OPEN

External Validation and Comparison of a General Ward Deterioration Index Between Diversely Different Health Systems

Brandon C. Cummings, MHI[1,2]

Joseph M. Blackmer, BS[1,2]

Jonathan R. Motyka, MS[1,2]

Negar Farzaneh, PhD[1,2]

Loc Cao, MS[1,2]

Erin L. Bisco, BA[1,2]

James D. Glassbrook, AAS RT[3]

Michael D. Roebuck, MD[2,4]

Christopher E. Gillies, PhD[1,2]

Andrew J. Admon, MD, MPH, MS[1,5,6]

Richard P. Medlin Jr, MD, MSIS[1,2]

Karandeep Singh, MD, MMS[5,7,8]

Michael W. Sjoding, MD[1,5,8]

Kevin R. Ward, MD[1,2,9]

Sardar Ansari, PhD[1,2]

OBJECTIVES: Implementing a predictive analytic model in a new clinical environment is fraught with challenges. Dataset shifts such as differences in clinical practice, new data acquisition devices, or changes in the electronic health record (EHR) implementation mean that the input data seen by a model can differ significantly from the data it was trained on. Validating models at multiple institutions is therefore critical. Here, using retrospective data, we demonstrate how Predicting Intensive Care Transfers and other UnfoReseen Events (PICTURE), a deterioration index developed at a single academic medical center, generalizes to a second institution with significantly different patient population.

DESIGN: PICTURE is a deterioration index designed for the general ward, which uses structured EHR data such as laboratory values and vital signs.

SETTING: The general wards of two large hospitals, one an academic medical center and the other a community hospital.

SUBJECTS: The model has previously been trained and validated on a cohort of 165,018 general ward encounters from a large academic medical center. Here, we apply this model to 11,083 encounters from a separate community hospital.

INTERVENTIONS: None.

MEASUREMENTS AND MAIN RESULTS: The hospitals were found to have significant differences in missingness rates (> 5% difference in 9/52 features), deterioration rate (4.5% vs 2.5%), and racial makeup (20% non-White vs 49% non-White). Despite these differences, PICTURE's performance was consistent (area under the receiver operating characteristic curve [AUROC], 0.870; 95% CI, 0.861–0.878), area under the precision-recall curve (AUPRC, 0.298; 95% CI, 0.275–0.320) at the first hospital; AUROC 0.875 (0.851–0.902), AUPRC 0.339 (0.281–0.398) at the second. AUPRC was standardized to a 2.5% event rate. PICTURE also outperformed both the Epic Deterioration Index and the National Early Warning Score at both institutions.

CONCLUSIONS: Important differences were observed between the two institutions, including data availability and demographic makeup. PICTURE was able to identify general ward patients at risk of deterioration at both hospitals with consistent performance (AUROC and AUPRC) and compared favorably to existing metrics.

KEY WORDS: clinical decision support; dataset shift; deterioration index; model generalization; predictive analytic

As predictive artificial intelligence models and Early Warning Systems (EWSs) are increasingly developed and deployed in the clinical environment (1–14), there is increasing interest in the "last-mile" challenges that delay implementation and often render these EWSs ineffective, or in the worst case, counterproductive when used clinically (15–17). One such problem is termed dataset shift (18). This occurs when the real-time data seen

DOI: 10.1097/CCM.0000000000005837

KEY POINTS

Question: Do the predictions made by Predicting Intensive Care Transfers and other UnfoReseen Events (PICTURE), a Clinical Decision Support tool developed at a single academic medical center, generalize to a second community-oriented hospital?

Findings: In this retrospective analysis, PICTURE's performance did not drop significantly (α = 5%) when moving to a second hospital, despite important clinical and information-technology differences between the institutions.

Meaning: Implementing a predictive analytic model in a new clinical environment is fraught with challenges, due in part to the phenomenon of dataset shift. PICTURE was able to overcome these differences and performed well at an outside institution.

by a model differs from that on which it was trained (19). In addition to intra-institution challenges such as changing reimbursement models, updated data acquisition devices, and new information technology practices that can cause changes in underlying data, differences between hospital systems can also change the ways in which a given predictive model behaves (18, 19). For example, different patient populations, different practice guidelines, and different equipment could all affect the types and quality of data being collected—among many other factors. Accounting for these differences are a necessary challenge whenever a model is implemented in a new hospital. As such, external validations is necessary to ensure these models remain useful outside their home institutions (20, 21).

