed on the 5% sample, but there is no reason to expect that relative ranking would differ much. Of models developed on the smaller 5% limited dataset, it appears that gradient boosting performed marginally better than logistic regression, but not by a clinically meaningful amount. This observation is consistent with previous work demonstrating that logistic regression provides results comparable to machine learning methods when large datasets are used.[36,37]

Table 1. Primary Model Performance on Out-of-sample 2019 Medicare Admissions

			Calibration Metrics				Discrimination Performance Metrics						
								Operating Point that Alerts for Highest 5% Risk Group (%)			Operating Point that Alerts for Highest Incidence % Risk Group (%)		
Period	Endpoint	Incidence (%)	Observed-to-Expected Ratio	R^2 Goodness of Fit*	Correlation Coefficient*	Calibration Estimates (Intercept/ Slope	AUC (95% CI)	Sensitivity	Positive Predictive Value	Lift	Sensitivity	Positive Predictive Value	Lift
In-hospital	Mortality	2.8	0.95	1.00	1.00	0.00/0.94	0.82 (0.82–0.82)	30.1	17.0	6.1	20.6	20.6	7.4
	Discharge to facility	35.9	0.98	1.00	1.00	0.00/1.00	0.79 (0.79–0.79)	12.2	87.7	2.4	63.9	63.9	1.8
90-days after admission	Pneumonia	3.9	0.90	1.00	1.00	0.00/0.89	0.72 (0.72–0.72)	18.5	14.6	3.7	15.5	15.5	4.0
	Acute kidney injury	5.5	0.99	0.99	1.00	0.01/0.87	0.73 (0.73–0.73)	16.3	17.8	3.2	17.8	17.8	3.2
	Sepsis	5.8	1.00	1.00	1.00	0.01/0.92	0.73 (0.73–0.74)	18.7	21.7	3.7	20.7	20.7	3.6
	Major cardiovascular complication	6.0	0.98	1.00	1.00	0.01/0.93	0.79 (0.79–0.79)	24.1	28.9	4.8	27.5	27.5	4.6
	Respiratory failure	6.3	0.98	1.00	1.00	0.01/0.90	0.73 (0.73–0.73)	20.0	25.3	4.0	23.9	23.9	3.8
	Mortality	13.4	0.96	1.00	1.00	0.01/0.95	0.82 (0.82–0.82)	21.1	56.7	4.2	44.6	44.6	3.3
	Unplanned admission	28.0	0.99	1.00	1.00	0.00/0.98	0.70 (0.70–0.70)	11.6	65.0	2.3	46.7	46.7	1.7
Overall mean (95% CI)		12.0 (3.7–20.2)	0.97 (0.95–0.99)	1.00 (1.00–1.00)	1.00 (1.00–1.00)	0.01 (0.00–0.01) 0.93 (0.90–0.96)	0.76 (0.73–0.79)	19.2 (15.2–23.2)	37.2 (19.1–55.3)	3.8 (3.0–4.6)	31.2 (19.7–42.7)	31.2 (19.7,42.7)	3.7 (2.5–4.9)

Metrics include endpoint, observation window (either in-hospital or within 90-days post-admission), incidence, observed-to-expected ratio, R^2 goodness-of-fit measure, correlation coefficient, area under the receiver operating characteristics curve (AUC), sensitivity, positive predictive value, and lift (enrichment ratio). Endpoints in the table are ordered by increasing incidence within their corresponding event period.

*R^2 and correlation coefficient 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. N.B. sensitivity and positive predictive value are equal when the detector operates at a threshold resulting in positive decisions (alerts) at a rate equal to the event incidence.

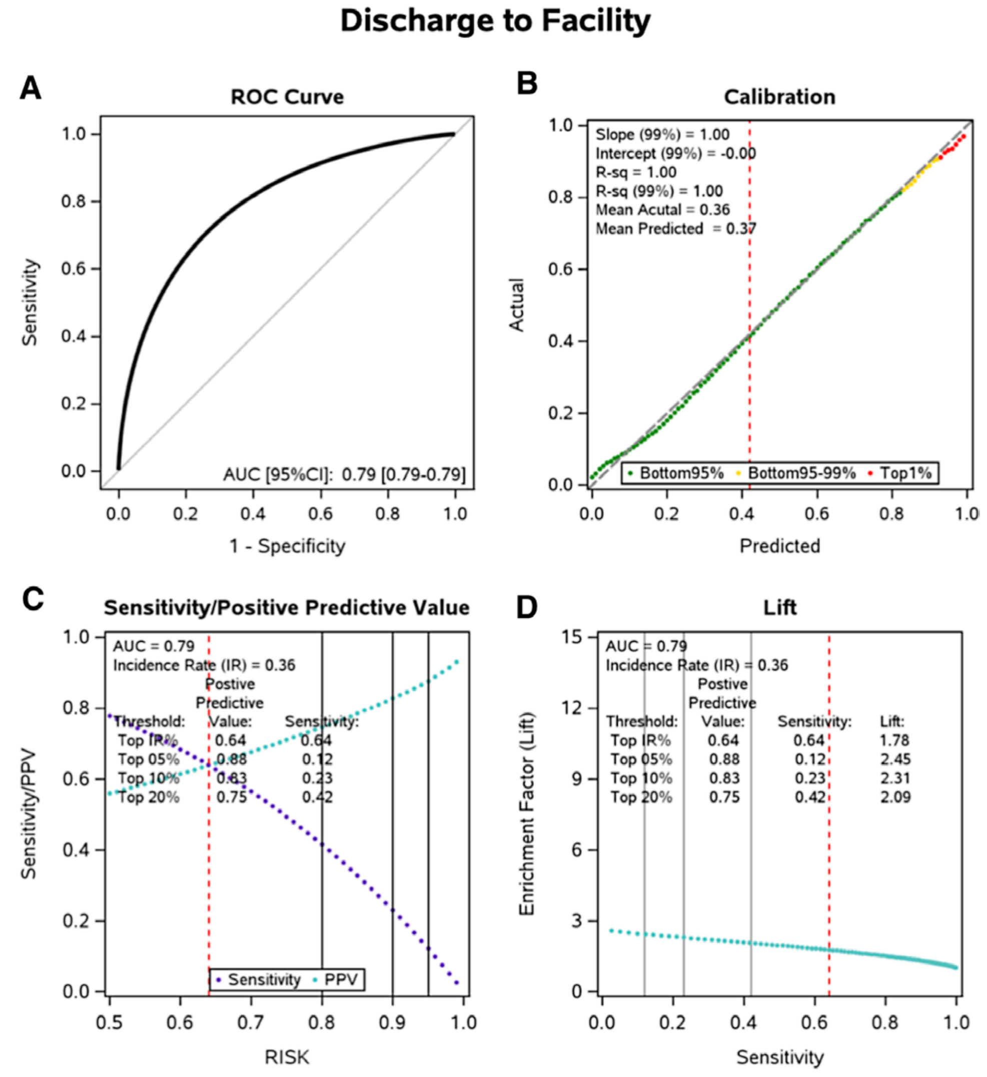


Fig. 2. Model performance for discharge to facility. *A*, Receiver operating characteristic curve. The curve displays the tradeoff between sensitivity and specificity over the range of possible detection thresholds. Tabulated metrics: mean and 95% CI of the area under the receiver operating characteristics curve (AUC). *B*, Calibration curve. The calibration plot displays mean actual incidence *versus* mean predicted risk of discharge to facility for populations clustered in 1% increments of the predicted risk. *Dark green*, *light green*, and *red dots* are populations of the lowest 95%, 95 to 99%, and top 1% risk. The *diagonal line* identifies the domain of ideal performance where actual and expected incidence are equal. The performance of this index is close to ideal for approximately 99% of the population. Tabulated metrics: The incidence of discharge to facility was 36%. The AUC was 0.79. Slope (99%) and Intercept (99%) are the estimates of slope and intercept of the best fit line for all subjects except the riskiest 1%. Rsq and Rsq (99%) are goodness-of-fit measures of individual results to the best fit line for all subjects and all subjects except the riskiest 1% (*i.e.*, *green dots*). *C*, Sensitivity/positive predictive value plot. Positive predictive value (*blue dots*) and sensitivity (*purple dots*) *versus* the fraction of population, sorted by the risk of discharge to facility. The *vertical red line* indicates where the number of patients above the risk threshold equals the incidence of the discharge to facility event in the population. Tabulated metrics: AUC and the incidence of discharge to facility (incidence rate). *Vertical bars* help identify the positive predictive value and sensitivity performance for detectors operating to identify the riskiest 5%, 10%, and 20% of patients. The positive predictive value and sensitivity are tabulated for these detector operating points. *D*, Enrichment factor (lift) plot. Lift (*i.e.*, positive predictive value/incidence) *versus* sensitivity. *Vertical bars* help identify the lift and sensitivity performance for detectors operating to identify the riskiest 5%, 10%, and 20% of patients. Positive predictive value, sensitivity, and lift are tabulated for these detector operating points. The AUC and incidence of mortality (incidence rate) are also tabulated.

Complication-specific models consistently identified more patients who experienced specific complications than patients selected only for mortality risk (fig. 3). Model information, including open source coefficient files, is made available at https://my.clevelandclinic.org/departments/anesthesiology/depts/outcomes-research/risk-stratification

Table 2. Comparative Classification Performance across Modeling Techniques

		Endpoint and Period								
		In-Hospital		90-Days after Admission						
Modeling Technique	Development Database	Mortality	Discharge to Facility	Mortality	Acute Kidney Injury	Sepsis	Respiratory Failure	Pneumonia	Major Cardiovascular Complication Composite	Unplanned Admission
Logistic	5% LDS	0.808 (0.804–0.811)	0.774 (0.773–0.775)	0.817 (0.815–0.819)	0.705 (0.701–0.708)	0.713 (0.710–0.716)	0.734 (0.731–0.737)	0.690 (0.686–0.694)	0.768 (0.765–0.771)	0.700 (0.698–0.701)
xgboost		0.815 (0.812–0.819)	0.780 (0.778–0.781)	0.821 (0.820–0.823)	0.715 (0.712–0.718)	0.720 (0.717–0.723)	0.741 (0.738–0.744)	0.701 (0.697–0.704)	0.775 (0.773–0.778)	0.702 (0.701–0.704)
xgboostRF		0.820 (0.817–0.823)	0.786 (0.784–0.787)	0.823 (0.822–0.825)	0.720 (0.717–0.723)	0.723 (0.720–0.726)	0.745 (0.742–0.748)	0.697 (0.693–0.700)	0.778 (0.775–0.780)	0.704 (0.703–0.706)
rangerRF		0.806 (0.803–0.810)	0.772 (0.771–0.773)	0.815 (0.814–0.817)	0.722 (0.719–0.725)	0.723 (0.720–0.726)	0.743 (0.740–0.746)	0.710 (0.707–0.714)	0.777 (0.775–0.780)	0.702 (0.701–0.704)
autoML		0.817 (0.813–0.820)	0.782 (0.780–0.783)	0.812 (0.810–0.814)	0.710 (0.707–0.713)	0.701 (0.698–0.704)	0.725 (0.722–0.728)	0.689 (0.685–0.693)	0.767 (0.764–0.770)	0.701 (0.700–0.703)
RuleFit		0.811 (0.808–0.815)	0.776 (0.774–0.777)	0.819 (0.818–0.821)	0.711 (0.708–0.714)	0.717 (0.714–0.720)	0.738 (0.735–0.741)	0.695 (0.691–0.698)	0.773 (0.770–0.776)	0.701 (0.699–0.702)
Logistic	100% Research identifiable file	0.820 (0.816–0.823)	0.774 (0.773–0.776)	0.823 (0.822–0.825)	0.724 (0.721–,0.727)	0.730 (0.727–0.733)	0.746 (0.743–0.749)	0.715 (0.711–0.718)	0.781 (0.779–0.784)	0.705 (0.703–0.706)

Overall classification performance was assessed using the area under the receiving operating characteristics curve (AUC), with higher AUC values indicating better detection performance. Table cells report the AUC and 95% CI. The techniques evaluated were logistic regression (Logistic, the reference method), Extreme Gradient Boosting (XGBoost), Combination Gradient Boosting with Random Forests (XGBoostRF), Ranger Random Forest (RangerRF), Automated Machine Learning (autoML), and RuleFit (RuleFit). Two logistic models were developed (one on the 5% Limited Data Set (LDS) set, the other on the 100% Research Identifiable File). All models were prospectively evaluated on the Medicare 2019 Admissions 5% Limited Data Set.

