Assessing the Generalizability of a Clinical Machine Learning Model Across Multiple Emergency Departments

Alexander J. Ryu, MD; Santiago Romero-Brufau, MD, PhD; Ray Qian, MD; Heather A. Heaton, MD; David M. Nestler, MD; Shant Ayanian, MD; and Thomas C. Kingsley, MD

Abstract

Objective: To assess the generalizability of a clinical machine learning algorithm across multiple emergency departments (EDs).

Patients and Methods: We obtained data on all ED visits at our health care system's largest ED from May 5, 2018, to December 31, 2019. We also obtained data from 3 satellite EDs and 1 distant-hub ED from May 1, 2018, to December 31, 2018. A gradient-boosted machine model was trained on pooled data from the included EDs. To prevent the effect of differing training set sizes, the data were randomly downsampled to match those of our smallest ED. A second model was trained on this downsampled, pooled data. The model's performance was compared using area under the receiver operating characteristic (AUC). Finally, site-specific models were trained and tested across all the sites, and the importance of features was examined to understand the reasons for differing generalizability.

Results: The training data sets contained 1918-64,161 ED visits. The AUC for the pooled model ranged from 0.84 to 0.94 across the sites; the performance decreased slightly when Ns were downsampled to match those of our smallest ED site. When site-specific models were trained and tested across all the sites, the AUCs ranged more widely from 0.71 to 0.93. Within a single ED site, the performance of the 5 site-specific models was most variable for our largest and smallest EDs. Finally, when the importance of features was examined, several features were common to all site-specific models; however, the weight of these features differed.

Conclusion: A machine learning model for predicting hospital admission from the ED will generalize fairly well within the health care system but will still have significant differences in AUC performance across sites because of site-specific factors.

Many health care systems are becoming interested in operationalizing machine learning (ML) predictive models to guide resource allocation.[1,2] The use of ML to predict which emergency department (ED) patients may ultimately require hospital admission has the potential to reduce ED overcrowding and improve hospital efficiency.[3,4] Existing literature has reported that regression analysis[5,6] and ML approaches[7-9] can be suitable for this task, achieving excellent performance. Specifically, regression models, tree-based models, and neural network models performed well when structured data inputs were used, whereas neural networks displayed better performance when unstructured features, such as clinical notes, were included.[10]

Modern health care systems often include a network of hospitals, with lower-capacity community or critical-access hospitals serving more remote geographies while referring patients with complications to the network's larger secondary or tertiary care centers.[11] With these structures becoming increasingly common, it will be important to understand how an algorithm for predicting admissions generalizes across different sites in the health

From the Division of Hospital Internal Medicine (A.J.R., S.A., T.C.K.), Department of Medicine (S.R.-B.), Department of Laboratory Medicine and Pathology (R.Q.), and Department of Emergency Medicine (H.A.H., D.M.N.), Mayo Clinic, Rochester, MN.

care system, particularly given the context of mounting concerns regarding the generalizability of health care ML algorithms.[12] A better understanding of this issue will likely help guide model training and implementation for many health care systems looking to implement these types of algorithms.

In particular, we hypothesized that issues with model generalizability could arise from site-specific or model-specific factors. Site-specific factors contributing to model generalizability could include differing patient characteristics or ED practice patterns, which may also be related to specific capabilities of a given ED. Model-specific factors may include random errors or differences in model performance relating to the volume of training data available, which may be of particular concern for small-volume, rural EDs.

Here, we assessed the generalizability of a model that predicts the likelihood of hospital admission of ED patients. We elected to use a gradient-boosted machine (GBM) model for this task, given that we anticipated using structured data features for the ease of model implementation in practice. Neural network models were considered, but we suspected that performance gains would be minimal over GBM, given the lack of unstructured features, and there would be an increase in the complexity of model training and tuning. We hypothesized that a model training strategy that uses pooled data across ED sites with downsampling training data set sizes would lead to optimal model performance across the sites. Additionally, we suspected that a model trained at a referral center might generalize poorly to a small, rural ED and vice versa. Finally, we hypothesized that there would be some common clinical features that are found to be important across all predictive models.