To illustrate these differences, we validated our previously described predictive model, Predicting Intensive Care Transfers and other UnfoReseen Events (PICTURE), using a large retrospective dataset from an outside institution (22). PICTURE is a predictive model using structured data such as laboratory values and vital signs from the electronic health record (EHR) to identify patients at risk of deterioration such as death, ICU transfer, or mechanical ventilation. We compare against two commonly used deterioration metrics, the Epic Deterioration Index (EDI) and National Early Warning Score (NEWS), both in terms of predictive performance as well as prediction lead time (see **Supplement 1**, http://links.lww.com/CCM/H305, for descriptions of EDI and NEWS). However, one key limitation of our initial study is that the model was developed and tested at a single academic medical center. While we made several important choices when designing the model in an attempt to prioritize generalizability (a novel imputation mechanism to mask patterns in missingness which can change between institutions [23] and excluding variables such as medications which reflect clinician behavior [22]), these design choices were backed mostly by simulations.

Our single center, Michigan Medicine (MM), is a large academic research hospital with a level 1 trauma center and advanced transplant and cardiac care facilities, which has key differences with the community hospitals that account for most of the U.S. population's healthcare (24). To this end, we externally validated PICTURE at Hurley Medical Center (HMC) in Flint, MI. In contrast to MM's status as an academic referral hub, HMC is a large 443-bed community hospital. Although both hospitals use the same EHR vendor (Epic Systems, Verona, WI), differences in both patient care patterns and structural/informatics organization coalesce into changes in feature distribution, deterioration rate, and other factors affecting the model. Here, we quantify these differences to understand the degree of dataset shift between the two hospitals and investigate the similarities and differences in how the model performs considering these changes.

METHODS

Study procedures were followed in accordance with the Helsinki Declaration of 1975 and were approved with a waiver of informed consent by Institutional Review Boards (IRBs) of both institutions, with MM acting as the IRB of record. At MM, the study was initially approved on September 4, 2014, under HUM00092309: "Development of Clinical Decision Support Tools in Acute Care." At HMC, it was approved January 5, 2021, as 1686064: "External Validation of PICTURE-Suite Model Performance at Hurley Medical Center" (see **Supplement 2**, http://links.lww.com/CCM/H305, for details). PICTURE was developed using XGBoost v0.90 (https://xgboost.readthedocs.io/en/release_0.90/index.html) using a logistic objective function with a maximum depth of three layers, a learning rate of 0.005, and early stopping via area under the precision-recall curve (AUPRC) after 30 rounds. It was initially evaluated on a composite target of death, ICU transfer, mechanical ventilation, and cardiac arrest within 24 hours on a cohort of adult inpatients admitted to the general wards. Missing value imputation was done with a combination of forward fill and using the mean of the posterior distribution from a multivariate Bayesian regression model as described in Gillies et al (23). Further details on model construction and internal validation, data preprocessing, and prediction explanations have been previously reported and are also available in **Supplement 3–5** (http://links.lww.com/CCM/H305) (22). A full list of input features is available in **Supplement 6** (http://links.lww.com/CCM/H305). Although this study uses retrospective data, PICTURE is currently implemented in real-time at MM, with prospective data currently being collected to assess performance.

PICTURE's performance was evaluated on data consisting of patients between the ages of 18 and 89 whose status was inpatient or other observation status, and whose first ICU transfer, if present, was from a general ward. PICTURE, as well the EDI and NEWS, was first assessed using the area under the receiver operating characteristic curve (AUROC) and AUPRC. The latter metric reflects the balance between sensitivity and positive predictive value (PPV). These are calculated using two levels of granularity: observation level and encounter level. On the observation level, we consider any prediction in which the patient deteriorated within 24 hours of the observation to be positive. On the encounter level, we compare the maximum score over the patient's entire length of stay to the ultimate outcome of that encounter. Two additional analyses were applied to further investigate the behavior of the three models. First, lead time, which refers to the amount of advanced warning that clinicians receive between the score generation and the patient's deterioration, was assessed by censoring the data at increasing intervals leading up to the deterioration event. Second, we demonstrate the variability in threshold selection between institutions by quantifying the change in threshold performance both with and without calibration.