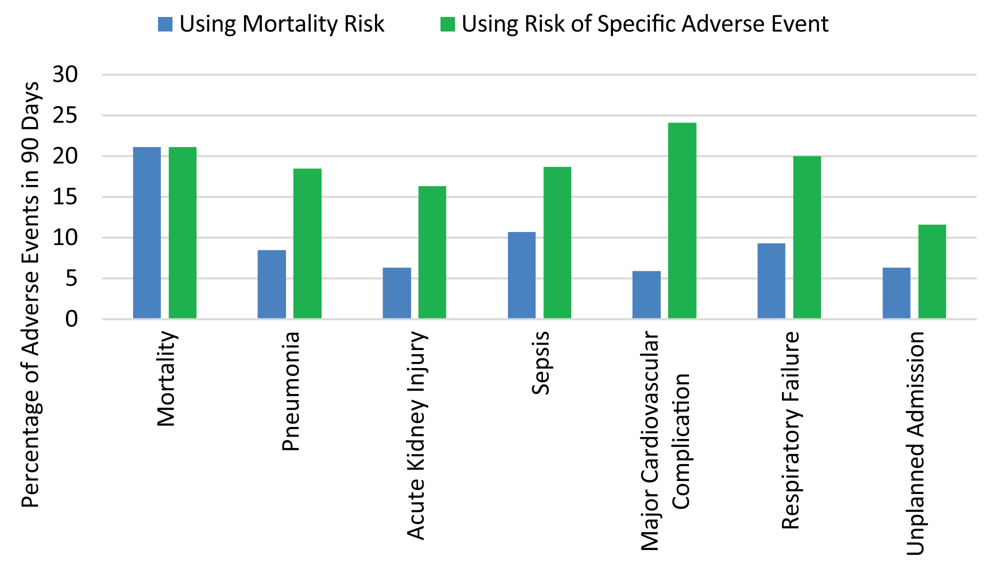


Fig. 3. Percentage of adverse events in the highest 5% risk group detected using complication-specific prediction models *versus* the highest 5% risk group for 90-day mortality. *Blue bars* show the incidence of various complications in patients selected for having the highest 5% mortality risk. *Green bars* show the incidence of complications for the highest 5% risk based on complication-specific models. Use of predictors specific to adverse events results in detecting far more adverse events than using the risk of mortality alone.

(accessed October 12, 2022). The repository includes descriptions of the models and instructions how to derive the predictors from International Classification of Diseases, Tenth Revision, codes, and how to use the provided equations to make predictions.

A reasonable question is whether available medical history is a sufficient substitute for present-on-admission codes. We therefore compared the two models using codes found the year before admission *versus* performance using only codes that were *post hoc* coded as present on admission. The AUC performance for the two methods is shown for each model in table 3. Models built with historical codes were preferable, with a P value of 0.006. We therefore conclude that using available codes from the year before admission is superior to using present-on-admission codes—which are actually not generally available at admission since they are usually coded *post hoc*.

Discussion

Unlike previous versions of the Risk Stratification Index, version 3.0 models are based on billing codes from the year before admission. The only information used from index admissions was admitting diagnosis, principal procedure (for planned admissions), and patient age. Consequently, our models can be implemented at admission with results immediately available to inform clinical decision-making (field tests of this approach are in progress at several institutions). Other major advances include use of International Classification of Diseases, Tenth Revision, codes, inclusion of medical as well as surgical admissions, and many new outcomes. Risk Stratification Index 3.0 is thus a substantial advance from previous versions.

Despite restricting inputs to our models to health events captured in billing codes during the year preceding admission and limited information about the pending admission, our models performed reasonably well. AUCs in the validation set exceeded 0.70 for all endpoints, indicating satisfactory discrimination power over the range of operating thresholds. The calibration goodness-of-fit measures, R^2, exceeded 0.95 for all models, indicating strong correlation between observed and predicted values along the full continuum of risk. Furthermore, the prospective observed-to-expected ratios were between 0.95 and 1.00 across all outcomes, indicating that our models predicted outcomes well in an out-of-sample population.

We chose a conservative approach to evaluate sensitivity, positive predictive value, and lift performance by prespecifying two standard thresholds for comparing models. The top 5% risk threshold represents a means to identify subjects at the highest level of risk for a particular outcome as in a recent Artificial Intelligence Health Outcomes Challenge competition.[38] We also considered a threshold tethered to the incidence for each endpoint.

Observed sensitivity, positive predictive value, and lift results for each model highlight both the limitations and potential clinical application of these types of predictive models. Focusing solely on subjects with the highest 5% of risk provides only modest sensitivity of between 12 to 30% of

Table 3. Comparative Performance between Models Based on Historical Codes *versus* Present-on-Admission Codes Based Out-of-Sample 2019 Medicare Admissions

		1-yr Hx (No POA) [Expected Use "Historical Method"]	POA Only (No 1-yr Hx) ["POA Method"]	Pairwise Difference (Historical – POA]
Period	**Endpoint**	**AUC (95% CI)**	**AUC (95% CI)**	**AUC Difference**
In-Hospital	Mortality	0.82 (0.82–0.82)	0.84 (0.84–0.84)	–0.02
	Discharge to facility	0.79 0.79–0.79)	0.76 (0.76–0.76)	0.03
90-days after admission	Pneumonia	0.72 (0.72–0.72)	0.67 (0.67–0.67)	0.05
	Acute kidney injury	0.73 0.73–0.73)	0.67 (0.67–0.67)	0.06
	Sepsis	0.73 (0.73–0.73)	0.69 (0.69–0.70)	0.04
	Major cardiovascular complication	0.79 (0.79–0.79)	0.75 0.75–0.76)	0.04
	Respiratory failure	0.73 (0.73–0.73)	0.69 0.69–0.69)	0.04
	Mortality	0.82 (0.82–0.82)	0.82 (0.82–0.82)	0.00
	Unplanned admission	0.70 (0.70–0.70)	0.65 (0.65–0.65)	0.05
Overall mean (95% CI)		0.76 (0.73–0.79)	0.73 (0.68–0.77)	0.03 (0.01–0.05)

Complications and resource utilization outcomes are ordered by increasing frequency within their event periods. The 1-yr historical method used available billing codes in the year before admission, and the present-on-admission method used only codes that were *post hoc* coded as being present on admission. Both methods used admitting diagnosis (not principal diagnosis) and principal procedure when appropriate. Variance is presented as 95% CI.

those who will experience each event. Although most people who had an adverse event were at lower risk levels, lift values for these top 5% riskiest patients exceeded 2.3, indicating that they were more than twice as likely as others to experience a future outcome event. The enrichment factor was particularly high for low-incidence (less than 5%) events, ranging from 3.7 to 6.1 for various models. In practical terms, this means that patients identified as being in this highest-risk category on admission are about five times more likely to experience adverse events than the general population. This 3.0 generation of Risk Stratification Index models, based solely on previous claims history and admitting diagnosis, therefore quantifies and ranks patient risk surprisingly well.

Both health trajectory and real time data will increasingly be available because Medicare is encouraging intraoperability and improved access to individual claims information through Blue Button individual access,[39] the Beneficiary Claims Data application program interface,[40] and Data at the Point of Care[41] initiatives. Real-time availability of automatically generated Risk Stratification Index profiles and associated alerts may help reduce cognitive load for the clinician by identifying key areas of concern in individual patients, thus potentially guiding monitoring and management.[42]

Although identifying patients at high risk of mortality helps to identify patients at high risk for specific adverse events associated with mortality, the use of predictors specific to adverse events results in detecting far more adverse events than using the risk of mortality alone. Accurate selection of patients at risk for specific complications therefore requires complication-specific models. A suite of predictors thus provides more information to guide risk-reducing treatment pathways in personalized care plans than overall risk of mortality.

Risk stratification models are conventionally developed from logistic regression models. Regression has the advantage of generating models that are portable and easy to deploy, apply, and interpret. Furthermore, regression models provide coefficients that identify factors that contribute most to specific adverse events, which is clinically valuable information, especially when contributing factors are modifiable. An additional consideration is that models can easily be re-run to accommodate new information obtained at a prehospitalization assessment, or even during hospitalization.

Although interpretable models are frequently preferred even among experts, machine learning models are increasingly popular.[43] Some reports suggest better performance for ML models compared to traditional clinical scoring systems,[17,18] but few have been validated on multiple external datasets. We evaluated five of the most commonly used methods. Interestingly, prediction model characteristics were similar with all six approaches—indicating that none was obviously superior and that any of the approaches is valid.[11–16] Therefore, an important corollary is that we appear to have identified both the value and ceiling for the amount of predictive information that can be derived from the medical claims history.

Our validation analysis was based on more than 9 million adult hospital admissions in 2019 among patients enrolled in the United States fee-for-service and Medicare/Medicaid program. Patients included in our validation represent approximately 70% of all hospital admissions in the Medicare-eligible population in the United States.[10] We excluded less than 0.4% of the available admissions because of missing and inconsistent values. Furthermore, data were missing nonsystematically, meaning that exclusion of these admissions was unlikely to introduce meaningful bias. Our results are therefore broadly applicable to Medicare-eligible adults. Although our 2019 sample included 1,025,099 dual eligible subjects younger than 65 yr (representing 16.3% of the 2019 dataset), our results should be cautiously generalized to younger and healthier populations.

Use of the Medicare claims database to represent individual medical histories remains controversial. For example, reliability of the Centers for Medicare and Medicaid Services registry depends completely on accurate coding. However, well-enforced federal laws promote accurate billing, and regional differences in billing appear to result from true local practice patterns rather than miscoding.[15] Potential errors and delays in Medicare claims processing are offset by large sample size, population diversity, and the highly structured and longitudinal nature of the dataset.

The Centers for Medicare and Medicaid Services usually uses a 30-day observation period. We used a 90-day period because previous reports suggest that 90-day outcomes may be more reliable than 30-day outcomes for measuring hospital performance.[22–24] Model performance at 30 days may of course differ.[40] A potential limitation of real-time use of the models may be incomplete or delayed access to codes in a patient's history, which may lead to underprediction of risk. However, analysis shows that predictions based on available 1-year coding history are slightly better than those based on codes for present-on-admission conditions.

In summary, we developed a suite of risk stratification models using methods similar to those used for earlier Risk Stratification Index versions. An important distinction from previous models is that version 3.0 is based on historical information coupled with admitting diagnosis, principal procedure (for planned admissions), and patient age. Consequently, our models can be implemented at admission with results immediately available to guide clinical decision-making. Other major advances include use of International Classification of Diseases, Tenth Revision, codes rather than International Classification of Diseases, Ninth Revision, codes; inclusion of medical as well as surgical admissions; and many new outcomes. Our models predicted outcomes well in an out-of-sample population and provide clinically meaningful guidance to clinicians.

Acknowledgments

The authors thank John Parks (Data Analyst, Health Data Analytics Institute, Dedham, Massachusetts) for his assistance in developing datasets and models. Dr. Greenwald had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. He conducted the data analysis and takes responsibility for it.

Data Sharing

The investigators' access to Medicare data is based on contracts with the Centers for Medicare and Medicaid Services (Baltimore, Maryland), which precludes sharing data. However, the data are readily available to other investigators *via* contract with the Center for Medicare and Medicaid Services. Requests for access to data to replicate these findings require a research protocol and data use agreement. For more information, contact the Research Data Assistance Center (Minneapolis, Minnesota; http://www.resdac.org). Statistical code will not be shared, but instructions for creating the model variables and the equations to apply provided model coefficients will be posted as described in the penultimate paragraph of the Results section.

Research Support

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

Competing Interests

G. F. Chamoun, N. G. Chamoun, D. Clain, Z. Hong, Dr. Greenwald, Dr. Manberg, and Dr. Jordan are employees of the Health Data Analytics Institute (Dedham, Massachusetts). Dr. Sessler is a consultant and shareholder in the Health Data Analytics Institute and is a consultant for Pacira biosciences (Parsippany, New Jersey). He serves on advisory boards and has equity interests in Calorint (Philadelphia, Pennsylvania), TransQtronics (Philadelphia, Pennsylvania, Medasense (Tel Aviv, Israel), Serenno

(Yokneam, Israel), Sensifree (Cupertino, California), Perceptive Medical (Newport Beach, California), and Neuroindex (Tel Aviv, Israel). He serves on the Board of the Foundation for Anesthesia Education and Research, and is a Senior International Fellow at the Population Health Research Institute, McMaster University (Ontario, Canada). The Department of Outcomes Research, which Dr. Sessler chairs, has research grants from dozens of companies. The other authors declare 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 File, http://links.lww.com/ALN/C923

Supplemental Table 1: Endpoint definitions.

Supplemental Table 2: Full Medicare population characteristics.

Supplemental Figure 1: Selection of candidate diagnostic codes.

Supplemental Figure 2. Development cohort selection.

Supplemental Figure 3. Performance characteristics: in-hospital mortality.

Supplemental Figure 4. Performance characteristics: discharge to facility.

Supplemental Figure 5. Performance characteristics: 90-day pneumonia.

Supplemental Figure 6. Performance characteristics: 90-day acute kidney injury.

Supplemental Figure 7. Performance characteristics: 90-day sepsis.

Supplemental Figure 8. Performance characteristics: 90-day cardiovascular complications.

Supplemental Figure 9. Performance characteristics: 90-day respiratory failure.