PATIENTS AND METHODS

Data Collection

We obtained data on all ED patients presenting to our health system's largest ED (Rochester, Minnesota) from May 5, 2018, to December 31, 2019. We also obtained data on all ED patients from 3 regional satellite EDs of various sizes (Eau Claire, Wisconsin; Austin, Minnesota; and New Prague, Minnesota) and 1 distant-hub site (Phoenix, Arizona) from May 1, 2018, to December 31, 2018. Our study was deemed exempt and granted a waiver of consent by the Mayo Clinic Institutional Review Board (#21-006291). The data were stored in our organization's electronic health record data warehouse and extracted using SAP Web Intelligence. Data on the hospital's surrounding communities were collected from public US Census Bureau reports.

Feature Selection

We captured and engineered a total of 47 model features pertaining to real-time clinical information about the patients (age and initial vital signs), the chief symptom, any ED protocols activated, and the mode and timing of arrival at the ED. Chief symptom labels were coded by our ED and, thus, did not have issues with accommodating free text. Emergency department protocol activation refers to our ED's use of several standardized documentation protocols for certain high-acuity situations such as ST-elevation myocardial infarction or trauma. We prioritized using features that would be available early in a patient's ED course, likely within the first 15 minutes, to ensure that our predictions are maximally forward looking. Of note, all EDs included in our study shared a high degree of data harmonization because of pre-existing data infrastructure work. However, local practices—and, thus, patterns in the data—may differ from site to site. All features included in the model, along with data type, are listed in the Supplemental Appendix (available online at http://www.mcpiqojournal.org).

Pooled Model

We then used the XGBoost package with a random 70/15/15 train/validation/test split of all the sites' data combined to train a model and subsequently test it against a 15% random test sample of each site's data. The area under the receiver operating characteristic (AUC) and its SD, using jackknife estimation, were calculated for each model test. For the jackknife estimation of SD, we calculated the AUC for each of n test sets, where n corresponded to the size of a given test set, with each test set consisting of n−1 observations, with each observation being excluded once. We then calculated the SD of this AUC distribution. Analysis of variance (ANOVA) was

TABLE 1. Characteristics of the Different Emergency Department Sites[a]

Characteristic	Rochester, Minnesota	Phoenix, Arizona	Eau Claire, Wisconsin	Austin, Minnesota	New Prague, Minnesota
ED visits/y	78,000	44,000	34,000	18,000	7,000
ED capabilities	76 beds, level 1 trauma center, inpatient psychiatry, ED observation unit	27 beds, level 4 trauma capabilities	27 beds, level 2 trauma center	17 beds, level 4 trauma center	4 beds, level 4 trauma center, critical-access hospital
City population	115,557	1,680,992	68,187	25,114	7,899
City % White	82%	66%	91%	93%	97%
City median household income (2019)	$73,106	$57,459	$55,477	$48,127	$77,949

[a]ED, emergency department.

performed to assess any significance of differences between the sites.

Pooled Model With Downsampled N

Because the number of visits sampled at each site differed by up to 30 times, we sought to remove this factor as a potential contributor to differing model performance by site. Therefore, each site's data were randomly sampled down to the size of our smallest ED site. The train/validation/test split was repeated, and a second model was trained on this downsampled, pooled data and again tested across each site. Furthermore, ANOVA was repeated.

Site-Specific Models With Cross-Site Testing

To further understand the potential reasons for the differing model performance, we then trained a model on each ED's downsampled data and tested each of those models across the sites to understand whether admission decisions were consistently easier or more difficult to predict at certain sites. Furthermore, ANOVA was performed for each of the 5 site-specific models to compare their AUC results across the sites.

Importance of Features

Finally, we examined the top 10 features that the model deemed most important across the 5 site-specific models and the pooled model to further understand how the model was making predictions at each site.

RESULTS

The characteristics of the 5 EDs studied are summarized in Table 1, along with basic demographic information about their surrounding communities. Specifically, in the US component of our health care system, there are 3 major EDs, in Rochester, Minnesota; Phoenix, Arizona; and Jacksonville, Florida, with the Rochester site being the largest. Surrounding the Rochester site is a network of smaller hospitals ranging from secondary referral sites to critical-access hospitals. As described in Table 1, we attempted to capture 1 ED of each of our institutional size designations.

Table 2 presents the characteristics of patient samples from each site. For context, we included information about the fraction of patients admitted, ED visit duration, and emergency severity index (ESI; lower numbers indicate greater patient acuity) as well as the number of visits used for the training, validation, and testing of the model per site.

