Predicting Hospitalizations From Electronic Health Record Data

Kyle Morawski, MD, MPH; Yoni Dvorkis, MPH; and Craig B. Monsen, MD, MS

ABSTRACT

OBJECTIVES: Electronic health record (EHR) data have become increasingly available and may help inform clinical prediction. However, predicting hospitalizations among a diverse group of patients remains difficult. We sought to use EHR data to create and internally validate a predictive model for clinical use in predicting hospitalizations.

STUDY DESIGN: Retrospective observational cohort study.

METHODS: We analyzed EHR data in patients 18 years or older seen at Atrius Health from June 2013 to November 2015. We selected variables among patient demographics, clinical diagnoses, medications, and prior utilization to train a logistic regression model predicting any hospitalization within 6 months and validated the model using a separate validation set. We performed sensitivity analysis on model performance using combinations of EHR-derived, claims-derived, or both EHR- and claims-derived data.

RESULTS: After exclusions, 363,855 patient-months were included for analysis, representing 185,388 unique patients. The strongest features included sickle cell anemia (odds ratio [OR], 52.72), lipidoses and glycogenosis (OR, 8.44), heart transplant (OR, 6.12), and age 76 years or older (OR, 5.32). Model testing showed that EHR-only data had an area under the receiver operating characteristic curve (AUC) of 0.84 (95% CI, 0.838-0.853), which was similar to the claims-only data (AUC, 0.84; 95% CI, 0.831-0.848) and combined claims and EHR data (AUC, 0.846; 95% CI, 0.838-0.853).

CONCLUSIONS: Prediction models using EHR-only, claims-only, and combined data had similar predictive value and demonstrated strong discrimination for which patients will be hospitalized in the ensuing 6 months.

Am J Manag Care. 2020;26(1):e7-e13

The healthcare system generates, collects, and stores a tremendous amount of data during the course of a patient's clinical encounter, with one study finding an average of more than 200,000 individual data points available during a single hospital stay.[1,2] These data are used to monitor a patient's progress, coordinate care among all members of the healthcare team, and provide documentation for billing and reporting activities. Although the use of data for these purposes has been long-standing, the availability of these data has increased substantially. The Health Information Technology for Economic and Clinical Health Act of 2009 was passed in part to assist healthcare professionals' transition to electronic health records (EHRs). A decade later, systematically collected data generated in the course of clinical care have created an opportunity to use such data to improve care practices.[3,4]

Retail entities have put forth strategic investments in data science, often with substantial return.[5,6] Accordingly, using data stored in EHRs to improve the lives of patients and lower total medical costs is one approach to transforming care. Big data, machine learning, and predictive analytics are some of the ways that clinicians hope to anticipate patients' needs and improve outcomes, evidenced by the myriad of organizations working in this field.[7] However, this is an evolving field with improving techniques, accuracy, and actionability of predictions. We need more precise prediction models and better integration of data into clinical care[4,8-10] to focus care resources and, in doing so, provide higher value.[11]

The morbidity[12,13] and healthcare costs[13] associated with hospital admissions underscore the need for hospitalization prevention activities including patient outreach, review of recent discharges, and case management. Unfortunately, acute hospital care needs remain difficult to predict.[9] A recent review evaluating accuracy of EHR-based prediction modeling showed that hospitalization and service utilization were more difficult to predict than mortality or disease-specific outcomes. Whereas mortality and clinical prediction models demonstrated C statistics ranging above 0.8, the discrimination of models built to predict hospitalization and service utilization was lower, at 0.71.[8]

TAKEAWAY POINTS

We aimed to develop a rigorous technique for predicting hospitalizations using data that are already available to most health systems.

- Our research can be used to provide clinicians with a risk score for a given patient, which can help guide care.
- Using the predictive tool, clinicians may be able to more accurately triage patients' concerns and respond to those concerns to prevent worsening of their condition and need for hospitalization.

Several approaches to improve hospitalization prediction exist, such as using new data sources, new variable types, more complete data, more timely data, or more advanced statistical methods. Data sets capable of linking EHR and claims data at the patient level remain uncommon. We hypothesized that when combined, these 2 data sources would complement each other and lead to stronger prediction than that observed previously. We set out to develop and test a model that uses EHR and claims data to predict patient hospitalizations in such a way that it can be implemented in an outpatient practice setting.

METHODS

Study Design

We performed a retrospective analysis of data generated in the course of clinical care and healthcare operations to develop a logistic regression model predicting a patient's future risk of hospitalization. Data were extracted from Atrius Health's unified data warehouse, which marries clinical data from Atrius Health's EHR (Epic version 2015; Epic Systems; Verona, Wisconsin) to normalized administrative claims data received from Medicare, Medicaid, and commercial payers. Variables were ascertained at the patient-month level. To reflect seasonality in hospitalization outcomes, 4 dates of prediction—referred to as index dates—were selected throughout the study period: September 1, 2014; December 1, 2014; March 1, 2015; and June 1, 2015. Sensitivity testing was performed to determine how the inclusion of certain variable categories or data sources (ie, EHR vs claims) would influence model performance. The analysis was performed as part of a quality improvement effort at Atrius Health and did not undergo institutional review board review.

Study Population

The study population was selected among patients seen from June 2013 to November 2015 at Atrius Health, a large multispecialty group in eastern Massachusetts. The population included patients insured under Medicare, Medicaid, and commercial contracts. Patients younger than 18 years were excluded from analysis because adult primary care was the focus of this effort.

Outcome Variable

We selected a binary outcome variable indicating if a patient had experienced any medical/surgical admission within 6 months of the index date of prediction. We chose to predict hospitalizations within 6 months to best match the prediction interval with the timeline of likely future downstream interventions. For example, to assist in the care of a complex patient, a relationship with a case manager is often established. This potential intervention requires a period of time to plausibly affect risk of hospitalization. Longer prediction intervals would potentially dilute the impact of future interventions or else necessitate interventions spanning very long time horizons. We excluded obstetrical admissions because these would not be targets for anticipated interventions.

Feature Development

The initial set of features included 651 variables defined among sociodemographics, diagnoses, medications, and prior utilization of both inpatient and outpatient services.