Supplemental Figure 10. Performance characteristics: 90-day mortality.

Supplemental Figure 11. Performance characteristics: 90-day unplanned admission.

References

1. Lane-Fall MB, Neuman MD: Outcomes measures and risk adjustment. Int Anesthesiol Clin 2013; 51:10–21
2. Imperial MZ, Phillips PPJ, Nahid P, Savic RM: Precision-enhancing risk stratification tools for selecting optimal treatment durations in tuberculosis clinical trials. Am J Respir Crit Care Med 2021; 204:1086–96
3. Haas LR, Takahashi PY, Shah ND, Stroebel RJ, Bernard ME, Finnie DM, Naessens JM: Risk-stratification methods for identifying patients for care coordination. Am J Manag Care 2013; 19:725–32
4. Levin S BS, Toerper M, Debraine A, DeAngelo A, Hamrock E, Hinson J, Hoyer E, Dungarani T, Howell E: Machine-learning-based hospital discharge predictions can support multidisciplinary rounds and decrease hospital length-of-stay. BMJ Innov 2021; 7:414–21
5. 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
6. Dalton JE, Glance LG, Mascha EJ, Ehrlinger J, Chamoun N, Sessler DI: Impact of present-on-admission indicators on risk-adjusted hospital mortality measurement. Anesthesiology 2013; 118:1298–306
7. Sigakis MJ, Bittner EA, Wanderer JP: Validation of a risk stratification index and risk quantification index for predicting patient outcomes: In-hospital mortality, 30-day mortality, 1-year mortality, and length-of-stay. Anesthesiology 2013; 119:525–40
8. Wahl KM, Moretti, E., White, W., Hale B., Gan, T.J.: Validation of a risk-stratification index for predicting 1-year mortality, 2011 Annual Meeting of the American Society of Anesthesiologists. Anesthesiology 2011; pp A014
9. Chamoun GF, Li L, Chamoun NG, Saini V, Sessler DI: Validation and calibration of the Risk Stratification Index. Anesthesiology 2017; 126:623–30
10. Chamoun GF, Li L, Chamoun NG, Saini V, Sessler DI: Comparison of an updated Risk Stratification Index to hierarchical condition categories. Anesthesiology 2018; 128:109–16
11. Kheterpal S, Shanks A, Tremper KK: Impact of a novel multiparameter decision support system on intraoperative processes of care and postoperative outcomes. Anesthesiology 2018; 128:272–82
12. Lee CK, Hofer I, Gabel E, Baldi P, Cannesson M: Development and validation of a deep neural network model for prediction of postoperative in-hospital mortality. Anesthesiology 2018; 129:649–62
13. Sonny A, Kurz A, Skolaris LA, Boehm L, Reynolds A, Cummings KC 3rd, Makarova N, Yang D, Sessler DI: Deficit accumulation and phenotype assessments of frailty both poorly predict duration of hospitalization and serious complications after noncardiac surgery. Anesthesiology 2020; 132:82–94
14. Turan A, Cohen B, Adegboye J, Makarova N, Liu L, Mascha EJ, Qiu Y, Irefin S, Wakefield BJ, Ruetzler K, Sessler DI: Mild acute kidney injury after noncardiac

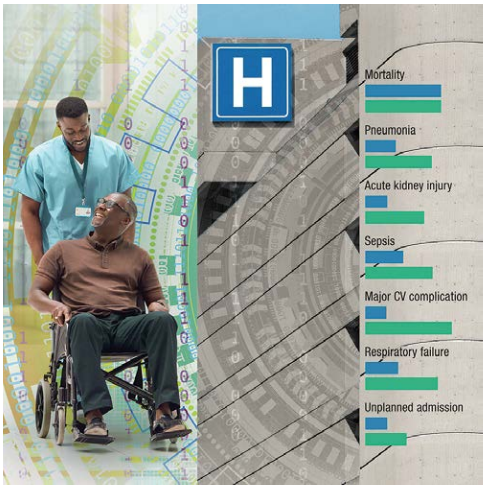
surgery is associated with long-term renal dysfunction: A retrospective cohort study. Anesthesiology 2020; 132:1053–61

15. Li L, Chamoun GF, Chamoun NG, Sessler D, Gopinath V, Saini V: Elucidating the association between regional variation in diagnostic frequency with risk-adjusted mortality through analysis of claims data of Medicare inpatients: A cross-sectional study. BMJ Open 2021; 11:e054632
16. Greenwald SD, Chamoun NG, Manberg PJ, Gray J, Clain D, Maheshwari K, Sessler DI: Covid-19 and excess mortality in Medicare beneficiaries. PLoS One 2022; 17:e0262264
17. Hill BL, Brown R, Gabel E, Rakocz N, Lee C, Cannesson M, Baldi P, Olde Loohuis L, Johnson R, Jew B, Maoz U, Mahajan A, Sankararaman S, Hofer I, Halperin E: An automated machine learning-based model predicts postoperative mortality using readily-extractable preoperative electronic health record data. Br J Anaesth 2019; 123:877–86
18. Rellum SR, Schuurmans J, van der Ven WH, Eberl S, Driessen AHG, Vlaar APJ, Veelo DP: Machine learning methods for perioperative anesthetic management in cardiac surgery patients: A scoping review. J Thorac Dis 2021; 13:6976–93
19. Jing B, Boscardin WJ, Deardorff WJ, Jeon SY, Lee AK, Donovan AL, Lee SJ: Comparing machine learning to regression methods for mortality prediction using Veterans Affairs electronic health record clinical data. Med Care 2022; 60:470–9
20. 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
21. CMS: International Classification of Diseases, Tenth Revision (ICD-10). Baltimore, Centers for Medicare & Medicaid Services, 2021
22. Hirji S, McGurk S, Kiehm S, Ejiofor J, Ramirez-Del Val F, Kolkailah AA, Berry N, Sobieszczyk P, Pelletier M, Shah P, O'Gara P, Kaneko T: Utility of 90-day mortality vs 30-day mortality as a quality metric for transcatheter and surgical aortic valve replacement outcomes. JAMA Cardiol 2020; 5:156–65
23. Mizushima T, Yamamoto H, Marubashi S, Kamiya K, Wakabayashi G, Miyata H, Seto Y, Doki Y, Mori M: Validity and significance of 30-day mortality rate as a quality indicator for gastrointestinal cancer surgeries. Ann Gastroenterol Surg 2018; 2:231–40
24. Damhuis RA, Wijnhoven BP, Plaisier PW, Kirkels WJ, Kranse R, van Lanschot JJ: Comparison of 30-day, 90-day and in-hospital postoperative mortality for eight different cancer types. Br J Surg 2012; 99:1149–54
25. AHRQ: Clinical Classification Software Refined (CCSR), Agency for Healthcare Research and Quality, 2021
26. AHRQ: Clinical Classification Software ICD-10-PCS (CCS), Agency for Healthcare Research and Quality, 2021
27. ResDAC: Patient Discharge Status Table, Research Data Assistance Center (ResDAC), 2021
28. 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
29. Schmueli G: Lift up and Act! Classifier Performance in Resource-constrained Applications. arXiv, 2019; abs/1906.03374
30. 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
31. Breiman L: Random forests. Machine Learning 2001; 45:5–32
32. Wright MN, Ziegler, A.: Ranger: A fast implementation of random forests for high dimensional data in C++ and R. J Stat Softw 2017; 77:1–17
33. Friedman JH: Stochastic gradient boosting. Comput Stat Data Anal 2002; 38:367–78
34. Friedman JH, Popescu, B.E.: Predictive learning via rule ensembles. Ann Appl Stat 2008; 2:916–54
35. H20.ai: R Interface for H20, 2016
36. Cowling TE, Cromwell DA, Bellot A, Sharples LD, van der Meulen J: Logistic regression and machine learning predicted patient mortality from large sets of diagnosis codes comparably. J Clin Epidemiol 2021; 133:43–52
37. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B: A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 2019; 110:12–22
38. CMS: Artificial Intelligence (AI) Health Outcomes Challenge, Centers for Medicare & Medicaid Services, 2018
39. CMS: CMS Blue Button 2.0, 2021
40. CMS: Beneficiary Claims Data API, 2021
41. CMS: Data at the Point of Care, 2021
42. Chi EA, Chi G, Tsui CT, Jiang Y, Jarr K, Kulkarni CV, Zhang M, Long J, Ng AY, Rajpurkar P, Sinha SR: Development and validation of an artificial intelligence system to optimize clinician review of patient records. JAMA Netw Open 2021; 4:e2117391
43. Rudin C: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 2019; 1:206–15

Prediction Algorithms: Is Peer Review Enough?

Laurent G. Glance, M.D., Laszlo Vutskits, M.D., Ph.D., Andrew Davidson, M.B.B.S., M.D., F.A.N.Z.C.A., F.A.H.M.S.

Many risk prediction models have been developed, some of which serve as the foundation for healthcare reform and clinical decision-making. Many of these tools were developed based on diagnoses and procedure codes from the index hospitalization, meaning that most of the information used as the inputs for these prediction models is only available after patient discharge. The inherent disadvantage of this approach is that it does not allow accurate and individualized risk stratification at the time of hospital admission when such an evaluation is of particular clinical relevance.



"Before embedding predictive analytics in the electronic medical record, should we require independent testing to show that they are 'safe' and improve outcomes—or at a minimum, that the models accurately predict outcomes?"

In this issue, Greenwald *et al.*[1] present an updated version of their Risk Stratification Index. The authors are to be congratulated for creating and validating a robust set of prediction models based on 4,426 International Classification of Diseases codes out of a possible 69,000 diagnostic codes. The authors suggest that the current revision will be more useful than previous versions because it uses International Classification of Diseases codes coded the year before hospital admission, thus making the revised index usable during the index admission. However, the decision to include only International Classification of Diseases codes that are present the year before admission may be both a strength and a limitation, since some patients may develop new diagnoses that are present on admission but not available in historical data. Furthermore, patients may be admitted to hospitals that do not have access to their historical data. Nonetheless, the predictions of the Risk Stratification Index (risk of death, major complications [acute kidney injury, sepsis, respiratory failure], excess length of stay, and unplanned readmission) could be used to guide patient management in important ways. For example, the Risk Stratification Index could be used to (1) identify patients who may benefit from step-down or intensive care, (2) triage surgical patients to match the skill set of anesthesiologists and trainees, (3) guide medical therapy to optimize outcomes, and (4) save costs by reducing the use of unnecessary levels of care.

The Risk Stratification Index is one of many prediction models that are now widespread in medicine. Prediction models have become a fundamental driver of healthcare reform and clinical practice. The quality and performance of hospitals and physicians cannot be fairly measured without first adjusting for differences in patient case mix and surgical complexity using risk adjustment models. The Centers for Medicare and Medicaid Services publicly report hospital risk-adjusted outcomes to promote transparency and patient choice. The American College of Surgeons[2] and the Society of Thoracic Surgeons[3] provide their members with nonpublic performance reporting to guide quality improvement. The Centers for Medicare and Medicaid Services is redesigning the healthcare system to deliver higher quality care at a lower cost using pay-for-performance (*e.g.*, Hospital Readmission Reduction Program[4]); episode-based payments (*e.g.*, Comprehensive Care for Joint Replacement,[5] Bundled

Image: A. Johnson, Vivo Visuals Studio.

This editorial accompanies the article on p. 673. This article has an audio podcast.

Accepted for publication October 17, 2022.

Laurent G. Glance, M.D.: Department of Anesthesiology and Perioperative Medicine, University of Rochester School of Medicine, Rochester, New York; and RAND Health, RAND, Boston, Massachusetts.

Laszlo Vutskits, M.D., Ph.D.: Department of Anesthesiology, Pharmacology, Intensive Care, and Emergency Medicine, University Hospitals of Geneva, Geneva, Switzerland; and Geneva Neuroscience Center, University of Geneva, Geneva, Switzerland.

Andrew Davidson, M.B.B.S., M.D., F.A.N.Z.C.A., F.A.H.M.S.: Department of Anesthesia, Royal Children's Hospital, Melbourne, Australia; and Murdoch Children's Research Institute, Melbourne, Australia.

Anesthesiology 2022; 137:661–3. DOI: 10.1097/ALN.0000000000004421

<zdoi;. DOI: 10.1097/ALN.0000000000004421>

Payments for Care Improvement[6]); and accountable care organizations (Medicare Shared Savings Program[7]). How much the Centers for Medicare and Medicaid Services pay hospitals depends on their risk-adjusted performance. Prediction models are also used to guide clinical decision-making: (CHA_2DS_2-VASc score for atrial fibrillation stroke risk[8]) and risk stratification before surgery (American College of Surgeons Surgical Risk Calculator[9]).