Table 3 presents the results of AUC testing across the sites of models trained on pooled data before and after the downsampling of data set Ns. The downsampled models matched each site's data to the same N as that of New Prague, Minnesota, before pooling the data. The AUC performance decreased slightly, and its SD increased at each site in which the downsampling of the data occurred. The same was true when the models on the pooled samples were tested. The differences in AUC by site were found to be significant, with $P<.01$, determined using ANOVA. New Prague, Minnesota, was not included in ANOVA for the second model because its AUC was identical to that of Austin, Minnesota.

TABLE 2. Patient Sample Characteristics by Emergency Department Site[a]

Characteristic	Rochester, Minnesota	Phoenix, Arizona	Eau Claire, Wisconsin	Austin, Minnesota	New Prague, Minnesota
Training N	64,161	18,233	12,506	6,558	1,919
Validation N	12,322	3,907	2,680	1,405	411
Test N	12,322	3,907	2,680	1,405	411
Patients admitted (%)	35%	49%	27%	18%	21%
Median ED visit duration (h)	4.4	3.7	3.3	2.8	3.0
Patients ESI 1-3 (%)	80%	88%	82%	67%	75%
Mean patient age (SD)	55.9±20.7	58.6±19.7	52.4±21.4	52.5±22.3	54.0±21.8

[a]ED, emergency department; ESI, emergency severity index; SD, standard deviation.

Table 4 presents the results of the testing of site-specific models, each trained using training sets of downsampled Ns, across all the sites and pooled test data. All AUCs were significantly different from each other. When the models were tested at the same site at which they were trained, the performance was highest at Eau Claire, Wisconsin (0.93), and lowest at New Prague, Minnesota (0.79), suggesting that the features included in our model were particularly well suited to patients and clinical practices in Eau Claire, Wisconsin, and less so in New Prague, Minnesota. For 4 of the 5 site-specific models, the AUC performance was the best at Eau Claire, Wisconsin. Interestingly, the AUC was higher in Eau Claire, Wisconsin, than that at "home" sites where the models were trained. The lowest AUC was noted for a model trained at our tertiary referral center, Rochester, Minnesota, and was tested at our critical-access hospital, New Prague, Minnesota. A model trained on data from New Prague, Minnesota, similarly performed poorly when tested on data from Rochester, Minnesota, suggesting that the

patients and practices at these sites were most dissimilar.

Table 5 presents the top 10 important features for each model, along with the associated feature weight. Specifically, we reported the average gain across all splits in which a feature was used. There appeared to be substantial overlap in features that the model determined to be of high importance, but the weights of these features varied substantially across the sites. The ESI score, arrival by an ambulance, patient's age, and having had an electrocardiogram were among the top 10 features at all the sites. In the pooled model, having been a patient in Eau Claire, Wisconsin, was the only site-specific indicator (there was 1 indicator feature per site) among the top 10 features.

DISCUSSION

This article presents a strategy for assessing the generalizability of a clinical artificial intelligence model across multiple ED sites. Although previous literature has reported the suitability of GBMs for similar clinical prediction tasks, few, if any, studies have proposed

TABLE 3. Performance of Pooled Model Before and After N Downsampling[a]

Test sites listed to the right	Rochester, Minnesota	Phoenix, Arizona	Eau Claire, Wisconsin	Austin, Minnesota	New Prague, Minnesota	All sites combined
Pooled, not downsampled AUC ± SD	0.89±.00002[b]	0.84±.0001[b]	0.94±.0001[b]	0.92±.0003[b]	0.86±.001[b]	0.88±.00002
Pooled, downsampled AUC ± SD	0.87±.0007[c]	0.83±.0008[c]	0.92±.0006[c]	0.86±.0008[c]	0.86±.001	0.87±.0001

[a]AUC, area under the receiver-operator characteristic; SD, standard deviation.
[b]Indicates P<.001 by analysis of variance when tested across the means with matching symbols.
[c]Indicates P<.001 by analysis of variance when tested across the means with matching symbols.