RESULTS

Cohort Details

Our initial data pull contained data from patients who had been admitted between January 1, 2015, and November 23, 2020. Since data from before December 31, 2018, had been previously used to train the model, these patients were excluded from the MM portion of the analysis. After selecting patients meeting our inclusion criteria, there remained 59,863 encounters from MM and 36,947 encounters from HMC. To facilitate comparison with the EDI, the PICTURE and NEWS scores were aligned from their native frequency (updated each time new data are resulted) to the existing EDI scores at both institutions (see **Supplement 7**, http://links.lww.com/CCM/H305, and [22] for details). This effectively resampled them to an update frequency of 15 minutes at MM and 20 minutes at HMC, and reduced our cohort to 44,202 encounters at MM and 11,083 at HMC that had overlapping PICTURE and EDI scores. PICTURE scores were aligned to the EDI, rather than vice versa, to give the EDI any potential benefit from the alignment procedure. PICTURE's performance on the full, nonaligned cohort can be found in **Supplement 8** (http://links.lww.com/CCM/H305), and performance on each component of the target label (e.g., ICU transfer, death) in **Supplement 9** (http://links.lww.com/CCM/H305). **Table 1** below presents the demographic makeup of our cohort across both institutions.

Characterization of Institutional Differences

Differences in Input Variables. One method of characterizing the differences between institutions is by comparing the distributions of the input features. **Figure 1** displays the standardized mean differences between MM and HMC for each of the numeric features. With the exception of one variable (mean platelet volume, MPV, Fig. 1*C*), the mean of all differences was within a single SD. There were also large differences in the missingness rates between institutions for individual variables, reflecting differences in care patterns between the two hospitals.

Differences in Patient Transfer Patterns. An additional key difference between the two institutions was the source of patients—while in both cases, the Emergency Department (ED) represented the most common ways for patients to enter the hospital system, this proportion was much lower at MM (66.0%) than HMC (91.7%). We believe this is due to two factors: first, there are multiple other level 1 trauma centers within close proximity to MM. Second, MM sees a much higher proportion of patients come through surgical pathways such as

TABLE 1.
Cohort Demographics

Dataset	Michigan Medicine	Hurley Medical Center	*p*
Encounters, *n*	44,202	11,083	NA
Patients, *n*	30,374	8,010	NA
Date range (date of admission)	January 1, 2019, to November 23, 2020	January 1, 2015, to November 23, 2020	NA
Age (yr), median (interquartile range)	61.5 (47.6–71.8)	58.2 (44.2–70.3)	< 0.001
Race, *n* (%)			
Asian	855 (2.0)	18 (0.2)	< 0.001
Black	5,575 (12.6)	5,044 (45.5)	< 0.001
White	35,480 (80.3)	5,641 (50.9)	< 0.001
Other	2,292 (5.2)	380 (3.4)	< 0.001
Ethnicity, *n* (%)			
Hispanic/Latino	1,196 (2.7)	258 (2.3)	0.026
Non-Hispanic/Latino	42,210 (95.5)	10,776 (97.2)	< 0.001
Other	652 (1.5)	49 (0.4)	< 0.001
Female sex, *n* (%)	21,592 (48.8)	6,000 (54.1)	< 0.001
Event rate, *n* (%)	1,998 (4.5)	278 (2.5)	< 0.001
Death	360 (0.8)	69 (0.6)	0.040
ICU transfer	1,610 (3.6)	200 (1.8)	< 0.001
Mechanical ventilation	28 (0.1)	9 (0.1)	0.516

NA = not available.

Cohort size, sex, racial and ethnic makeup, and target prevalence. Statistics are calculated per-encounter. For the event rate percentages, in the case that multiple outcome criteria were met, only the first event was counted. *p* values are calculated across the two datasets using a Mann-Whitney *U* test for continuous variables (age) and a χ^2 test for categorical variables.

operating rooms (ORs) and catheterization laboratory values (19.6%) than HMC (5.9%). To further visualize these workflow differences, directed network graphs were constructed based on the Admission, Discharge, and Transfer (ADT) table for each encounter (**Supplement 10**, http://links.lww.com/CCM/H305) (25).

Differences in Demographic Makeup. The demographic makeup of the two institutions are very distinct in terms of both race and biological sex. A subset analysis was performed to ensure we perform equally well across groups. **Table 2** presents the results of these subsets. The primary institution (MM) had a noticeably higher event rate in non-White patients (4.8–5.5% compared with 4.4% in White patients). This effect was less noticeable at the second institution (HMC, 2.6 vs 2.4%). Despite these institutional differences, the model performed similarly at both hospitals. The model performed similarly well across ethnicities at the first institution, but there were too few deteriorations (4/258) in Hispanic/Latino patients at the second institution to provide a reliable estimate of model performance in this population. At MM, all three models performed better in females than in males. This is also true at HMC, but there is overlap between the CI estimations, possibly due to the smaller sample size. In both cases, the proportion of males who deteriorated is greater than that of females.