However, to be useful, a prediction model must accurately predict outcomes. Hospital performance is quantified by comparing its performance (*e.g.*, the observed mortality rate) to its predicted performance (*e.g.*, the expected [predicted] mortality rate). Suppose a prediction model does not accurately predict outcomes. In that case, patients may unintentionally be guided to select low-performance hospitals, the Centers for Medicare and Medicaid Services may penalize high-performance hospitals while rewarding low-performance hospitals, and clinicians may decide to place high-risk patients on the ward immediately after surgery. The performance of prediction models can be assessed using standard statistical criteria (*e.g.*, model discrimination, model calibration) in patient samples that are independent of the sample used to create the prediction model. Using best-in-class prediction models is important because whether a hospital is classified as either a high- or low-quality hospital can depend on which prediction model is used for risk adjustment and not just on the intrinsic quality of the hospital.[10] Similarly, the decision to pursue invasive testing before surgery is also a function of which prediction model is used for risk stratification.[11]

Before the Centers for Medicare and Medicaid Services use a performance measure for quality reporting or value-based purchasing, the measure and the underlying risk adjustment methodology must first be evaluated and endorsed by the National Quality Forum.[12] There is no formal mechanism to evaluate and endorse most prediction models before they are used clinically. Although the Food and Drug Administration is responsible for the regulation of "Software as a Medical Device," it has only recently issued guidance that prediction models like the Epic Sepsis Model (developed by the commercial electronic health record vendor Epic) should be subject to regulatory review.[13] This prediction model, which is widely used at hundreds of U.S. hospitals without first undergoing independent validation, was recently shown to miss 67% of patients with sepsis.[14] Before embedding predictive analytics in the electronic medical record, should we require independent testing to show that they are "safe" and improve outcomes—or at a minimum, that the models accurately predict outcomes? We believe that the answer is a resounding "yes." There are currently best practices for the reporting of prediction models.[15] However, peer review should only be the first step before a prediction model is used to guide clinical care.

Some critics of these algorithms point out that the code's details are often kept proprietary and not published with the article.[16] Anesthesiology encourages authors to describe the code in sufficient detail so that readers can consider whether the fundamentals of the algorithm are built on robust designs and data. What is the right balance between protection of intellectual property *versus* transparency? We believe that journals need to ask more from authors. In the absence of widespread regulation of prediction models, journals are the first and only line of defense to ensure that valid prediction models are disseminated to front-line clinicians. We propose that journals strongly encourage developers to make code available to outside researchers to allow the independent evaluation of prediction models. Alternatively, developers could provide a working version of the prediction model (without sharing proprietary code) to allow independent validation. Greenwald *et al.*[1] are to be commended for providing code for each of the Risk Stratification Index models along with sample data. It is worth noting that if the algorithms are already patented, then the details of the algorithm may already be in the public domain as part of the regulatory requirement for obtaining the patent. Journals should also promote the validation of prediction models by publishing the work of independent teams who evaluate and replicate published models. One challenge in the evaluation of these models is that they are likely to be regularly updated or tweaked. It would be hoped that these models are upgraded as new validation data emerge, as medical practice changes, and as we have access to new types of data (for example, physiologic data from wearables). Like a new phone, the software behind the prediction scores may—and perhaps should—be upgraded every year. Whatever process we have for rigorous evaluation will need to be nimble enough to accommodate regular upgrades.

Last, some worry about the ethical implications of these predictive algorithms. Instead of being used to identify patients who need escalated care, could they instead be used to identify patients who would be denied care because they are predicted to have an increased risk of major complications, extended length of stay, and readmission? Could insurance companies and hospitals use them to selectively avoid patients deemed to be at too high a financial or reputational risk? Race, ethnicity, and insurance status (*e.g.*, Medicaid coverage) are frequently a proxy for unmeasured disease severity. Will including race and ethnicity in prediction models unfairly disadvantage vulnerable populations by encouraging hospitals to selectively avoid these vulnerable individuals?

Risk prediction models have become embedded in the healthcare system. They are the centerpiece of healthcare reform and will play an increasingly important role in clinical decision-making. Anesthesiology welcomes manuscripts that help our readers understand the important features of these algorithms and especially those studies that provide more evidence for when and where they can improve the outcomes for our patients.

Research Support

Supported by the Department of Anesthesiology and Perioperative Medicine at the University of Rochester School of Medicine and Dentistry (Rochester, New York).

Competing Interests

Drs. Vutskits and Davidson are Editors for Anesthesiology. Dr. Glance is a member of the Scientific Methods Panel for the National Quality Forum and an Associate Editor for Anesthesiology. Dr. Glance reports funding from the National Institutes of Health (Bethesda, Maryland; grant Nos. R01AG074492 and RO1NRO16865).

Correspondence

Address correspondence to Dr. Glance: laurent_glance@urmc.rochester.edu

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. Hall BL, Hamilton BH, Richards K, Bilimoria KY, Cohen ME, Ko CY: Does surgical quality improve in the American College of Surgeons National Surgical Quality Improvement Program: An evaluation of all participating hospitals. Ann Surg 2009; 250:363–76
3. Bowdish ME, D'Agostino RS, Thourani VH, Schwann TA, Krohn C, Desai N, Shahian DM, Fernandez FG, Badhwar V: STS Adult Cardiac Surgery Database: 2021 update on outcomes, quality, and research. Ann Thorac Surg 2021; 111:1770–80
4. Desai NR, Ross JS, Kwon JY, Herrin J, Dharmarajan K, Bernheim SM, Krumholz HM, Horwitz LI: Association between hospital penalty status under the hospital readmission reduction program and readmission rates for target and nontarget conditions. JAMA 2016; 316:2647–56
5. Finkelstein A, Ji Y, Mahoney N, Skinner J: Mandatory Medicare bundled payment program for lower extremity joint replacement and discharge to institutional postacute care: Interim analysis of the first year of a 5-year randomized trial. JAMA 2018; 320:892–900
6. Joynt Maddox KE, Orav EJ, Zheng J, Epstein AM: Year 1 of the bundled payments for care improvement—Advanced model. N Engl J Med 2021; 385:618–27
7. McWilliams JM, Hatfield LA, Landon BE, Hamed P, Chernew ME: Medicare spending after 3 years of the Medicare Shared Savings Program. N Engl J Med 2018; 379:1139–49
8. Melgaard L, Gorst-Rasmussen A, Lane DA, Rasmussen LH, Larsen TB, Lip GY: Assessment of the CHA_2DS_2-VASc score in predicting ischemic stroke, thromboembolism, and death in patients with heart failure with and without atrial fibrillation. JAMA 2015; 314:1030–8
9. Cohen ME, Liu Y, Ko CY, Hall BL: An examination of American College of Surgeons NSQIP Surgical Risk Calculator accuracy. J Am Coll Surg 2017; 224:787–95.e1
10. Iezzoni LI: The risks of risk adjustment. JAMA 1997; 278:1600–7
11. Glance LG, Faden E, Dutton RP, Lustik SJ, Li Y, Eaton MP, Dick AW: Impact of the choice of risk model for identifying low-risk patients using the 2014 American College of Cardiology/American Heart Association perioperative guidelines. Anesthesiology 2018; 129:889–900
12. Glance LG, Joynt Maddox K, Johnson K, Nerenz D, Cella D, Borah B, Kunisch J, Kurlansky P, Perloff J, Stoto M, Walters R, White S, Lin Z: National Quality Forum guidelines for evaluating the scientific acceptability of risk-adjusted clinical outcome measures: A report from the national quality forum scientific methods panel. Ann Surg 2020; 271:1048–55
13. Ross C: In New Guidance, FDA Says AI Tools to Warn of Sepsis Should Be Regulated as Devices. 2022. Available at: https://www.statnews.com/2022/09/27/health-fda-artificial-intelligence-guidance-sepsis/. Accessed October 12, 2022
14. Wong A, Cao J, Lyons PG, Dutta S, Major VJ, Otles E, Singh K: Quantification of sepsis model alerts in 24 US hospitals before and during the COVID-19 pandemic. JAMA Netw Open 2021; 4:e2135286
15. Collins GS, Reitsma JB, Altman DG, Moons KG: Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD). Ann Intern Med 2015; 162:735–6
16. Wanderer JP, Ehrenfeld JM: Toward external validation and routine clinical use of the American College of Surgeons NSQIP Surgical Risk Calculator. J Am Coll Surg 2016; 223:674

Glance *et al.* Anesthesiology 2022; 137:661–3 663

Supplemental Table 1: Endpoint definitions and References.

ID	Name	Definition Conditions/Codes[1]	References
MORX	Mortality	In-hospital mortality is identified from the patient discharge status code. 90-day mortality is identified if the beneficiary death date is within 90 days of admission.	[1]
MCEX	Major adverse cardiac events	AMI ICD-10 Codes: I21.x CVA ICD-10 Codes: I60.x, I61.x, I63.x, I64.x, G45.x (except G45.4) CHF ICD-10 Codes: I11.0, I13.0, I13.2, I50.1, I50.2x, I50.3x, I50.4x, I50.8x, I50.9	MCEX is a composite of hospitalizations for: Acute myocardial infarction (AMI)[2] Stroke (CVA) (AIS/ICH definition)[3] Chronic Heart Failure (CHF)[4]
AKIX	Acute kidney injury	Subjects with N17.x, excluding those with Z49.3x	[5]
PNEX	Pneumonia	A48.1, J11.00, J12.x, J13, J14, J15.x, J16.x, J18.0, J18.9	[6]
IPAU	Unplanned hospital readmission	The admission status was classified as "planned" if the reason for admission was elective, and otherwise designated as "unplanned."	[7]

[1] ICD-10-CM codes provided are used to identify the associated endpoint. The letter 'x' indicates any value is acceptable.

ELSX	Excess length of stay	Excess length of stay was defined as the difference between the observed duration of hospitalization and the geometric mean length of stay associated with the default 2021 v1 Clinical Classifications Software Refined (CCSR) [8] category for the admitting diagnosis if the admission was unplanned (i.e., urgent or emergent), or the 2020 Clinical Classifications Software (CCS)[9] category associated with the primary procedure if the admission was planned.	[10]
INFX	Infection	See extensive list of ICD-10-CM infection codes provided in the reference [11].	[11]
IPDF	In-patient discharge location (Facility vs Not)	Discharge status to a facility was defined as discharge to a location other than home (with or without organized home health care.)	[12]

References

1.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*. Nov 2010;113(5):1026-37. doi:10.1097/ALN.0b013e3181f79a8d

2.Metcalfe A, Neudam A, Forde S, et al. Case definitions for acute myocardial infarction in administrative databases and their impact on in-hospital mortality rates. *Health Serv Res*. Feb 2013;48(1):290-318. doi:10.1111/j.1475-6773.2012.01440.x

3.Kumamaru H, Judd SE, Curtis JR, et al. Validity of claims-based stroke algorithms in contemporary Medicare data: reasons for geographic and racial differences in stroke (REGARDS) study linked with medicare claims. *Circ Cardiovasc Qual Outcomes*. Jul 2014;7(4):611-9. doi:10.1161/CIRCOUTCOMES.113.000743

4.Bonow RO, Bennett S, Casey DE, Jr., et al. ACC/AHA Clinical Performance Measures for Adults with Chronic Heart Failure: a report of the American College of Cardiology/American Heart Association Task Force on Performance Measures (Writing Committee to Develop Heart Failure Clinical Performance Measures): endorsed by the Heart Failure Society of America. *Circulation*. Sep 20 2005;112(12):1853-87. doi:10.1161/CIRCULATIONAHA.105.170072

5.Logan R, Davey P, De Souza N, Baird D, Guthrie B, Bell S. Assessing the accuracy of ICD-10 coding for measuring rates of and mortality from acute kidney injury and the impact of electronic alerts: an observational cohort study. *Clin Kidney J*. Dec 2020;13(6):1083-1090. doi:10.1093/ckj/sfz117

6.Yu O, Nelson JC, Bounds L, Jackson LA. Classification algorithms to improve the accuracy of identifying patients hospitalized with community-acquired pneumonia using administrative data. *Epidemiol Infect*. Sep 2011;139(9):1296-306. doi:10.1017/S0950268810002529

7.Ellimoottil C, Khouri RK, Dhir A, Hou H, Miller DC, Dupree JM. An Opportunity to Improve Medicare's Planned Readmissions Measure. *J Hosp Med*. Oct 2017;12(10):840-842. doi:10.12788/jhm.2833

8.AHRQ. Clinical Classification Software Refined (CCSR). Agency for Healthcare Research and Quality. Updated March 5, 2021. Accessed August 21, 2021, https://www.hcup-us.ahrq.gov/toolssoftware/ccsr/ccs_refined.jsp

9.AHRQ. Clinical Classification Software ICD-10-PCS (CCS). Agency for Healthcare Research and Quality. Updated October 2019. Accessed August 21, 2021, https://www.hcup-us.ahrq.gov/toolssoftware/ccs10/ccs10.jsp