Sociodemographic variables, such as age, insurance type, body mass index, and smoking status, were for the most part obtained from the EHR. In the case of claims sensitivity testing, age and insurance status were obtained from payer roster files. Missing data were considered as a separate class within each categorical variable.

We aggregated diagnoses among EHR encounter- and claims-level *International Classification of Diseases, Ninth Revision* (*ICD-9*) and *Tenth Revision* (*ICD-10*) codes and mapped them to a smaller set of features by grouping them into 1 of 87 HHS–Hierarchical Condition Categories (HHS-HCC) diagnosis groups.[14] A patient needed just 1 instance of an *ICD-9* or *ICD-10* code within an HHS-HCC group at any point during the retrospective period to ascertain that categorical variable as positive. Missing data were interpreted as the patient not having the clinical condition.

Uses of medications were similarly aggregated by National Drug Code across EHR data and pharmacy claims using commercially available therapeutic class codes (First Databank, Inc; South San Francisco, California). As with diagnoses, just 1 occurrence of an order or a prescription for a medication belonging to a given class was needed to ascertain that categorical variable as positive. Missing data were interpreted as the patient not having used the medication class.

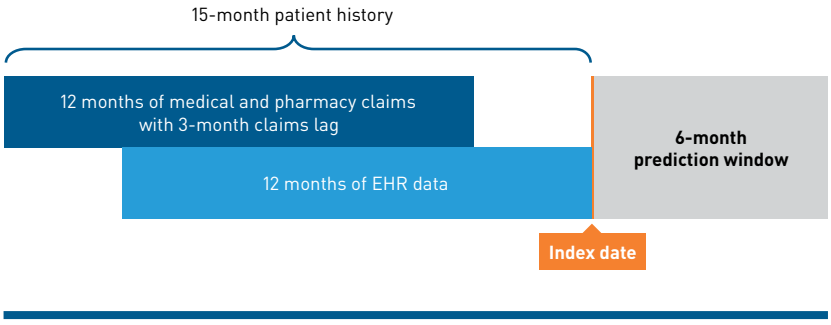
Utilization variables included indicator variables of prior admissions, emergency department (ED) visits, and outpatient visits. These variables were further categorized based on the timing of the occurrence relative to the index date. For example, hospitalization utilization variables included those indicating if the patient had been hospitalized in the past 1 month, hospitalized in the past 1 to 3 months, hospitalized in the past 3 to 6 months, and hospitalized in the past 6 to 12 months.

Any variables that did not occur in more than 30 patient-months in the data set were removed prior to model training to provide stable coefficients for the logistic regression model. For example, if there were just 10 patient-months in the sample during which

any patient was on a medication represented as a binary variable, this variable was dropped from the model.

Although EHR data are readily available within 24 hours of an index date, claims data are often received at a 3-month delay called claims lag. To simulate this claims lag, we ascertained historical variables during a 12-month period starting 15 months prior to the index date up until to 3 months prior to the index date. This avoids advantaging models with data that would not normally be available. Data from the EHR, which do not experience this lag, were obtained during a partially overlapping 12-month period starting 12 months prior to the index date until the day before the index date. This is illustrated in **Figure 1**.

FIGURE 1. Data Schema



EHR indicates electronic health record.

Model Development

We randomly selected 80% of the data to serve as the training set, reserving the remaining 20% of the data as a testing set. We then regressed our selected variables onto our hospitalization outcome using a logistic model with the canonical link. Variables were included in the final model if their odds ratio (OR) was greater than or equal to 1 (see **eAppendix** [available at **ajmc.com**]). This decision was made to be consistent with our organization's goal to identify predictors of increased risk of hospitalization and aid with model interpretability, as clinicians would be appropriately skeptical of a disease state conferring a protective effect. Previous unpublished work informed our approach here, as machine learning algorithms such as random forest, support vector machines, and neural networks did not consistently improve model performance and were less interpretable than the logistic regression approach. This has since been corroborated in recent literature for general outcomes such as mortality and disease-specific outcomes such as HIV incidence.[8,15,16] All analysis was performed in R version 3.2.1 (R Foundation; Vienna, Austria).

Model Performance

We measured performance on the training and testing data sets using area under the receiver operating characteristic curve (AUC) and model calibration.[17] We calculated 95% CIs around the AUC using the DeLong method (R pROC package, version 1.10.0). For model calibration, we plotted calibration curves and calculated the Hosmer-Lemeshow statistic (R Resource Selection Package, version 0.3-2).

Other Statistical Tests

For continuous variables, we report means and SDs. For noncontinuous variables, we report counts and percentages. For normally distributed data, we applied the *t* test. For nonnormally distributed data, we applied the Wilcoxon test. For comparisons between categorical variables, we used the Fisher test.

Sensitivity Testing

Although the canonical model included EHR and claims data, we sought to identify which category of variables most contributed to model performance. We trained 15 models testing 2 dimensions of model characteristics.

The first dimension compared models developed from different data sources: EHR data only, claims data only, or both. The EHR data–only models used information drawn from the EHR (eg, medication use categories were ascertained as positive if the patient had a medication order placed by a provider). The claims data–only models used information drawn from claims (eg, medication use categories were ascertained as positive if a patient had a medication dispense claim in the administrative data). In the models using both data sources, a categorical feature was ascertained to be positive if there was evidence from either the EHR or claims data.

The second dimension considered was variable types. Separate models were trained to include demographic variables only, diagnoses only, medications only, prior utilization only, or all variables combined. Model performance was assessed for training and testing sets using the C statistic.

RESULTS

Study Population

After exclusions, 363,855 patient-months were included for analysis, corresponding to 185,388 unique patients. Selected patient characteristics ascertained by combining EHR and claims data are summarized in **Table 1**. In aggregate, 5% of the study population had been hospitalized within 6 months of an index date.