Comparison of PICTURE Versus EDI and NEWS

The performance of the three predictive tools is presented in **Table 3** below. PICTURE outperformed both models at both institutions, even when resampled to the EDI's native frequency. On the observation level, PICTURE's AUROC and adjusted AUPRC increased slightly at HMC when compared with MM. The encounter level performance metrics fell well within the CIs for both AUROC and AUPRC.

Prediction Lead Time Simulation

Lead time refers to the amount of advanced warning clinicians receive between the score generation and the patient's deterioration. It is an important component of a predictive model's utility in a clinical environment. Here, it was assessed in a threshold-independent manner by comparing the classification performance of both PICTURE and the EDI while excluding varying intervals extending out before the deterioration event. PICTURE's performance was higher than the EDI at all timepoints across both institutions, although the larger CIs at the second institution shared some overlap. PICTURE's AUROC remained above 0.8 at both hospitals even when limited to predictions 24 hours in advance of the deterioration. **Figure 2** displays the performance of PICTURE and the EDI at each interval.

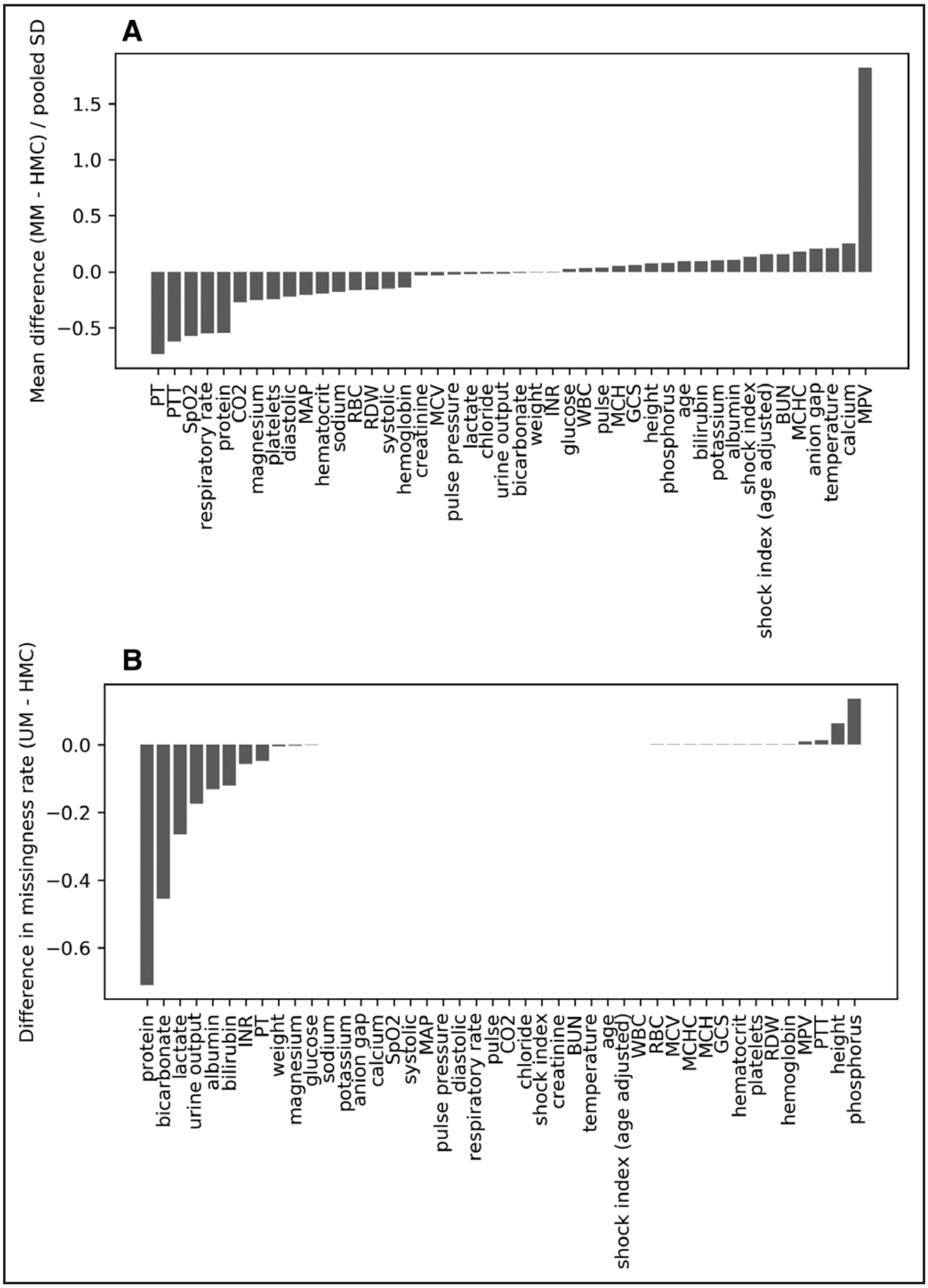

Figure 1. Standardized mean difference (Cohen's *d*) between the two institutions at encounter level. **A**, For each feature, the median value of each hospital encounter was first calculated to avoid biasing the calculations toward patients with more frequently drawn laboratory values, which may