TABLE 4. Performance of Site-specific Models With Downsampled Ns[a]

Training sites listed below; test sites listed to the right	Rochester, Minnesota	Phoenix, Arizona	Eau Claire, Wisconsin	Austin, Minnesota	New Prague, Minnesota	All sites combined
Rochester, Minnesota AUC ± SD	0.85±.0009[b]	0.78±.001[b]	0.89±.0007[b]	0.84±.0009[b]	0.71±.001[b]	0.84±.0002
Phoenix, Arizona AUC ± SD	0.77±.0009[c]	0.81±.0008[c]	0.90±.0007[c]	0.85±.0008[c]	0.81±.001	0.82±.0002
Eau Claire, Wisconsin AUC ± SD	0.79±.0009[d]	0.81±.0009[d]	0.93±.0005[d]	0.88±.0007[d]	0.75±.001[d]	0.84±.0002
Austin, Minnesota AUC ± SD	0.75±.0009[e]	0.79±.0009[e]	0.91±.0006[e]	0.88±.0007[e]	0.79±.001	0.83±.0002
New Prague, Minnesota AUC ± SD	0.74±.0009[f]	0.8±.0009[f]	0.86±.0007[f]	0.88±.0008[f]	0.82±.001[f]	0.83±.0002

[a]AUC, area under the receiver-operator characteristic; SD, standard deviation.
[b]Indicates $P<0.01$ by analysis of variance when tested across the means with matching symbols.
[c]Indicates $P<0.01$ by analysis of variance when tested across the means with matching symbols.
[d]Indicates $P<0.01$ by analysis of variance when tested across the means with matching symbols.
[e]Indicates $P<0.01$ by analysis of variance when tested across the means with matching symbols.
[f]Indicates $P<0.01$ by analysis of variance when tested across the means with matching symbols.

methods for investigating the generalizability of these models across multiple health care sites.[7-9,13] We used GBMs to train a model predicting the likelihood of a patient's admission from the ED early in their ED course. We trained a model on pooled data from 5 diverse ED sites and tested the model's performance at each of the sites, noting high, but different, AUC performance across the sites. We then downsampled the data set Ns across the sites to correct for any effect of site overrepresentation on differing AUCs and found small but significant decreases in the model's performance. This investigation revealed significant differences in the AUC performance when examined using ANOVA, even after downsampling the data set Ns; the AUCs for our downsampled, pooled model ranged from 0.83 to 0.92 when tested across the sites, which is likely attributable to a combination of patient-specific and clinical practice (ie, sitespecific) factors. We also found that after the downsampling of Ns, the AUC performance decreased by 0.01-0.06 per site. Importantly, the decrease in the model performance by site was not proportional to the degree of downsampling that occurred. Additionally, the pooled, downsampled model performed better than site-specific models at all the sites when compared with how these performed at their home sites, with the exception of Eau Claire, Wisconsin. This suggests that

patient data from the other sites aided the pooled model in detecting patterns that were generalizable, thus raising the performance for all the sites. Overall, this suggests that training a model on site-pooled data, over as long a time period as possible, will lead to the best model performance; however, even with significant downsampling, from pooled training N=103,377 to downsampled N=9595, the performance decreases only modestly.

We then trained site-specific models using the downsampled data sets for each of our 5 ED sites and tested those across all 5 ED sites. This exercise suggested that at 1 site in particular (Eau Claire, Wisconsin), admission decisions were easier to predict than at any other site, which is likely because of a combination of patient factors and practice patterns, although the reasons for this were not explicitly captured in our study. Additionally, this suggested that a model trained at a tertiary referral ED would perform particularly poorly at a critical-access ED and vice versa. This seems like an expected finding, given the differences in patient complexity and ED capabilities.

Upon examining the top 10 important features at each site, we noted that there was significant overlap of all the models, including the ESI score, arrival by an ambulance, patient age, and having had an electrocardiogram. This suggests that these factors are indeed

TABLE 5. Top 10 Important Features for Each Model[a]