Model Features

After excluding variables with low counts or protective factors, 169 variables were included in the final model. Diagnoses, demographics, and prior utilization were well represented among the top predictors (**Figure 2**). The features with the highest ORs for predicting future hospitalization were sickle cell anemia (OR, 52.72), lipidoses and glycogenosis (OR, 8.44), heart transplant (OR, 6.12), and age 76

TABLE 1. Selected Characteristics of the Cohort Stratified by Hospitalization Within 6 Months[a]

	Total Patient-Months (%) by Hospitalization Within 6 Months	
	No	Yes
Total patient-months	351,603	12,252
Age category in years		
18-20	14,806 (4.2)	137 (1.1)
21-25	28,146 (8.0)	178 (1.5)
26-30	28,714 (8.2)	143 (1.2)
31-35	30,680 (8.7)	227 (1.9)
36-40	27,238 (7.7)	256 (2.1)
41-45	27,184 (7.7)	364 (3.0)
46-50	29,302 (8.3)	551 (4.5)
51-55	31,303 (8.9)	677 (5.5)
56-60	29,260 (8.3)	848 (6.9)
61-65	25,194 (7.2)	982 (8.0)
66-70	25,959 (7.4)	1362 (11.1)
71-75	19,676 (5.6)	1337 (10.9)
≥76	34,141 (9.7)	5190 (42.4)
Race/ethnicity		
Asian	18,332 (5.2)	251 (2.0)
Black or African American	31,768 (9.0)	1014 (8.3)
Hispanic or Latino	12,681 (3.6)	281 (2.3)
Other	30,274 (8.6)	574 (4.7)
White	258,548 (73.5)	10,132 (82.7)
Insurance type		
Commercial	245,733 (69.9)	3420 (27.9)
Medicaid	26,043 (7.4)	715 (5.8)
Medicare	79,827 (22.7)	8117 (66.3)
Marital status		
Divorced	15,453 (4.4)	864 (7.1)
Married	149,054 (42.4)	5076 (41.4)
Other	123,453 (35.1)	4774 (39.0)
Single	63,643 (18.1)	1538 (12.6)

	Total Patient-Months (%) by Hospitalization Within 6 Months	
	No	Yes
BMI category		
<30	237,741 (67.6)	7447 (60.8)
30-34	59,342 (16.9)	2377 (19.4)
35-40	23,628 (6.7)	1182 (9.6)
>40	14,173 (4.0)	920 (7.5)
Unknown BMI	16,719 (4.8)	326 (2.7)
Prior hospitalization		
Past month	1704 (0.5)	643 (5.2)
1-3 months	3399 (1.0)	944 (7.7)
3-6 months	5732 (1.6)	1245 (10.2)
6-12 months	9693 (2.8)	1711 (14.0)
0 hospitalizations	341,910 (97.2)	10,541 (86.0)
1 hospitalization	8364 (2.4)	1246 (10.2)
≥2 hospitalizations	1329 (0.4)	465 (3.8)
Prior outpatient visits in past 3-6 months		
0	191,396 (54.4)	5638 (46.0)
1	78,306 (22.3)	1911 (15.6)
2	39,388 (11.2)	1411 (11.5)
3-4	29,895 (8.5)	1734 (14.2)
5-8	10,852 (3.1)	1170 (9.5)
≥9	1766 (0.5)	388 (3.2)
Prior ICU stay in past 12 months	2339 (0.7)	735 (6.0)
Select diagnoses		
Diabetes with chronic complications	13,724 (3.9)	2136 (17.4)
Diabetes without complications	17,186 (4.9)	1128 (9.2)
Cirrhosis of liver	724 (0.2)	127 (1.0)
Drug dependence	1961 (0.6)	150 (1.2)
Major depressive and bipolar disorders	19,480 (5.5)	1112 (9.1)
Congestive heart failure	8254 (2.3)	2312 (18.9)
Specified heart arrhythmias	13,527 (3.8)	2622 (21.4)
Chronic obstructive pulmonary disease	8686 (2.5)	1690 (13.8)
Asthma	25,634 (7.3)	986 (8.0)

BMI indicates body mass index; ICU, intensive care unit.

[a]*P* <.001 for all variables.

years or older (OR, 5.32). A full list of final features is included in the eAppendix.

Model Performance

Model discrimination. Model discrimination varied widely, depending primarily on included variables. The predictive model using only prescription medications performed least well, with an AUC of 0.602. The model including all variable types, claims data, and EHR data performed best on the testing set, with an AUC of 0.846. There were no statistical differences in performance on the testing set among the 3 models including all variable types based on claims data alone (AUC, 0.840; 95% CI, 0.832-0.848), EHR data alone (AUC, 0.840; 95% CI, 0.831-0.848), or the claims and EHR data combined (AUC, 0.846; 95% CI, 0.838-0.853). **Table 2** illustrates these results in more detail.

Model calibration. The best-performing model, which included all variable types from claims and EHR data combined, appeared to be well calibrated (**Figure 3**). Predicted probability of hospitalization at 6 months corresponded closely to the observed proportion of hospitalized patients when sorted into 10 bins of equal size

(~7300 patients per bin). Further, the slope of the calibration was 0.96 (95% CI, 0.94-0.98) compared with a perfectly calibrated slope of 1.0. The model overestimated 6-month hospitalizations among those with the highest predicted risk.

DISCUSSION

Principal Findings

Using a combination of EHR and claims data describing patients' demographics, healthcare utilization behavior, medical diagnoses, and medications, we were able to develop a risk score that accurately predicted hospitalization in the ensuing 6 months. Although our results suggest some utility to combining EHR and claims data to inform predictive model creation, we find that even in scenarios in which only EHR or claims data are available, strong performance can be achieved provided that a diverse collection of variable types is represented. A variety of highly predictive characteristics were derived from all major domains evaluated. Consistent with traditional methods, age group was one of the strongest predictors, with the more elderly groups being at higher risk. Prior healthcare utilization was also a strong predictor and likely covaries with many other factors in the model. However, this collinearity improves the variance of the logistic regression approach and may allow unmeasured factors, such as healthcare literacy and choices among individuals of where to seek care, influence in the prediction.[18] Particular medical diagnoses also were found to be predictive, likely indicating frailty and rapid decline in health status that is unable to be adequately managed in the outpatient setting. For example, those with end-stage organ damage (renal or hepatic) have little functional reserve, necessitating precision with both health behaviors and medication adjustments. They are prone to imbalances in fluid or electrolytes that require the care of the inpatient setting for monitoring and correction.