indicate sicker patients. The mean difference of these encounter level statistics was taken (Michigan Medicine [MM]–Hurley Medical Center [HMC]) and normalized to the pooled SD. **B**, Difference in encounter level missing rate between institutions (MM–HMC).

Calibration and Simulated Alert Thresholds

To demonstrate the effect of dataset shift on the selection of an alerting threshold, two simulations were constructed: one in which the same threshold is used on both institutions, and a second using recalibrated values. **Figure 3** displays PPV, sensitivity, and specificity across varying thresholds at both institutions. In a live clinical environment, the threshold may be selected based on provider preference and feedback which likely vary between clinical settings. Work-up to detection ratio (WDR) has been suggested as a possible metric for threshold selection, defined as the number of alerts required for each true positive (the inverse of PPV) (26).

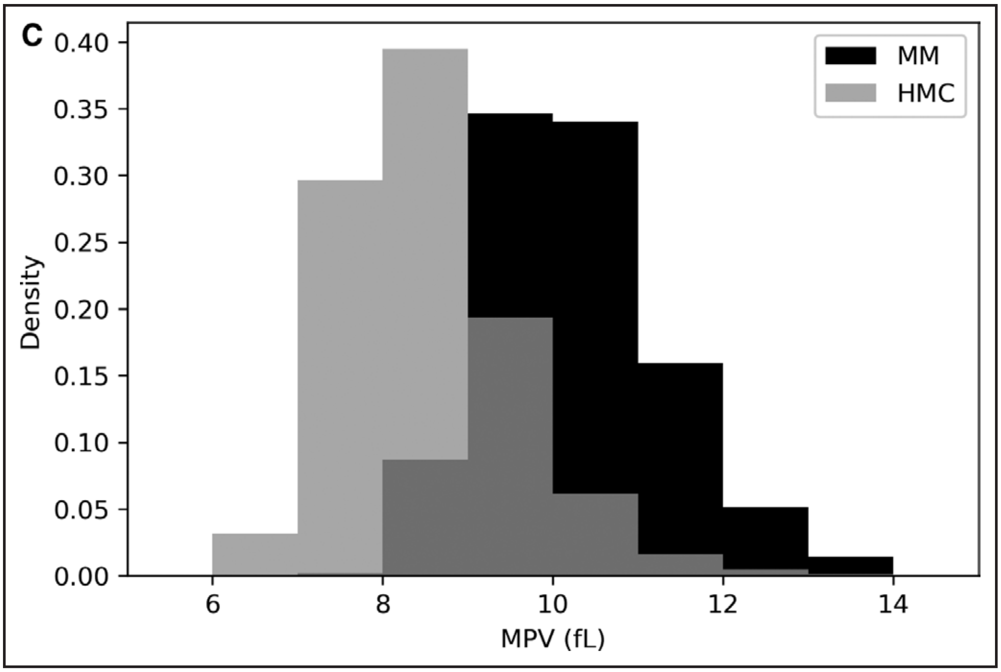

Figure 1. *(Continued).* **C**, Histogram of mean platelet volume (MPV), which had the largest mean difference between the two institutions. Features were summarized over the encounter level (e.g., did the patient have a measurement taken during the encounter). BUN = blood urea nitrogen, GCS = Glasgow Coma Score, INR = international normalized ratio, MAP = mean arterial pressure, MCH = mean corpuscular hemoglobin, MCHC = MCH concentration, MCV = mean corpuscular volume, RDW = red cell distribution width, PT = prothrombin time, PTT = partial thromboplastin time, Spo_2 = oxygen saturation.