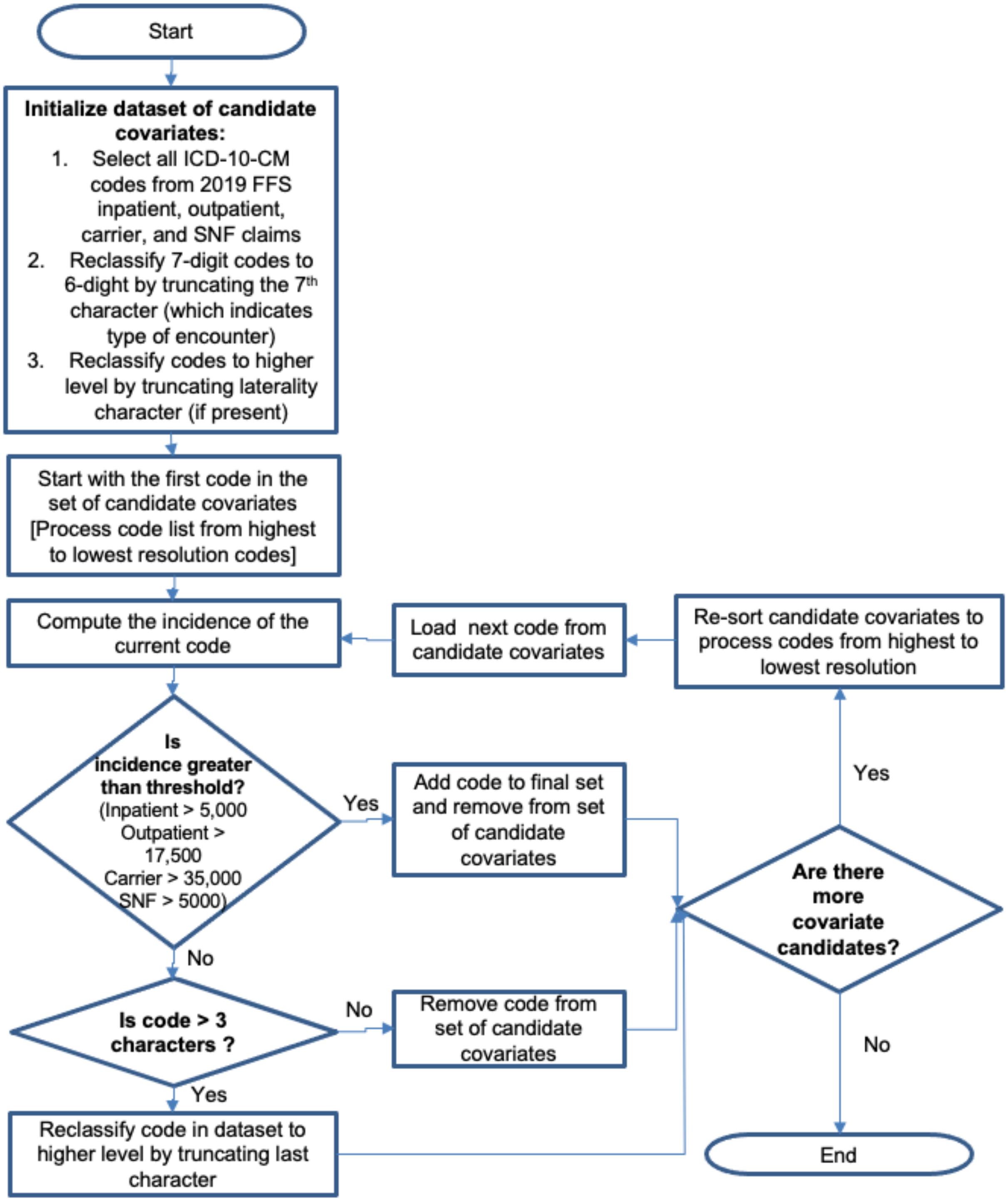
10.Hughes AH, Horrocks D, Jr., Leung C, Richardson MB, Sheehy AM, Locke CFS. The increasing impact of length of stay "outliers" on length of stay at an urban academic hospital. *BMC Health Serv Res*. Sep 9 2021;21(1):940. doi:10.1186/s12913-021-06972-6

11.AHRQ. AHRQ QI(tm) ICD-10-CM/PCS Specification v2021 Patient Safety Indicators, Appendix F. Agency for Healthcare Research and Quality. Updated July 2021. Accessed December 27,2021, https://qualityindicators.ahrq.gov/Downloads/Modules/PSI/V2019/TechSpecs/PSI_Appendix_F.pdf

12.ResDAC. Patient Discharge Status Table. Research Data Assistance Center (ResDAC). Updated Unspecified. Accessed December 28, 2021, https://requests.resdac.org/cms-data/variables/patient-discharge-status-code-encounter

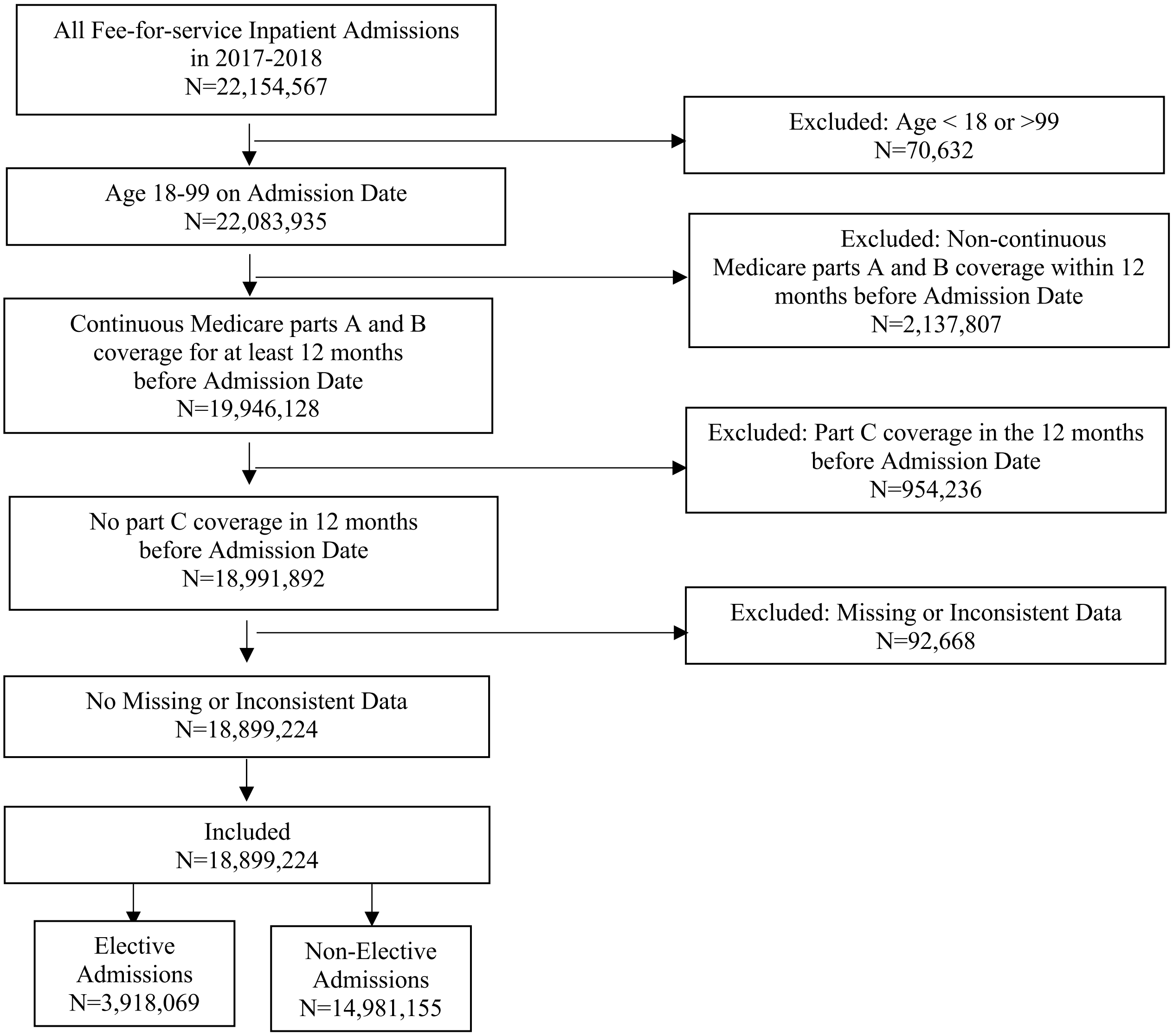
Database	Year	Learn or Test Set	Number of Admissions	Age [IQR]	Male (%)	Unplanned Admissions (%)	Race: White, Black, Asian, Other (%)
Development	2017	Learn	7,669,132	73 [67, 83]	46	79	82.0, 11.9, 1.3, 4.8
		Test	1,918,006	73 [67, 83]	46	79	82.0, 11.9, 1.3, 4.8
	2018	Learn	7,448,554	74 [67, 83]	46	80	82.0, 11.7, 1.4, 5.0
		Test	1,863,532	74 [67, 83]	46	80	82.0, 11.7, 1.4, 5.0
Prospective Validation	2019	Validation	9,205,835	74 [68, 83]	46	80	81.7, 11.7, 1.4, 5.0

Supplemental Table 2: Full Medicare Population Characteristics. The mean and interquartile of age, and percentage of men, unplanned admissions, and percentage race distributions are shown for the 2017-2018 Development Dataset and 2019 Validation Dataset. Models were developed using an 80/20 Learn/Test split of the Development Dataset.


Start
Initialize dataset of candidate covariates:
1. Select all ICD-10-CM codes from 2019 FFS inpatient, outpatient, carrier, and SNF claims
2. Reclassify 7-digit codes to 6-dight by truncating the 7th character (which indicates type of encounter)
3. Reclassify codes to higher level by truncating laterality character (if present)
Start with the first code in the set of candidate covariates [Process code list from highest to lowest resolution codes]
Compute the incidence of the current code
Load next code from candidate covariates
Re-sort candidate covariates to process codes from highest to lowest resolution
Is incidence greater than threshold? (Inpatient > 5,000 Outpatient > 17,500 Carrier > 35,000 SNF > 5000)
Yes
Add code to final set and remove from set of candidate covariates
Yes
Are there more covariate candidates?
No
No
Is code > 3 characters ?
No
Remove code from set of candidate covariates
Yes
Reclassify code in dataset to higher level by truncating last character
No
End

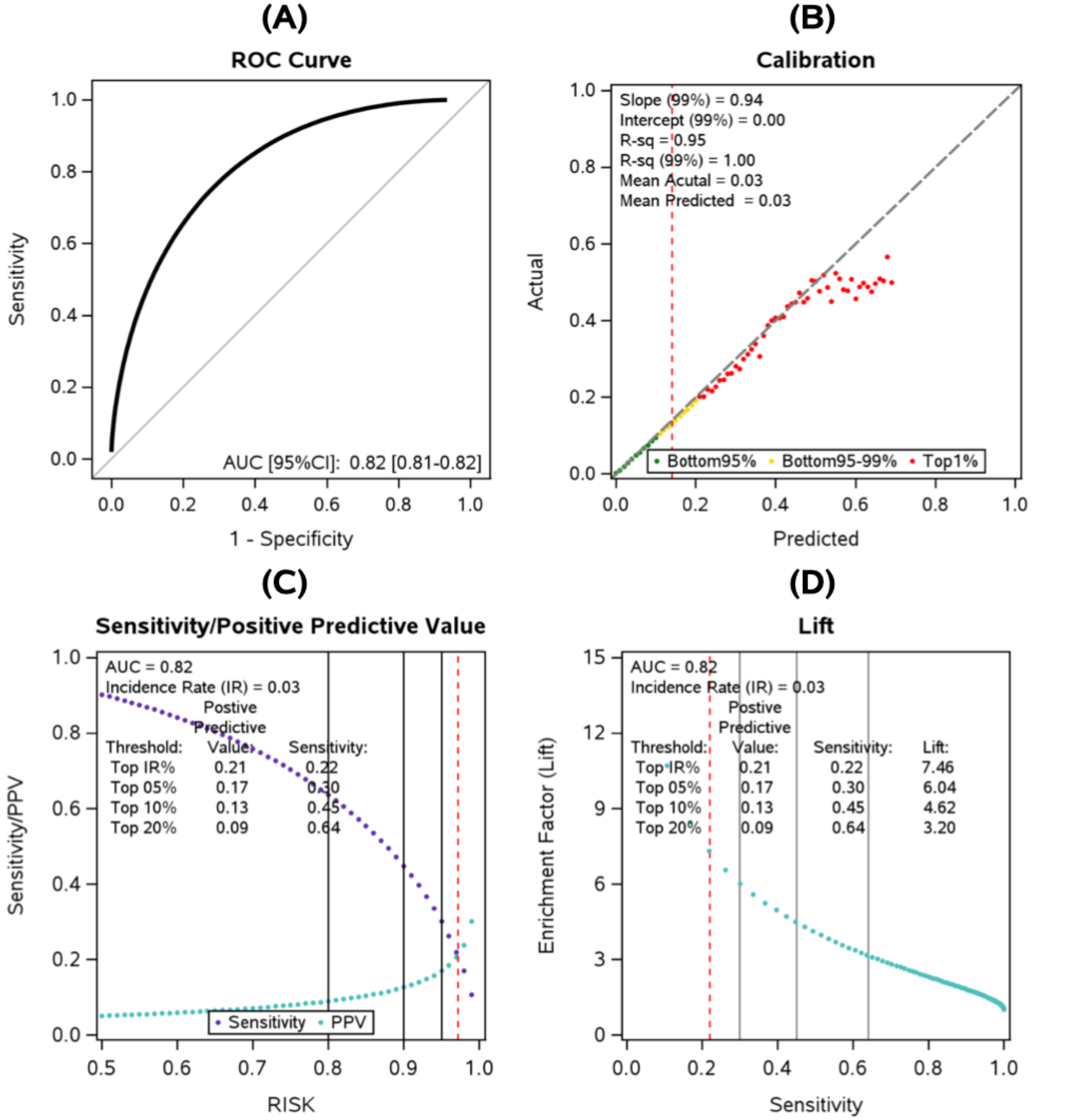
Supplemental Figure 1: Selection process for candidate diagnostic codes (D-Codes). The final set of diagnostic codes are identified by an iterative process that selects frequently occurring codes. ICD-10-CM = International Classification of Diseases, Tenth Edition, Clinical Modification.