The risk prediction score was also found to be well calibrated in those less likely to be hospitalized in the next 6 months, but it did become less accurate among those at higher risk of hospitalization. The model tended to overestimate the likelihood of hospitalization in those with higher than 30% predicted risk, likely owing to the small number of patients demonstrating such high risk.

Comparison With Prior Work

Although many risk scores have been created for individual disease entities[19] or certain groups of people,[20-24] ours is agnostic of clinical condition or demographic. Past efforts in predicting hospitalization have been limited in addressable ways.[25,26] Whereas other models are updated infrequently, as in the case of the QAdmissions model from the British National

FIGURE 2. Selected Top Predictors of Hospitalization Risk

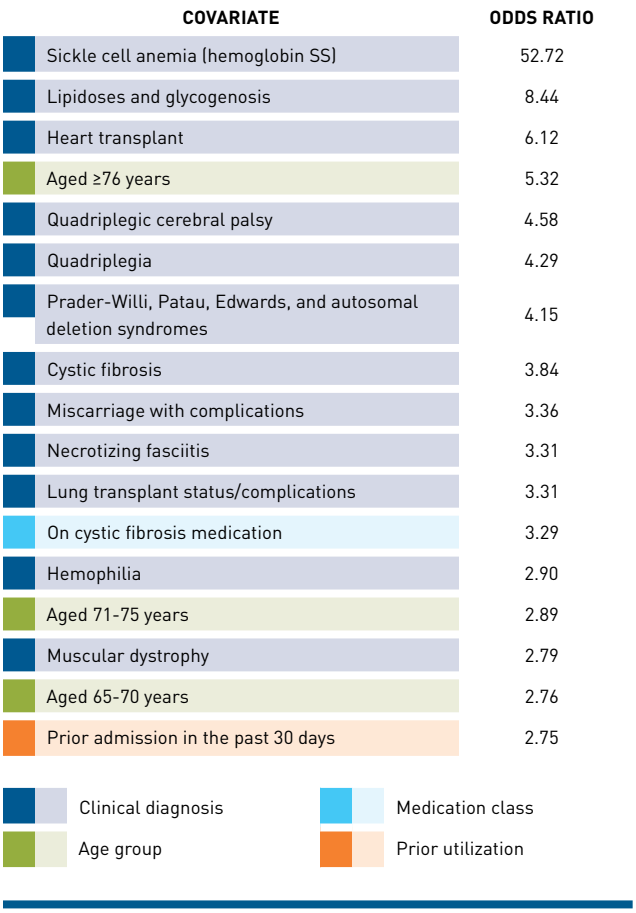
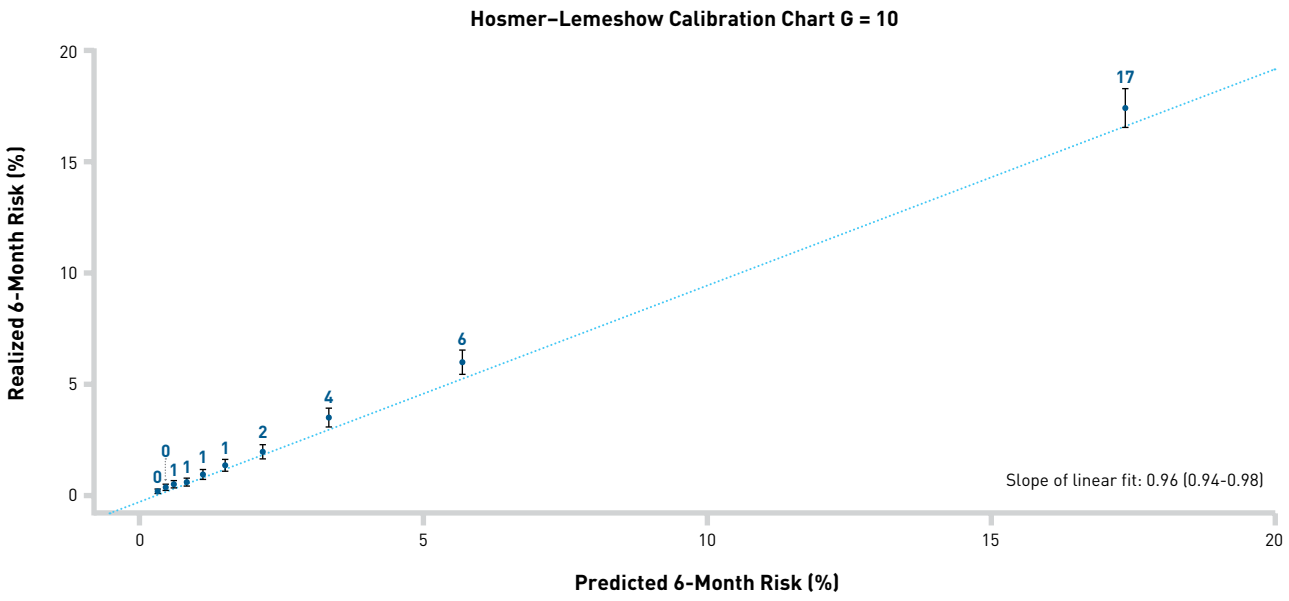


TABLE 2. Model Discrimination

Data Source	Variable Types Included	Training Set: AUC (95% CI)	Testing Set: AUC (95% CI)
Claims and EHR combined	Demographics alone	0.798 (0.795-0.813)	0.796 (0.788-0.805)
	Diagnoses alone	0.763 (0.749-0.772)	0.762 (0.751-0.772)
	Medications alone	0.742 (0.726-0.750)	0.746 (0.735-0.757)
	Utilization alone	0.749 (0.728-0.752)	0.749 (0.738-0.760)
	All together	0.848 (0.844-0.860)	0.846 (0.838-0.853)
EHR only	Demographics alone	0.798 (0.786-0.805)	0.796 (0.788-0.805)
	Diagnoses alone	0.735 (0.717-0.741)	0.737 (0.726-0.747)
	Medications alone	0.737 (0.729-0.753)	0.743 (0.732-0.754)
	Utilization alone	0.743 (0.740-0.763)	0.738 (0.727-0.749)
	All together	0.843 (0.845-0.862)	0.840 (0.831-0.848)
Claims only	Demographics alone	0.786 (0.778-0.797)	0.787 (0.778-0.795)
	Diagnoses alone	0.733 (0.721-0.745)	0.732 (0.721-0.742)
	Medications alone	0.602 (0.588-0.612)	0.602 (0.591-0.612)
	Utilization alone	0.724 (0.713-0.738)	0.721 (0.710-0.733)
	All together	0.843 (0.830-0.847)	0.840 (0.832-0.848)

AUC indicates area under the receiver operating characteristic curve; EHR, electronic health record.