Data aggregation	Rochester, Minnesota	Phoenix, Arizona	Eau Claire, Wisconsin	Austin, Minnesota	New Prague, Minnesota	Pooled model
Feature 1 (weight)	ESI321 (171)	ESI (27)	Weight (108)	Had EKG? (56)	Had EKG? (87)	Had EKG? (141)
Feature 2 (weight)	Ambulance (115)	Had EKG? (13)	Had EKG? (53)	ESI (26)	Ambulance (35)	ESI321 (56)
Feature 3 (weight)	Weight (65)	Fever (12)	Wheelchair (25)	Oxygen saturation (15)	ESI (33)	ESI (52)
Feature 4 (weight)	ESI (25)	Altered mental status (10)	ESI (25)	Age (14)	Wheelchair (25)	Weight (43)
Feature 5 (weight)	Had EKG? (19)	Wheelchair (7)	Ambulance (23)	Suicidal (11)	Oxygen saturation (22)	Ambulance (38)
Feature 6 (weight)	Wheelchair (17)	Age (7)	Suicidal (22)	Weight (11)	Temperature (21)	Wheelchair (25)
Feature 7 (weight)	Chest pain (17)	Chest pain (5)	Age (20)	Ambulance (10)	Age (16)	Eau Claire (23)
Feature 8 (weight)	From outside hospital (14)	Ambulance (5)	From outside hospital (18)	Pulse (8)	Weight (14)	Abdominal pain (20)
Feature 9 (weight)	Age (13)	Temperature (5)	Respiratory rate (16)	Respiratory rate (7)	Diastolic blood pressure (13)	Age (20)
Feature 10 (weight)	Resuscitation status (12)	Weakness (5)	Diastolic blood pressure (10)	Abdominal pain (6)	Respiratory rate (13)	Chest pain (18)

aEKG, electrocardiogram; ESI, emergency severity index.

highly relevant clinical features for our prediction, rather than site-specific outliers. The weight of the top 10 features, however, differed significantly between the site-specific models.

Comparing our results with those of the existing literature, our model's AUC results fit into the higher end of reported AUCs, supporting the suitability of GBMs for this task.[7-9,13] There also appear to be limited studies that examined model generalizability, particularly using ML models. One study examined the generalizability of a logistic regression model across disparate hospitals and retrained the model for each hospital.[14] Other similar studies focused on a single ED,[5,6,13,15-19] used curated survey data,[10,20] or grouped multiple EDs together for model training.[8]

Our study was limited in that only 1 health care system's EDs were represented, although the EDs were substantially variable in size and geography. All the EDs included shared a high degree of data harmonization, which facilitated model testing. Finally, our algorithm generated predictions at only 1 time point in the patients' ED course, thereby neglecting situations in which major changes occurring in a patient's ED course significantly alter their probability of admission.

With respect to model deployment, our health care system's largest ED is currently piloting an initiative in which hospital medicine physicians collaborate with emergency medicine physicians to help facilitate patient triage. The leadership of this initiative is interested in using this model to more quickly identify patients for interventions to expedite both admission and discharge. For other health care systems looking to implement similar models, our data appear to suggest that the best strategy for achieving high performance across multiple sites is to train a model on pooled data, without downsampling, that contain features identifying which patients came from which site, as was done in our study. This approach contributes to not only a high-performing model but also the ease of maintenance, with only 1 model to oversee, to which additional training data can be easily added over time.

CONCLUSION

Overall, this study provides a strategy for systematically assessing the generalizability of a

clinical ML model across multiple ED sites and provides the estimates of AUC differences across multiple scenarios. We determined that optimal GBM model performance is achieved when trained on multisite, pooled data; however, even with this strategy, the model will perform differently at some sites, which was not explained by the random errors or differences in the training set N. Instead, the differences in the model performance were likely due to site-specific factors. When the importance of the features was examined for our GBM models, the ESI score, arrival by an ambulance, patient's age, and having had an electrocardiogram appeared to be important clinical predictors across all the sites, although the importance of the features differed.

ACKNOWLEDGMENTS

SUPPLEMENTAL ONLINE MATERIAL

Supplemental material can be found online at http://www.mcpiqojournal.org. Supplemental material attached to journal articles has not been edited, and the authors take responsibility for the accuracy of all data.

Abbreviations and Acronyms: **ANOVA,** analysis of variance; **AUC,** area under the receiver-operator characteristic; **ED,** emergency department; **ESI,** emergency severity index; **GBM,** gradient-boosted machine; **ML,** machine learning

Correspondence: Address to Alexander J. Ryu, MD, 1216 2nd St SW, Old Marion Hall, 4th Floor, Rochester, MN 55902 (Ryu.alexander@mayo.edu).