As a starting point, an initial threshold value was chosen corresponding to an encounter level sensitivity of 0.5 using data from MM. This threshold resulted in an adjusted PPV of 0.257 (WDR 3.9) and specificity of 0.962. When the same threshold was applied to HMC, it resulted in higher PPV (0.452, WDR 2.2) and specificity (0.989), but at the cost of a lower sensitivity (0.360) than originally calibrated. However, if a new threshold is chosen at HMC (again using a sensitivity of 0.5), the PPV (0.279, WDR 3.6) and specificity (0.967) become more consistent with their intended values. For comparison, at a sensitivity of 0.5, the EDI only achieved a PPV of 0.146 (WDR 6.8) at MM, and 0.160 (WDR 6.3) at HMC after recalibration. A full table of threshold performance is available in **Supplement 11** (http://links.lww.com/CCM/H305).

DISCUSSION

Characterization of Institutional Differences

In the strictest sense, any difference between two institutions can result in significant dataset shift. As Subbaswamy and Saria (19) phrase it: “Even slight deviations from the training conditions can result in wildly different performance.” Understanding and quantifying these differences is an important step in countering this last-mile problem. Changes in how the input features are distributed is perhaps the most nebulous and have a direct effect on the ultimate model performance. One source of this shift can be information management. Examples include variables being reported in different units or with different names or mappings. Using standardized clinical ontologies such as Logical Observation Identifiers Names and Codes (LOINC) codes or the Observational Medical Outcomes Partnership common data model could significantly expedite this process (27, 28).

Data shift can also occur due to clinical differences that are more difficult to detect. Examples include using different equipment with different measurement properties and reference ranges; alternate treatment patterns, which affect the frequency and timing of orders; and differences in disease prevalence and patient demographics. In particular, this can affect missingness rates which can have an outsized effect on predictive performance if the model is allowed to rely on them (23).

Another key difference we identified was in transfer patterns. While both institutions have level 1 Trauma Center certification, MM has several other trauma centers within close proximity, whereas HMC is the only such hospital in the region. This may account for the larger percentage of HMC patients being admitted through the ED, while a larger proportion of admissions at MM are through the OR or catheterization laboratory values. This may have important implications for the types of patients seen by each hospital. Additionally, OR to general floor transfers were considerably higher at MM than at HMC (Supplement 10, http://links.lww.com/CCM/H305).

TABLE 2.
Evaluation of Predicting Intensive Care Transfers and Other Unforeseen Events, Epic Deterioration Index, and National Early Warning Score Across Demographic Groups