FIGURE 3. Model Calibration



Health Service that is updated quarterly,[27] the present model may be updated weekly to provide more timely information across a range of clinical applications. Another model uses a clinician's assessment to ask whether a patient is likely to be seen in the emergency ward,[25] whereas ours uses a multimodal, data-derived approach to create the risk prediction. Additionally, our model's C statistic of 0.846 compares favorably with those of previous models (0.67-0.77), which we attribute to its incorporation of a wide array of variables (demographics, clinical diagnoses, medications, and prior utilization). We believe that our model adds to the current literature by providing an example of EHR and claims data utilization that can routinely and in real time provide risk prediction for hospitalization among patients seen in a primary care setting.

Limitations

Our investigation has limitations. First, the retrospective analysis was performed using data from a single health system without an external center to validate our results. Although this threatens the generalizability of the model results, we believe the approach is one that can be reproduced at other centers to derive a more tailored model that reflects local patients, patient features, and care practices, all of which may also influence the risk of hospitalization. For instance, ED visits may occur with different frequencies and in different clinical scenarios in other parts of the country due to geographical characteristics of care providers. Other regions may have differing access to outpatient care, which may result in lower-acuity situations escalating to inpatient care. It is worth noting that we used data representing a large, diverse patient population, which offers some stability to the model coefficients and results. That said, we would expect that a given health system could apply these methods to calibrate the model for its own patients and system of care.

After creating our model, we used an internal validation strategy, testing its predictive ability on 20% of the data that were withheld during model creation. Other methods of validation include bootstrapping[28] and external validation.[29] We felt that the training/testing set approach was a sufficiently accurate and interpretable method for measuring discrimination, and we observe that it is commonly used in the literature.[30] Because these efforts were performed to improve the quality of care in a single health system, future research work would be helpful to validate our approach on an external population.[31]

The extent to which our predictive model can better target particular interventions and improve care remains to be proven. First, the strongest covariates in the model were those that are nonmodifiable, such as clinical diagnoses. For example, somebody with sickle cell anemia or a heart transplant cannot modify those factors. Second, for factors that are modifiable, such as medication use, the coefficients derived are correlative, not causative. One must be careful not to interpret the fact that a patient is on a medication associated with hospitalization to mean that the medication is a cause of future hospitalization. The net of this is that although identifying highest-risk patients seems a natural approach to prioritize interventions such as postdischarge education and case management, our model provides no evidence that such patients are amenable to these interventions or that their risk of hospitalization would be responsive to them.

Despite these limitations, we believe that our model approach is a meaningful step toward identifying patients at highest risk of hospitalization. Tying the model to care interventions that are likely to modify the risk of hospitalization represents a promising area for future research.

CONCLUSIONS

Prediction models using EHR-only, claims-only, and combined data had similar predictive value and demonstrated strong discrimination for which patients will be hospitalized in the ensuing 6 months. The resulting model offers additional benefits of interpretability and timeliness and may be reproduced with local data for greater accuracy. ■

Author Affiliations: Atrius Health (KM, YD, CBM), Newton, MA.

Source of Funding: Atrius Health institutional funding.

Author Disclosures: The authors report no relationship or financial interest with any entity that would pose a conflict of interest with the subject matter of this article.

Authorship Information: Concept and design (YD, CBM); acquisition of data (YD); analysis and interpretation of data (YD, CBM); drafting of the manuscript (KM, YD, CBM); critical revision of the manuscript for important intellectual content (KM, YD, CBM); statistical analysis (YD, CBM); administrative, technical, or logistic support (CBM); and supervision (KM, CBM).

Address Correspondence to: Kyle Morawski, MD, MPH, Atrius Health, 133 Brookline Ave, Boston, MA 02215. Email: Kyle_morawski@atriushealth.org.