Supplemental Figure 2. Cohort Selection Diagram for the Development Dataset


All Fee-for-service Inpatient Admissions in 2017-2018 N=22,154,567
Excluded: Age < 18 or >99 N=70,632
Age 18-99 on Admission Date N=22,083,935
Excluded: Non-continuous Medicare parts A and B coverage within 12 months before Admission Date N=2,137,807
Continuous Medicare parts A and B coverage for at least 12 months before Admission Date N=19,946,128
Excluded: Part C coverage in the 12 months before Admission Date N=954,236
No part C coverage in 12 months before Admission Date N=18,991,892
Excluded: Missing or Inconsistent Data N=92,668
No Missing or Inconsistent Data N=18,899,224
Included N=18,899,224
Elective Admissions N=3,918,069
Non-Elective Admissions N=14,981,155

Supplemental Figures 3-11: Performance characteristics for predicting endpoints. (A) Receiver Operating Characteristic (ROC) Curve. The ROC displays the tradeoff between sensitivity and specificity over the full range of possible detection thresholds. Tabulated metrics: Mean and 95% Confidence Interval (CI) of the area under the ROC (AUC). **(B) Calibration Curve.** The calibration plot displays the mean actual vs predicted endpoint for populations clustered in increments of 1% predicted risk. Dark green, light green, and red dots are populations of the lowest 95%, 95%-99%, and top 1% risk of adverse event. The diagonal line identifies the domain of ideal performance where actual and expected mortality rates are equal for a population. Tabulated metrics: Incidence (Mean Actual), average risk (Mean Predicted), and the area under the Receiver Operating Curve. Slope (99%) and Intercept (99%) are the estimates of slope and intercept of the best fit line for all subjects except the riskiest 1%. Rsq and Rsq (99%) are goodness of fit measures of individual results to the best fit line for all subjects and all subjects except the riskiest 1% (i.e., green dots), respectively. **(C) Sensitivity/Positive Predictive Value Plot**. Positive predictive value (blue dots) and sensitivity (purple dots) versus the fraction of population, sorted by the risk of in-hospital mortality. The vertical red line indicates where the number of patients above the risk threshold equals the incidence of mortality in the population. Tabulated metrics: the area under the Receiver Operating Characteristic (AUC) and the incidence of mortality (IR). Vertical bars help identify the PPV and sensitivity performance for detectors operating to identify the riskiest 5%, 10%, and 20% patients. The positive predictive accurary and sensitivity are tabulated for these detector operating points. **(D) Enrichment Factor (Lift) Plot.** Lift (i.e., positive predictive value/incidence) versus sensitivity. Vertical bars help identify the lift and sensitivity performance for detectors operating to identify the riskiest 5%, 10%, and 20% patients. Positive predictive value, sensitivity and lift are tabulated for these detector operating points. The AUC and incidence of mortality (IR) are also tabulated.

In-Hospital Mortality



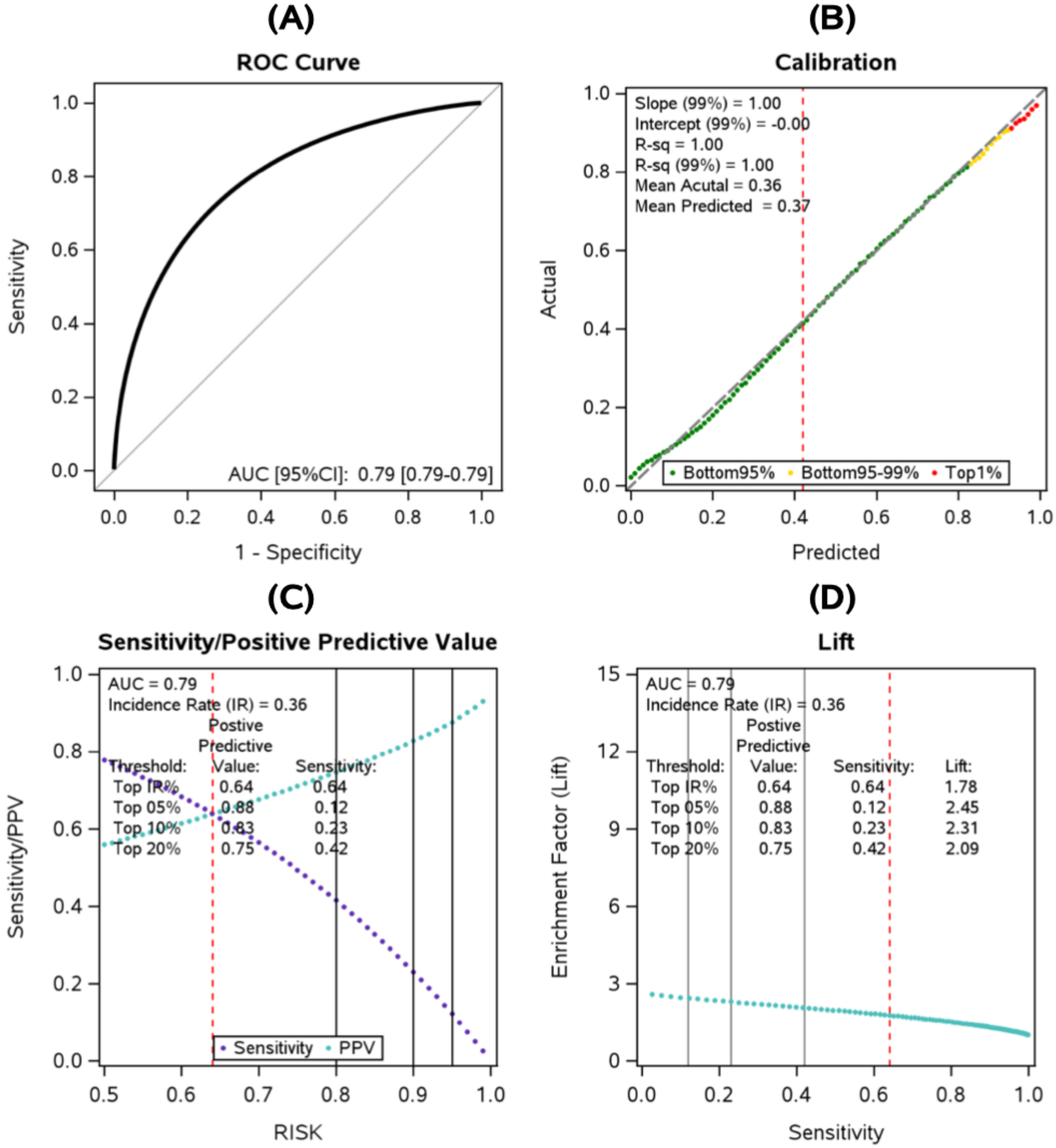
(A) ROC Curve

(B) Calibration

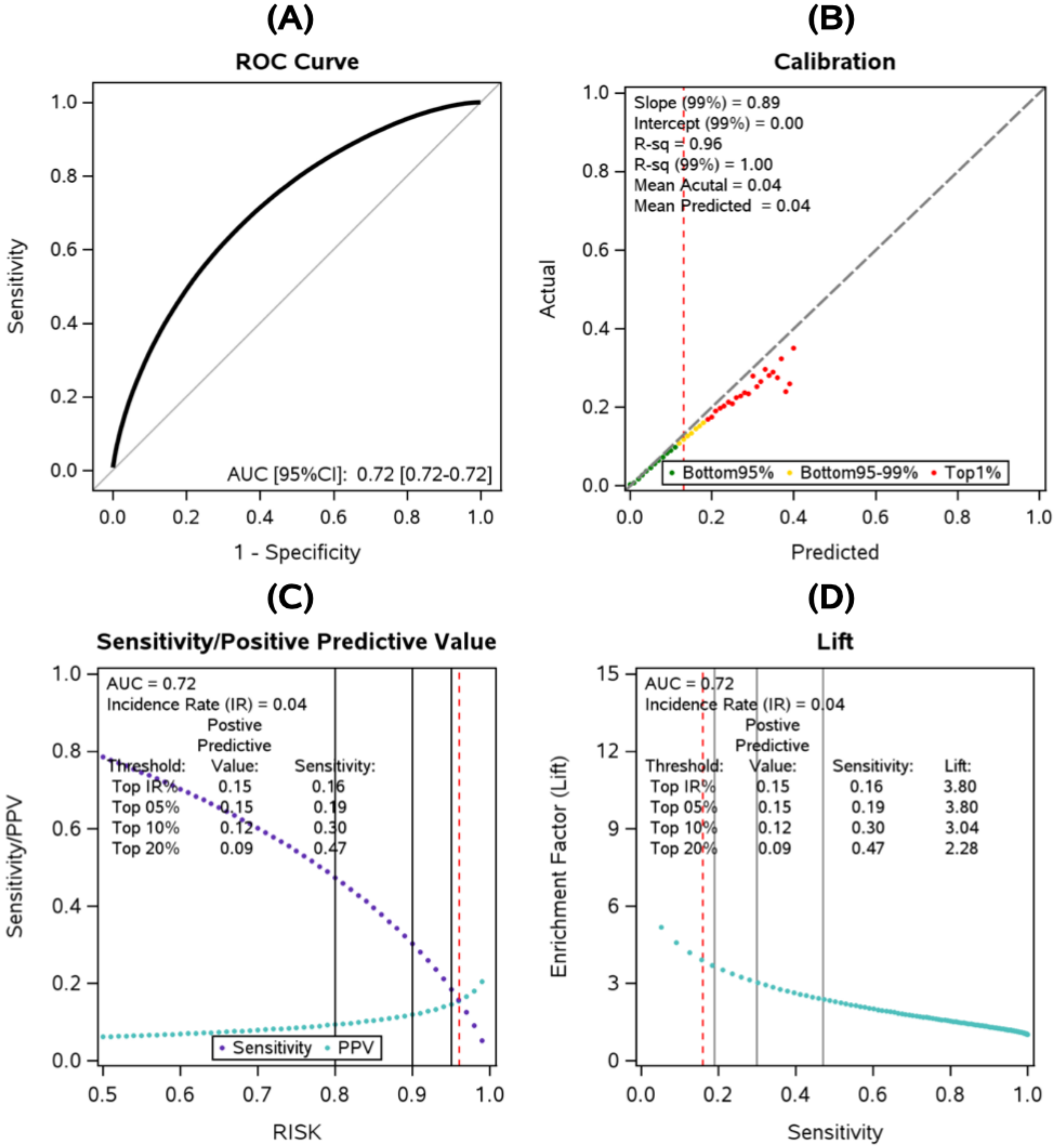
(C) Sensitivity/Positive Predictive Value