ORCID
Santiago Romero-Brufau: [ID] https://orcid.org/0000-0002-9922-0083; Shant Ayanian: [ID] https://orcid.org/0000-0001-9319-9001

REFERENCES

1. The Johns Hopkins Hospital launches capacity command center to enhance hospital operations. The Johns Hopkins Hospital. Accessed November 16, 2020. https://www.hopkinsmedicine.org/news/media/releases/the_johns_hopkins_hospital_launches_capacity_command_center_to_enhance_hospital_operations
2. Improving patient-care with hospital command centers. Medtech impact on wellness. Medtech Impact. Accessed November 16, 2020. https://www.medtechimpact.com/improving-patient-care-with-hospital-command-centers/
3. Hoot NR, Aronsky D. Systematic review of emergency department crowding: causes, effects, and solutions. Ann Emerg Med. 2008;52(2):126-136.
4. Asplin BR, Magid DJ, Rhodes KV, Solberg LI, Lurie N, Camargo CA Jr. A conceptual model of emergency department crowding. Ann Emerg Med. 2003;42(2):173-180.
5. Parker CA, Liu N, Wu SX, Shen Y, Lam SS, Ong ME. Predicting hospital admission at the emergency department triage: a novel prediction model. Am J Emerg Med. 2019;37(8):1498-1504.
6. Lucke JA, de Gelder J, Clarijs F, et al. Early prediction of hospital admission for emergency department patients: a comparison between patients younger or older than 70 years. Emerg Med J. 2018;35(1):18-27.
7. Lucini FR, S. Fogliatto FS, da Silveira GJ, et al. Text mining approach to predict hospital admissions using early medical records from the emergency department. Int J Med Inform. 2017; 100:1-8.
8. Graham B, Bond R, Quinn M, Mulvenna M. Using data mining to predict hospital admissions from the emergency department. IEEE Access. 2018;6:10458-10469.
9. Roquette BP, Nagano H, Marujo EC, Maiorano AC. Prediction of admission in pediatric emergency department with deep neural networks and triage textual data. Neur Netw. 2020; 126:170-177.
10. Zhang X, Kim J, Patzer RE, Pitts SR, Patzer A, Schrager JD. Prediction of emergency department hospital admission based on natural language processing and neural networks. Methods Inf Med. 2017;56(5):377-389.
11. Hospital consolidation: trends, impacts & outlook. National Institute for Health Care Management. Accessed June 12, 2021. https://nihcm.org/publications/hospital-consolidation-trends-impacts-outlook
12. Kakarmath S, Esteva A, Arnaout R, et al. Best practices for authors of healthcare-related artificial intelligence manuscripts. NPJ Digit Med. 2020;3(1):1-3.
13. Leegon J, Jones I, Lanaghan K, Aronsky D. Predicting hospital admission in a pediatric emergency department using an artificial neural network. AMIA Annu Symp Proc. 2006;2006:1004.
14. Peck JS, Gaehde SA, Nightingale DJ, et al. Generalizability of a simple approach for predicting hospital admission from an emergency department. Acad Emerg Med. 2013;20(11):1156-1163.
15. Hoot NR, LeBlanc LJ, Jones I, et al. Forecasting emergency department crowding: a prospective, real-time evaluation. J Am Med Inform Assoc. 2009;16(3):338-345.
16. Golmohammadi D. Predicting hospital admissions to reduce emergency department boarding. Int J Prod Econ. 2016;182: 535-544.
17. Peck JS, Benneyan JC, Nightingale DJ, Gaehde SA. Predicting emergency department inpatient admissions to improve same-day patient flow. Acad Emerg Med. 2012;19(9):E1045-E1054.
18. Sun Y, Heng BH, Tay SY, Seow E. Predicting hospital admissions at emergency department triage using routine administrative data. Acad Emerg Med. 2011;18(8):844-850.
19. Kim SW, Li JY, Hakendorf P, Teubner DJ, Ben-Tovim DI, Thompson CH. Predicting admission of patients by their presentation to the emergency department. Emerg Med Australas. 2014;26(4):361-367.
20. Kocher KE, Meurer WJ, Desmond JS, Nallamothu BK. Effect of testing and treatment on emergency department length of stay using a national database. Acad Emerg Med. 2012;19(5):525-534.