REFERENCES

1. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. *Health Inf Sci Syst*. 2014;2:3. doi: 10.1186/2047-2501-2-3.
2. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digit Med*. 2018;1:18. doi: 10.1038/s41746-018-0029-1.
3. Parikh RB, Kakad M, Bates DW. Integrating predictive analytics into high-value care: the dawn of precision delivery. *JAMA*. 2016;315(7):651-652. doi: 10.1001/jama.2015.19417.
4. Roski J, Bo-Linn GW, Andrews TA. Creating value in health care through big data: opportunities and policy implications. *Health Aff (Millwood)*. 2014;33(7):1115-1122. doi: 10.1377/hlthaff.2014.0147.
5. LaRiviere J, McAfee P, Rao J, Narayanan VK, Sun W. Where predictive analytics is having the biggest impact. *Harvard Business Review* website. hbr.org/2016/05/where-predictive-analytics-is-having-the-biggest-impact. Published May 25, 2016. Accessed September 11, 2017.
6. Siegel E. *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die*. Hoboken, NJ: John Wiley & Sons, Inc; 2013.
7. Hileman G, Steele S. Accuracy of claims-based risk scoring models. Society of Actuaries website. soa.org/globalassets/assets/files/research/research-2016-accuracy-claims-based-risk-scoring-models.pdf. Published October 2016. Accessed February 5, 2019.
8. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JPA. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *J Am Med Inform Assoc*. 2017;24(1):198-208. doi: 10.1093/jamia/ocw042.
9. Monsen KA, Swanberg HL, Oancea SC, Westra BL. Exploring the value of clinical data standards to predict hospitalization of home care patients. *Appl Clin Inform*. 2012;3(4):419-436. doi: 10.4338/ACI-2012-05-RA-0016.
10. Snooks H, Bailey-Jones K, Burge-Jones D, et al. Predictive risk stratification model: a randomised stepped-wedge trial in primary care (PRISMATIC). *Health Serv Deliv Res*. 2018;6(1):1-164.
11. Berwick DM, Nolan TW, Whittington J. The Triple Aim: care, health, and cost. *Health Aff (Millwood)*. 2008;27(3):759-769. doi: 10.1377/hlthaff.27.3.759.
12. Reynolds MR, Morais E, Zimetbaum P. Impact of hospitalization on health-related quality of life in atrial fibrillation patients in Canada and the United States: results from an observational registry. *Am Heart J*. 2010;160(4):752-758. doi: 10.1016/j.ahj.2010.06.034.
13. Krumholz HM. Post-hospital syndrome—an acquired, transient condition of generalized risk. *N Engl J Med*. 2013;368(2):100-102. doi: 10.1056/NEJMp1212324.
14. Kautter J, Pope GC, Ingber M, et al. The HHS-HCC risk adjustment model for individual and small group markets under the Affordable Care Act. *Medicare Medicaid Res Rev*. 2014;4(3). doi: 10.5600/mmrr2014-004-03-a03.
15. Krakower DS, Gruber S, Hsu K, et al. Development and validation of an automated HIV prediction algorithm to identify candidates for pre-exposure prophylaxis: a modelling study. *Lancet HIV*. 2019;6(10):e696-e704. doi: 10.1016/S2352-3018(19)30139-0.
16. Sahni N, Simon G, Arora R. Development and validation of machine learning models for prediction of 1-year mortality utilizing electronic medical record data available at the end of hospitalization in multicondition patients: a proof-of-concept study. *J Gen Intern Med*. 2018;33(6):921-928. doi: 10.1007/s11606-018-4316-y.
17. Gareth J, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning: With Applications in R*. 8th ed. New York, NY: Springer; 2017.
18. Rasu RS, Bawa WA, Suminski R, Snella K, Warady B. Health literacy impact on national healthcare utilization and expenditure. *Int J Health Policy Manag*. 2015;4(11):747-755. doi: 10.15171/ijhpm.2015.151.
19. Álvarez-García J, Ferrero-Gregori A, Puig T, et al; Investigators of the Spanish Heart Failure Network (REDINSCOR). A simple validated method for predicting the risk of hospitalization for worsening of heart failure in ambulatory patients: the Redin-SCORE. *Eur J Heart Fail*. 2015;17(8):818-827. doi: 10.1002/ejhf.287.
20. Inouye SK, Zhang Y, Jones RN, et al. Risk factors for hospitalization among community-dwelling primary care older patients: development and validation of a predictive model. *Med Care*. 2008;46(7):726-731. doi: 10.1097/MLR.0b013e3181649426.
21. Tabak YP, Sun X, Nunez CM, Gupta V, Johannes RS. Predicting readmission at early hospitalization using electronic clinical data: an early readmission risk score. *Med Care*. 2017;55(3):267-275. doi: 10.1097/MLR.0000000000000654.
22. Morris JN, Howard EP, Steel K, et al. Predicting risk of hospital and emergency department use for home care elderly persons through a secondary analysis of cross-national data. *BMC Health Serv Res*. 2014;14:519. doi: 10.1186/s12913-014-0519-z.
23. Coleman EA, Wagner EH, Grothaus LC, Hecht J, Savarino J, Buchner DM. Predicting hospitalization and functional decline in older health plan enrollees: are administrative data as accurate as self-report? *J Am Geriatr Soc*. 1998;46(4):419-425. doi: 10.1111/j.1532-5415.1998.tb02460.x.
24. Kansagara D, Englander H, Salanitro A, et al. Risk prediction models for hospital readmission: a systematic review. *JAMA*. 2011;306(15):1688-1698. doi: 10.1001/jama.2011.1515.
25. Hwang AS, Ashburner JM, Hong CS, He W, Atlas SJ. Can primary care physicians accurately predict the likelihood of hospitalization in their patients? *Am J Manag Care*. 2017;23(4):e127-e128.
26. Haas LR, Takahashi PY, Shah ND, et al. Risk-stratification methods for identifying patients for care coordination. *Am J Manag Care*. 2013;19(9):725-732.
27. Hippisley-Cox J, Coupland C. Predicting risk of emergency admission to hospital using primary care data: derivation and validation of QAdmissions score. *BMJ Open*. 2013;3(8):e003482. doi: 10.1136/bmjopen-2013-003482.
28. O'Mahony C, Jichi F, Pavlou M, et al; Hypertrophic Cardiomyopathy Outcomes Investigators. A novel clinical risk prediction model for sudden cardiac death in hypertrophic cardiomyopathy (HCM risk-SCD). *Eur Heart J*. 2014;35(30):2010-2020. doi: 10.1093/eurheartj/eht439.
29. Markaki M, Tsamardinos I, Langhammer A, Lagani V, Hveem K, Røe OD. A validated clinical risk prediction model for lung cancer in smokers of all ages and exposure types: a HUNT study. *EBioMedicine*. 2018;31:36-46. doi: 10.1016/j.ebiom.2018.03.027.
30. Steyerberg EW, Harrell FE Jr, Borsboom GJJM, Eijkemans MJC, Vergouwe Y, Habbema JDF. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol*. 2001;54(8):774-781. doi: 10.1016/s0895-4356(01)00341-9.
31. Steyerberg EW, Harrell FE Jr. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol*. 2016;69:245-247. doi: 10.1016/j.jclinepi.2015.04.005.

Visit **ajmc.com/link/4438** to download PDF and eAppendix

eAppendix Table

Variables used in combined EHR and claims logistic regression model..