(D) Lift

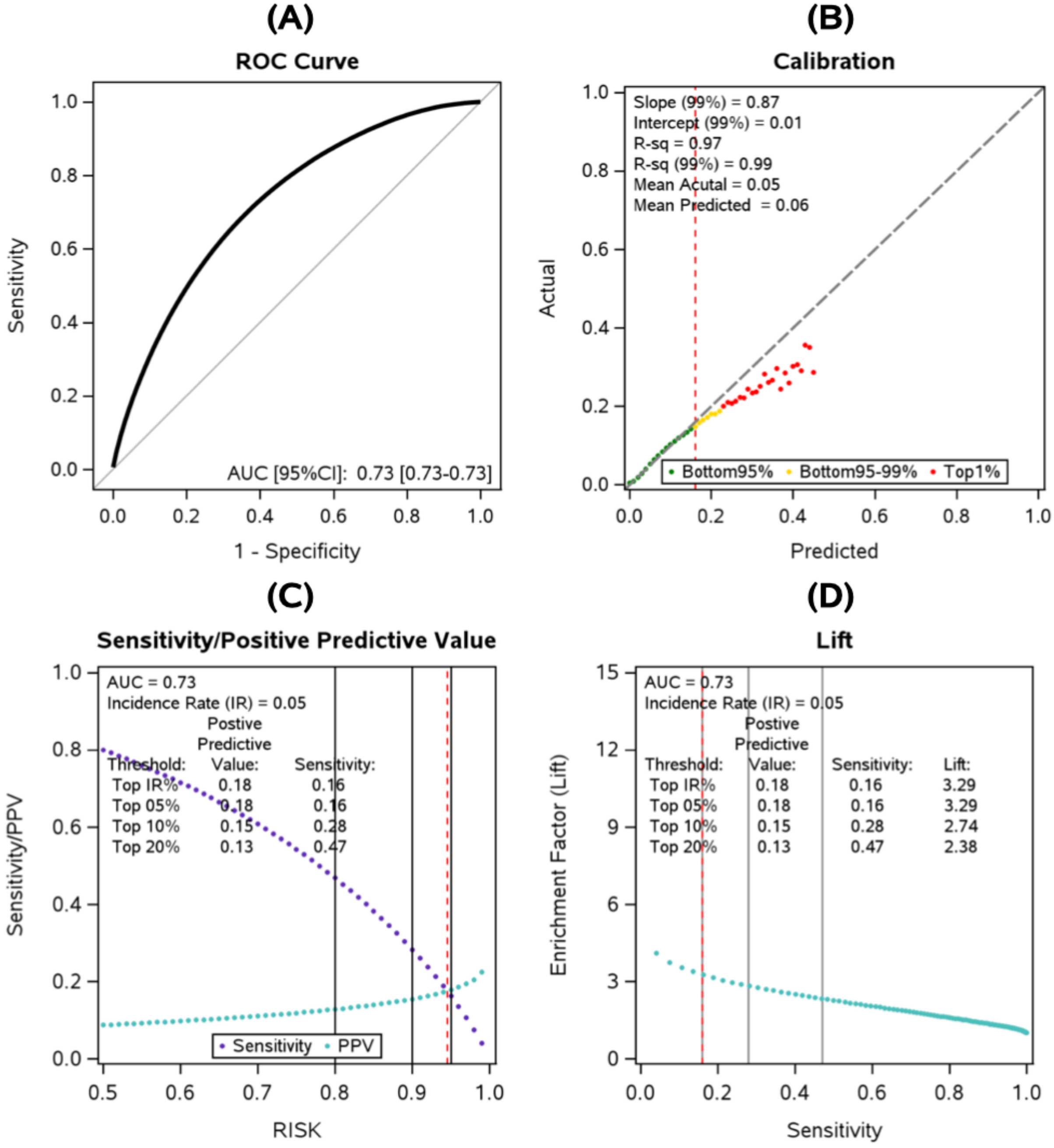
Discharge to Facility



90-Day Pneumonia



90-Day Acute Kidney Injury



(A)

ROC Curve

(B)

Calibration

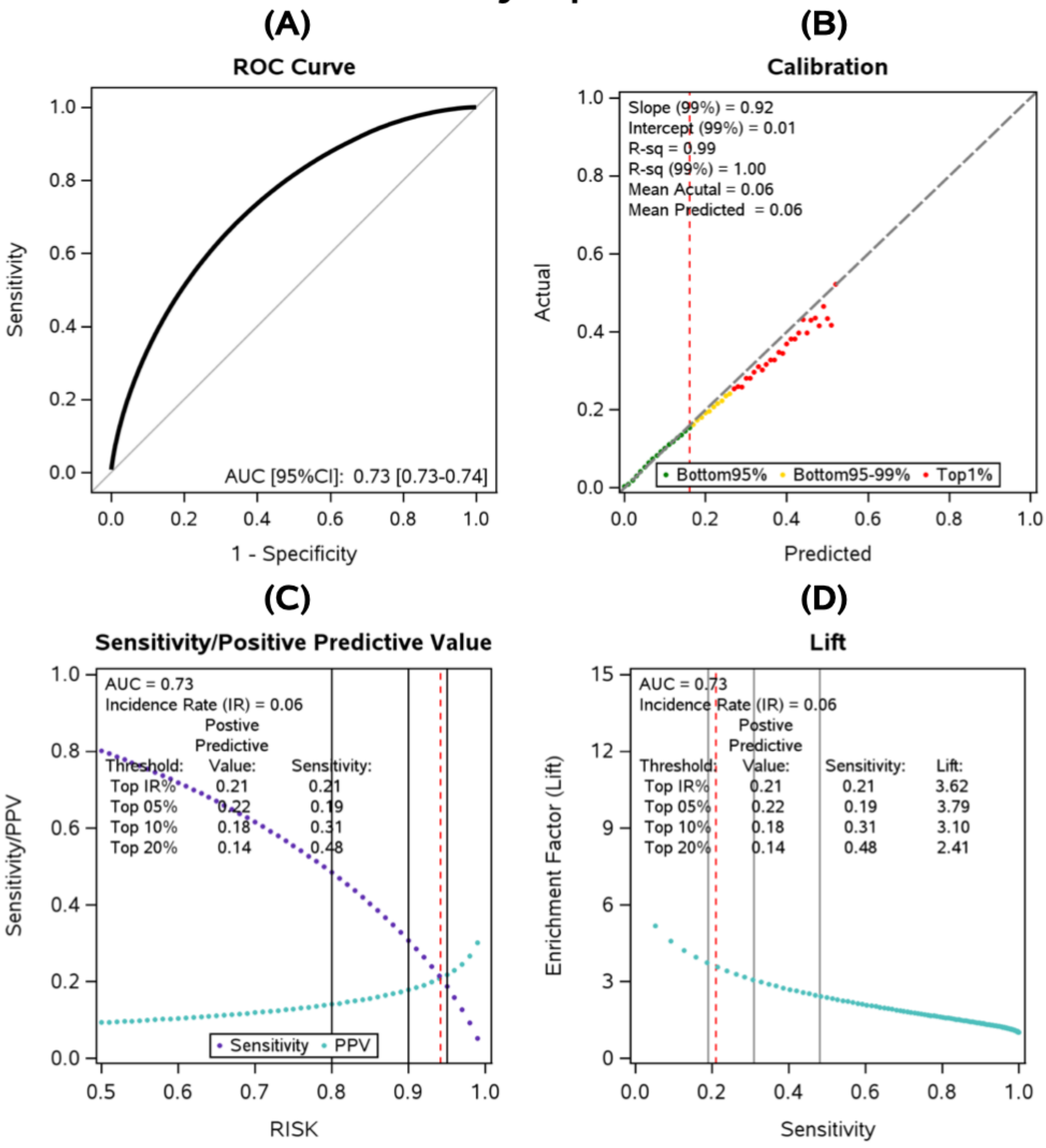
(C)

Sensitivity/Positive Predictive Value

(D)

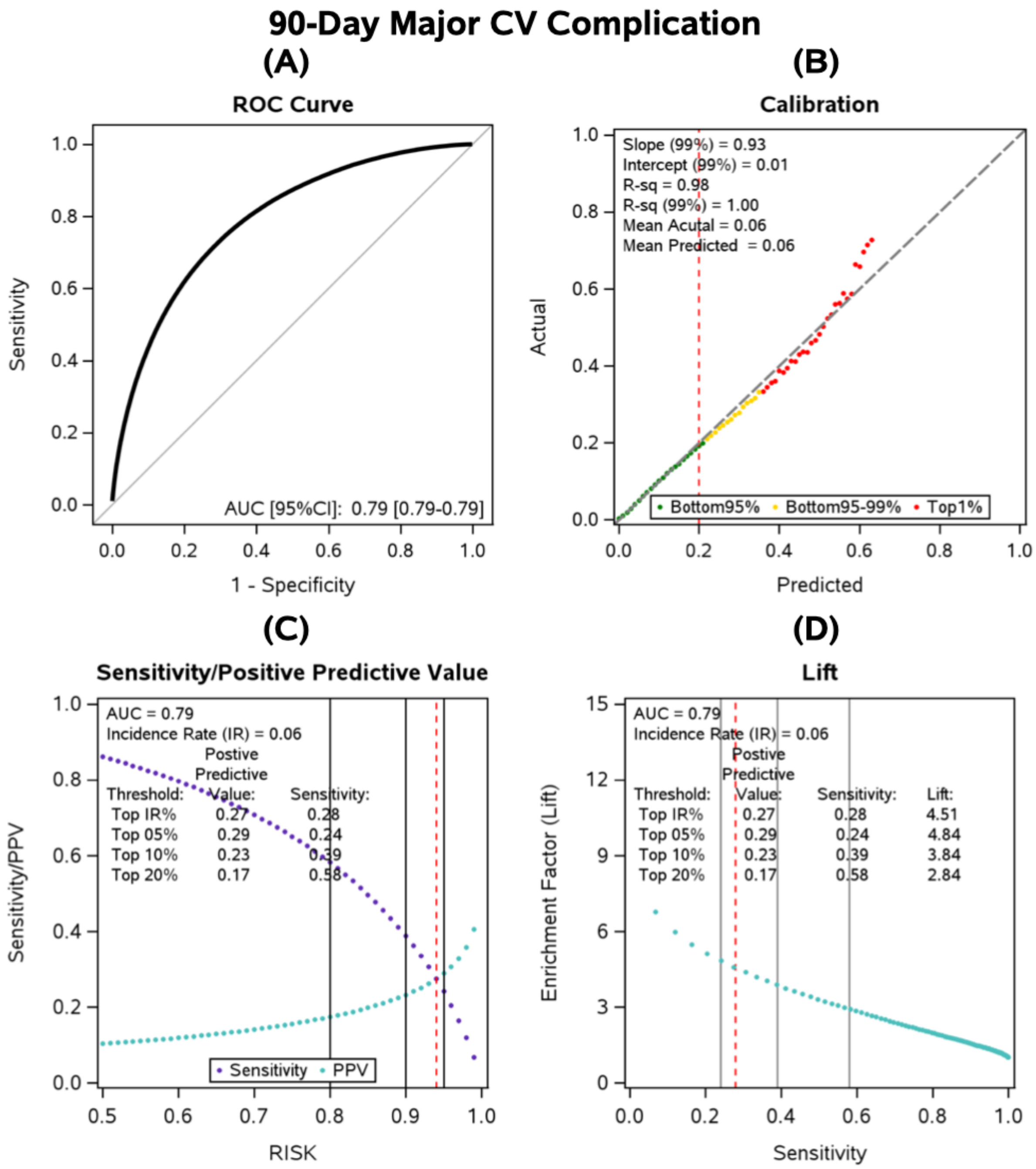
Lift

90-Day Sepsis

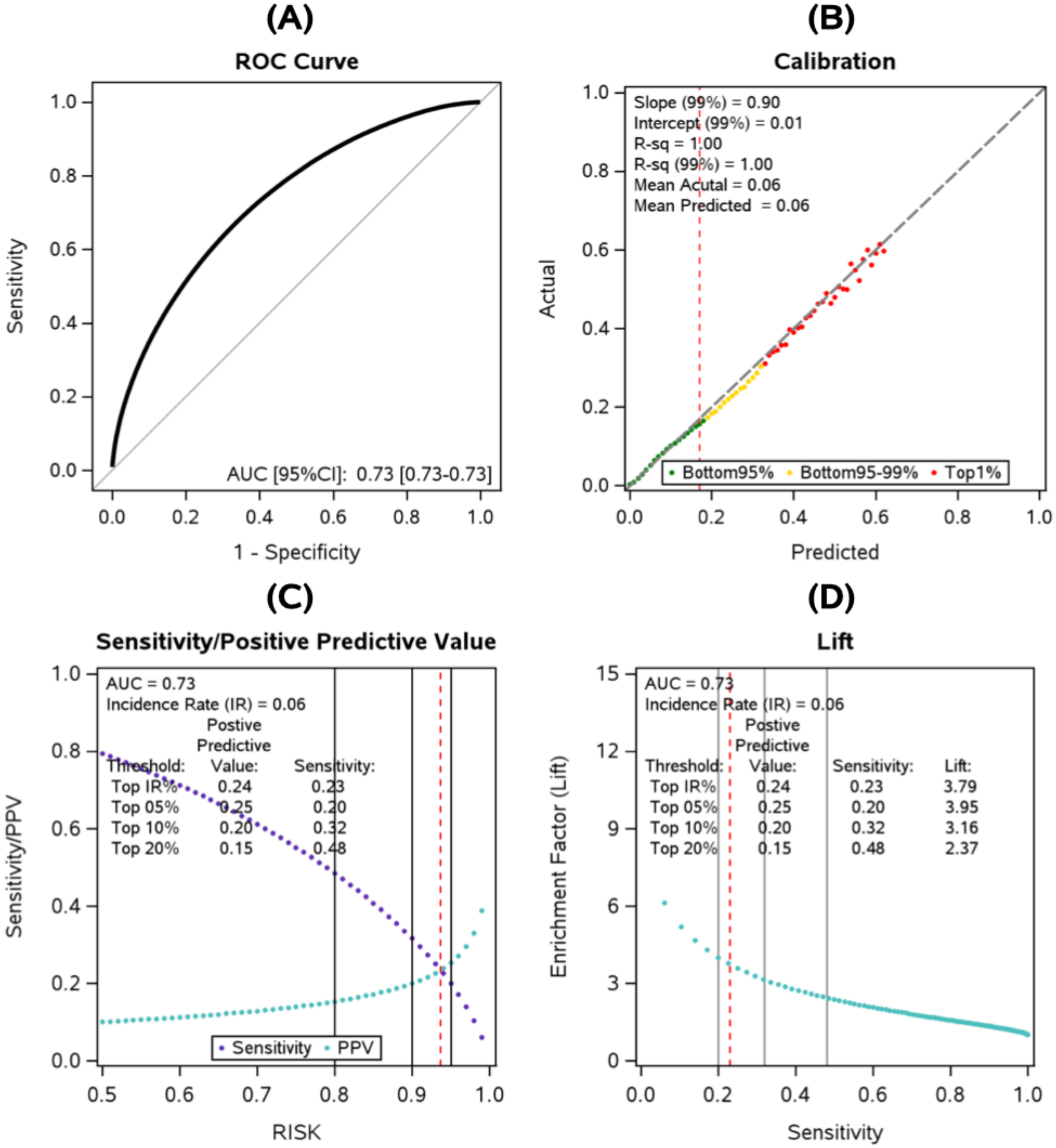


Threshold:	Postive Predictive Value:	Sensitivity:
Top IR%	0.21	0.21
Top 05%	0.22	0.19
Top 10%	0.18	0.31
Top 20%	0.14	0.48

Threshold:	Postive Predictive Value:	Sensitivity:	Lift:
Top IR%	0.21	0.21	3.62
Top 05%	0.22	0.19	3.79
Top 10%	0.18	0.31	3.10
Top 20%	0.14	0.48	2.41


90-Day Major CV Complication
(A)
ROC Curve
Sensitivity
1 - Specificity
0.0
0.2
0.4
0.6
0.8
1.0
AUC [95%CI]: 0.79 [0.79-0.79]
(B)
Calibration
Actual
Predicted
Slope (99%) = 0.93
Intercept (99%) = 0.01
R-sq = 0.98
R-sq (99%) = 1.00
Mean Acutal = 0.06
Mean Predicted = 0.06
● Bottom95% ● Bottom95-99% ● Top1%
(C)
Sensitivity/Positive Predictive Value
Sensitivity/PPV
RISK
0.5
0.6
0.7
0.8
0.9
1.0
AUC = 0.79
Incidence Rate (IR) = 0.06
Threshold: Postive Predictive Value: Sensitivity:
Top IR% 0.27 0.28
Top 05% 0.29 0.24
Top 10% 0.23 0.39
Top 20% 0.17 0.58
● Sensitivity ● PPV
(D)
Lift
Enrichment Factor (Lift)
Sensitivity
0
3
6
9
12
15
AUC = 0.79
Incidence Rate (IR) = 0.06
Threshold: Postive Predictive Value: Sensitivity: Lift:
Top IR% 0.27 0.28 4.51
Top 05% 0.29 0.24 4.84
Top 10% 0.23 0.39 3.84
Top 20% 0.17 0.58 2.84

90-Day Respiratory Failure



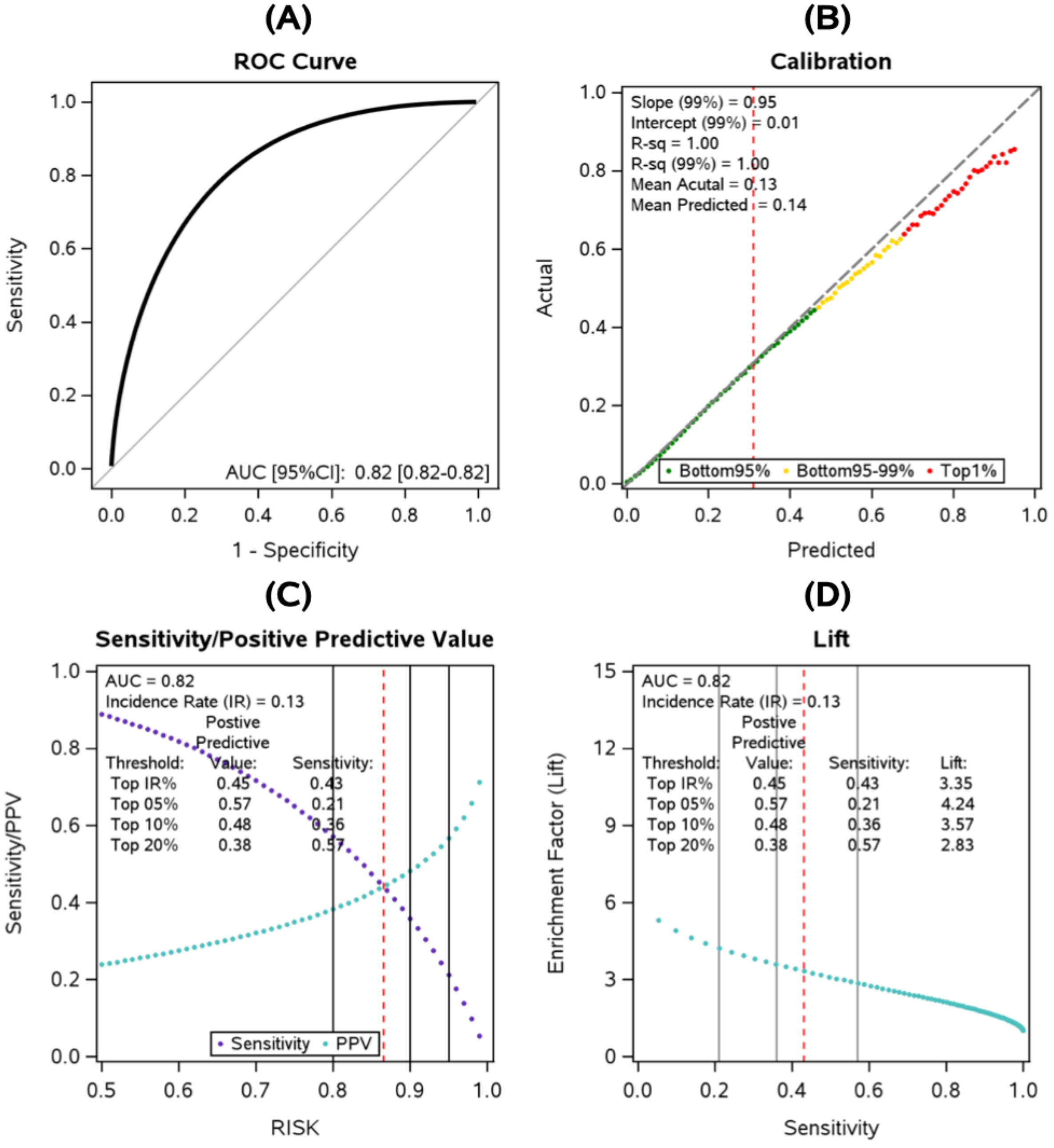
AUC = 0.73
Incidence Rate (IR) = 0.06

Threshold:	Postive Predictive Value:	Sensitivity:
Top IR%	0.24	0.23
Top 05%	0.25	0.20
Top 10%	0.20	0.32
Top 20%	0.15	0.48

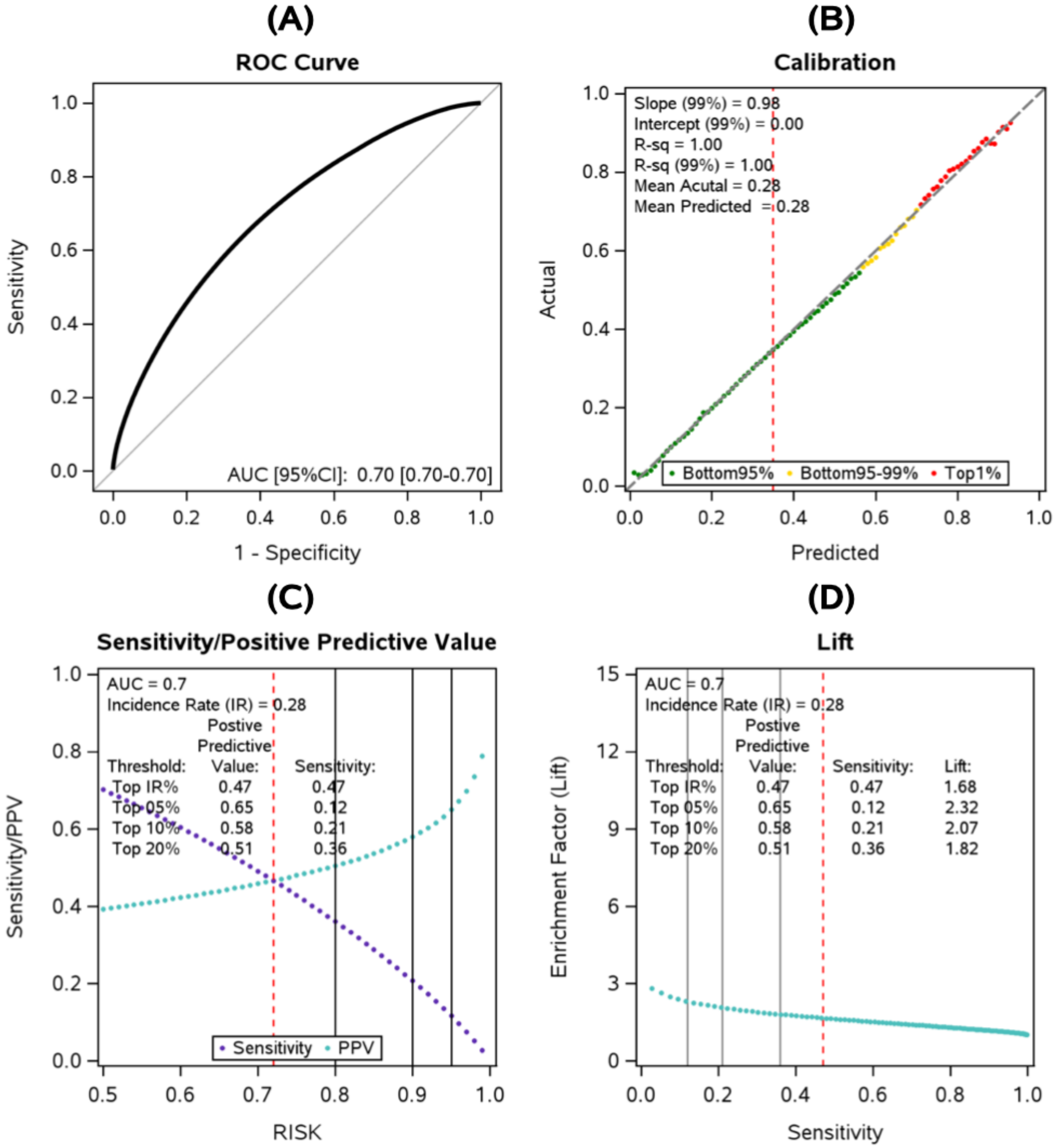
AUC = 0.73
Incidence Rate (IR) = 0.06

Threshold:	Postive Predictive Value:	Sensitivity:	Lift:
Top IR%	0.24	0.23	3.79
Top 05%	0.25	0.20	3.95
Top 10%	0.20	0.32	3.16
Top 20%	0.15	0.48	2.37

90-Day Mortality



90-Day Unplanned Admission


(A)
ROC Curve
Sensitivity
1 - Specificity
AUC [95%CI]: 0.70 [0.70-0.70]
(B)
Calibration
Slope (99%) = 0.98
Intercept (99%) = 0.00
R-sq = 1.00
R-sq (99%) = 1.00
Mean Acutal = 0.28
Mean Predicted = 0.28
Actual
Predicted
Bottom95%
Bottom95-99%
Top1%
(C)
Sensitivity/Positive Predictive Value
AUC = 0.7
Incidence Rate (IR) = 0.28
Threshold: Postive Predictive Value: Sensitivity:
Top IR% 0.47 0.47
Top 05% 0.65 0.12
Top 10% 0.58 0.21
Top 20% 0.51 0.36
Sensitivity/PPV
RISK
Sensitivity
PPV
(D)
Lift
AUC = 0.7
Incidence Rate (IR) = 0.28
Threshold: Postive Predictive Value: Sensitivity: Lift:
Top IR% 0.47 0.47 1.68
Top 05% 0.65 0.12 2.32
Top 10% 0.58 0.21 2.07
Top 20% 0.51 0.36 1.82
Enrichment Factor (Lift)
Sensitivity