Covariate Name	OR	p value	
Sickle Cell Anemia (Hb-SS)	52.71793	0.000	***
Lipidoses and Glycogenosis	8.439548	0.000	***
Heart Transplant	6.115212	0.000	***
Age Category (75,120]	5.323958	0.000	***
Quadriplegic Cerebral Palsy	4.579942	0.001	**
Quadriplegia	4.29012	0.000	***
Prader-Willi, Patau, Edwards, and Autosomal Deletion Syndromes	4.151376	0.011	*
Cystic Fibrosis	3.840861	0.013	*
Miscarriage with Complications	3.363122	0.235	
Necrotizing Fasciitis	3.314922	0.127	
Lung Transplant Status/Complications	3.314348	0.125	
Medication for Cystic Fibrosis	3.294892	0.000	***
Hemophilia	2.895673	0.140	
Age Category (70,75]	2.891736	0.000	***
Muscular Dystrophy	2.791991	0.014	*
Age Category(65,70]	2.756039	0.000	***
Prior Hospitalization within 1 Month	2.746199	0.000	***
Age Category (60,65]	2.688919	0.000	***
Medication Omalizumab	2.655246	0.202	
End Stage Renal Disease	2.373742	0.000	***
Metastatic Cancer	2.179432	0.000	***
Age Category (55,60]	2.078689	0.000	***
Chronic Pancreatitis	2.023375	0.000	***
Down Syndrome, Fragile X, Other Chromosomal Anomalies, and Congenital Malformation Syndromes	1.997219	0.003	**
End-Stage Liver Disease	1.991574	0.000	***
Medication Erythropoietin	1.878834	0.002	**
Hypoplastic Left Heart Syndrome and Other Severe Congenital Heart Disorders	1.878607	0.567	
Stem Cell, Including Bone Marrow, Transplant Status/Complications	1.866307	0.012	*
BMI Category > 40	1.753408	0.000	***
Emergency Department Visit in Last Month	1.731379	0.000	***
Frequent Outpatient Visits in Last 3-6 Months Category (8,1e+03]	1.707119	0.000	***
Disease Modifying Anti-Rheumatologic Drug	1.686556	0.009	**
Amputation Status, Lower Limb/Amputation Complications	1.683726	0.005	**
Age Category (50,55]	1.677519	0.000	***
Anorexia/Bulimia Nervosa	1.656824	0.084	
Hospitalization in Last 3-6 Months	1.648454	0.000	***

Paraplegia	1.636469	0.075	
Medicine Growth Hormone	1.608941	0.688	
Non-Hodgkin's Lymphomas and Other Cancers and Tumors	1.595556	0.000	***
Age Category (45,50]	1.588617	0.000	***
Medicare Insurance	1.564792	0.000	***
Emergency Department Visit in Last 1-3 Months	1.553831	0.000	***
Artificial Openings for Feeding or Elimination	1.53228	0.000	***
Frequent Outpatient Visits in Last 1 Month Category(2,3]	1.517004	0.000	***
Myelodysplastic Syndromes and Myelofibrosis	1.515705	0.013	*
Aplastic Anemia	1.501304	0.340	
Inflammatory Bowel Disease	1.495952	0.000	***
Medicaid Insurance	1.48954	0.000	***
Lung, Brain, and Other Severe Cancers, Including Pediatric Acute Lymphoid Leukemia	1.488767	0.000	***
Cirrhosis of Liver	1.475679	0.002	**
Hydrocephalus	1.472466	0.033	*
Colorectal, Breast (Age < 50), Kidney, and Other Cancers	1.4717	0.000	***
Autistic Disorder	1.467215	0.448	
Medication Creon	1.460441	0.080	
Medication Alzheimer Agent	1.439566	0.000	***
Kidney Transplant Status	1.436893	0.134	
Frequent Outpatient Visits in Last 1 Month Category (3,1e+03]	1.425477	0.000	***
Frequent Outpatient Visits in Last 3-6 Months Category (4,8]	1.414148	0.000	***
Medication Multiple Sclerosis Agent	1.412587	0.277	
Parkinson's, Huntington's, and Spinocerebellar Disease, and Other Neurodegenerative Disorders	1.410698	0.000	***
Bone/Joint/Muscle Infections/Necrosis	1.408601	0.001	**
Medication Immunosuppressant	1.402747	0.019	*
Drug Dependence	1.37458	0.004	**
Acute Liver Failure/Disease, Including Neonatal Hepatitis	1.37293	0.175	
Smoker	1.370368	0.000	***
Medication Loop Diuretic	1.361409	0.000	***
Chronic Kidney Disease, Severe (Stage 4)	1.359154	0.000	***
Peritonitis/Gastrointestinal Perforation/Necrotizing Enterocolitis	1.356138	0.066	
Disorders of the Immune Mechanism	1.352204	0.007	**
No Show Category (4,6]	1.339778	0.000	***
Medication Amiodarone	1.338088	0.001	***
Drug Psychosis	1.337407	0.107	
No Show Category (2,4]	1.335649	0.000	***
Prior Emergency Department Visit in Last 6-12 Months (1,100]	1.335611	0.000	***
Medication Opioid	1.326979	0.000	***

BMI Category 35 - 40	1.314562	0.000	***
Prior Hospitalization in Last 6-12 Months (0,1]	1.299306	0.000	***
Osteogenesis Imperfecta and Other Osteodystrophies	1.293402	0.560	
Medication for Hepatitis C Virus	1.286659	0.361	
Smoking Status Other	1.283733	0.010	*
Medication Steroid	1.282474	0.000	***
Respirator Dependence/Tracheostomy Status	1.281544	0.349	
Frequent Outpatient Visits in Last 1 Month Category (1,2]	1.270009	0.000	***
Seizure Disorders and Convulsions	1.266444	0.001	***
Amyloidosis, Porphyria, and Other Metabolic Disorders	1.263766	0.260	
Congestive Heart Failure	1.262549	0.000	***
Diabetes with Chronic Complications	1.256726	0.000	***
Medication Antipsychotic	1.250484	0.000	***
No Show Category (1,2]	1.244587	0.000	***
Hospitalization in Last 3-6 Months	1.235502	0.000	***
Intestinal Obstruction	1.230755	0.015	*
Age Category (40,45]	1.229785	0.071	
Cerebral Aneurysm and Arteriovenous Malformation	1.226791	0.141	
Frequent Outpatient Visits in Last 3-6 Months Category [2,4]	1.226348	0.000	***
HIV/AIDS	1.222857	0.301	
No Show Category (6,100]	1.220685	0.030	*
Multiple Sclerosis	1.218816	0.268	
Breast (Age 50+) and Prostate Cancer, Benign/Uncertain Brain Tumors, and Other Cancers and Tumors	1.216918	0.000	***
Ischemic or Unspecified Stroke	1.210894	0.001	**
Chronic Kidney Disease, Stage 5	1.209473	0.376	
Medication for Parkinson’s Disease	1.209271	0.013	*
Medication for Depression	1.208949	0.000	***
Chronic Obstructive Pulmonary Disease, Including Bronchiectasis	1.199998	0.000	***
Emergency Department Visit in Last 3-6 Months	1.195912	0.000	***
Specified Heart Arrhythmias	1.193885	0.000	***
Unstable Angina and Other Acute Ischemic Heart Disease	1.192848	0.010	*
Atrial and Ventricular Septal Defects, Patent Ductus Arteriosus, and Other Congenital Heart/Circulatory Disorders	1.190884	0.144	
Frequent Outpatient Visits in Last 1 Month Category (0,1]	1.185211	0.000	***
Chronic Ulcer of Skin, Except Pressure	1.168494	0.020	*
Medication for Seizures	1.164102	0.000	***
Emergency Department Visit in Last 6-12 Months	1.156446	0.000	***
Hospitalization in Last 6-12 Months	1.15427	0.067	
Hip Fractures and Pathological Vertebral or Humerus Fractures	1.153932	0.095	

Acute Pancreatitis/Other Pancreatic Disorders and Intestinal Malabsorption	1.151537	0.054	
Spinal Cord Disorders/Injuries	1.148862	0.313	
Medication Inhaled Anti-Cholinergic	1.147153	0.020	*
Medication GCSF	1.139953	0.575	
Frequent Outpatient Visits in Last 1-3 Months Category (1,2]	1.133237	0.001	***
Frequent Outpatient Visits in Last 3-6 Months Category (1,2]	1.130698	0.001	***
Unknown BMI	1.128808	0.111	
Frequent Outpatient Visits in Last 1-3 Months Category (4,8]	1.126941	0.007	**
Frequent Outpatient Visits in Last 1-3 Months Category (0,1]	1.116216	0.000	***
Cardio-Respiratory Failure and Shock, Including Respiratory Distress Syndromes	1.116085	0.107	
Diabetes without Complication	1.11408	0.006	**
Previous Smoker	1.101892	0.000	***
Respiratory Arrest	1.100168	0.861	
Medication Insuline	1.096623	0.076	
Frequent Outpatient Visits in Last 1-3 Months Category (2,4]	1.094936	0.015	*
Frequent Outpatient Visits in Last 3-6 Months Category (0,1]	1.092356	0.005	**
Adrenal, Pituitary, and Other Significant Endocrine Disorders	1.089472	0.139	
Medication Inhaled Steroid	1.081609	0.173	
No Show Category (0,1]	1.074999	0.016	*
BMI 30 - 34	1.072585	0.017	*
Intracranial Hemorrhage	1.071832	0.556	
ICU Stay in Last 12 Months	1.071463	0.270	
Medication Montelukast	1.071407	0.348	
Major Congenital Heart/Circulatory Disorders	1.064123	0.639	
Acquired Hemolytic Anemia, Including Hemolytic Disease of Newborn	1.0511	0.882	
Medication Coumadin	1.043868	0.364	
Spina Bifida and Other Brain/Spinal/Nervous System Congenital Anomalies	1.039102	0.877	
Medication DOAC	1.03286	0.802	
Fibrosis of Lung and Other Lung Disorders	1.02544	0.754	
Medication for Hyperparathyroidism	1.021379	0.963	
Medication Chemotherapy	1.012046	0.873	
Medication Phosphate Binder	0.99251	0.970	
Medication Colchicine	0.986139	0.841	

Systemic Lupus Erythematosus and Other Autoimmune Disorders	0.983554	0.828	
Coagulation Defects and Other Specified Hematological Disorders	0.983165	0.815	
Thyroid Cancer, Melanoma, Neurofibromatosis, and Other Cancers and Tumors	0.978338	0.823	
Frequent Outpatient Visits in Last 1-3 Months Category (8,1e+03]	0.974163	0.682	
Non-Traumatic Coma, Brain Compression/Anoxic Damage	0.962989	0.841	
Hemiplegia/Hemiparesis	0.955542	0.702	
Protein-Calorie Malnutrition	0.954147	0.686	
Aspiration and Specified Bacterial Pneumonias and Other Severe Lung Infections	0.953691	0.667	
Septicemia, Sepsis, Systemic Inflammatory Response Syndrome/Shock	0.940973	0.467	
Combined and Other Severe Immunodeficiencies	0.886293	0.900	
Chronic Hepatitis	0.866152	0.291	
Age Category (35,40]	0.860383	0.217	
Congenital Metabolic Disorders, Not Elsewhere Classified	0.758346	0.328	
Age Category (30,35]	0.721685	0.008	**
Medication Anti-TNF	0.697307	0.394	
Central Nervous System Infections, Except Viral Meningitis	0.657062	0.054	
Age Category (20,25]	0.646884	0.001	***
Medication for ALS	0.594239	0.645	
Personality Disorders	0.586983	0.258	
Age Category(25,30]	0.507528	0.000	***
Ectopic and Molar Pregnancy, Except with Renal Failure, Shock, or Embolism	0.419348	0.391	
Cleft Lip/Cleft Palate	0.000261	0.922	
Thalassemia Major	9.31E-05	0.866	

Description of Model Details:

We used each variable's odds ratios (available in the supplementary data) as the ß coefficients for a logistic regression where the variables are regressed using generalized linear modeling onto the hospitalization outcome variable using the logit link. This methodology has been used prior, and is felt to perform as well as machine learning approaches[1]. Producing the mathematical formula including all 169 variables would be quite long, however the basic formula is shown

[1] Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, *110*, 12–22. https://doi.org/10.1016/j.jclinepi.2019.02.004

below with each beta coefficient represented by “ß” and each variable represented by “x.”

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m$$

All of the variables carry their own ß coefficients, and all are combined in the model to produce a score between 0 and 1.0, corresponding to a likelihood of hospitalization.