Demographic Group	*n* (%)	Predicting Intensive Care Transfers and Other Unforeseen Events	Epic Deterioration Index	National Early Warning Score	Event Rate
Michigan Medicine					
Total	44,202	0.870 (0.859–0.880)	0.830 (0.819–0.841)	0.817 (0.805–0.828)	4.5% (1,998/44,202)
Race					
Asian	855 (2.0%)	0.845 (0.773–0.917)	0.843 (0.771–0.916)	0.817 (0.740–0.893)	5.5% (45/855)
Black	5,575 (12.6%)	0.862 (0.834–0.889)	0.802 (0.771–0.833)	0.820 (0.789–0.850)	5.1% (282/5,575)
White	35,480 (80.3%)	0.871 (0.860–0.882)	0.835 (0.822–0.847)	0.816 (0.803–0.829)	4.4% (1,561/35,480)
Other	2,292 (5.2%)	0.878 (0.835–0.920)	0.830 (0.783–0.878)	0.824 (0.775–0.872)	4.8% (110/2,292)
Ethnicity					
Hispanic/ Latino	1,196 (2.7%)	0.908 (0.870–0.952)	0.823 (0.766–0.891)	0.821 (0.756–0.888)	3.9% (47/1,196)
Non-Hispanic/ Latino	42,210 (95.5%)	0.868 (0.859–0.877)	0.830 (0.820–0.840)	0.816 (0.807–0.826)	4.5% (1,910/42,210)
Other	652 (1.5%)	0.885 (0.821–0.972)	0.815 (0.731–0.911)	0.809 (0.729–0.909)	4.9% (32/652)
Sex					
Female	21,592 (48.8%)	0.885 (0.870–0.899)	0.850 (0.833–0.867)	0.838 (0.821–0.855)	3.8% (829/21,592)
Male	22,610 (50.5%)	0.855 (0.842–0.869)	0.811 (0.796–0.827)	0.802 (0.786–0.817)	5.2% (1,169/22,610)
Hurley Medical Center					
Total	11,083	0.875 (0.848–0.902)	0.835 (0.805–0.864)	0.819 (0.789–0.850)	2.5% (278/11,083)
Race					
Asian	18 (0.2%)	NA	NA	NA	0% (0/18)
Black	5,044 (45.5%)	0.887 (0.850–0.924)	0.857 (0.816–0.897)	0.834 (0.792–0.877)	2.6% (133/5,044)
White	5,641 (50.9%)	0.863 (0.824–0.903)	0.815 (0.771–0.859)	0.801 (0.755–0.846)	2.4% (135/5,641)
Other	380 (3.4%)	0.865 (0.720–1)	0.804 (0.638–0.969)	0.867 (0.721–1)	2.6% (10/380)
Ethnicity					
Hispanic/ Latino	258 (2.3%)	0.703 (0.426–1.00)	0.698 (0.496–0.912)	0.732 (0.506–1.00)	1.6 (4/258)
Non-Hispanic/ Latino	10,776 (97.2%)	0.878 (0.856–0.902)	0.838 (0.812–0.869)	0.821 (0.793–0.855)	2.5% (271/10,776)
Other	49 (0.4%)	0.989 (0.979–1.00)	0.883 (0.765–1.00)	0.648 (0.506–0.795)	6.1% (3/49)
Sex					
Female	6,000 (54.1%)	0.899 (0.864–0.935)	0.869 (0.830–0.909)	0.859 (0.818–0.900)	2.2% (131/6,000)
Male	5,083 (45.9%)	0.850 (0.810–0.889)	0.803 (0.760–0.846)	0.786 (0.741–0.830)	2.9% (147/5,083)

NA = not applicable.

Encounter level area under the receiver operating characteristic curve (AUROC) for each of the three analytics is presented below, separated by subgroup. Performance metrics are reported as encounter level AUROC (95% CI) on the Epic Deterioration Index-matched cohort. Event rate is reported as both a percentage and fraction of encounters. AUROC values are reported as NA if no patient in the subgroup met a deterioration outcome.

TABLE 3.
Evaluation of Predicting Intensive Care Transfers and Other Unforeseen Events, Epic Deterioration Index, and National Early Warning Score

		Predicting Intensive Care Transfers and Other Unforeseen Events		Epic Deterioration Index		National Early Warning Score	
Granularity	Metric	MM	HMC	MM	HMC	MM	HMC
Observation	AUROC (95% CI)	0.813[a,b] (0.812–0.815)	0.844[a,b] (0.841–0.848)	0.769[b] (0.768–0.770)	0.776 (0.771–0.780)	0.751 (0.749–0.752)	0.777 (0.773–0.781)
	AUPRC (95% CI)	0.077[a,b] (0.075–0.078)	0.094[a,b] (0.089–0.098)	0.051[b] (0.050–0.052)	0.060 (0.056–0.063)	0.040 (0.039–0.041)	0.056 (0.053–0.059)
	Event rate	0.8%	0.6%	0.8%	0.6%	0.8%	0.6%
Encounter	AUROC (95% CI)	0.870[a,b] (0.861–0.878)	0.875[a,b] (0.851–0.902)	0.830[b] (0.821–0.840)	0.835 (0.808–0.863)	0.817 (0.806–0.827)	0.819 (0.792–0.851)
	AUPRC (95% CI)	0.298[a,b] (0.275–0.320)	0.339[a,b] (0.281–0.398)	0.201[b] (0.182–0.218)	0.231 (0.180–0.276)	0.171 (0.154–0.184)	0.233 (0.180–0.281)
	Event rate	4.5%	2.5%	4.5%	2.5%	4.5%	2.5%

AUPRC = area under the precision-recall curve, AUROC = area under the receiver operating characteristic curve, HMC = Hurley Medical Center, MM = Michigan Medicine.

If no simulations had a difference less than 0, the *p* value is reported as $p < 0.001$, indicated by

[a](vs Epic Deterioration Index [EDI]) and

[b](vs National Early Warning Score [NEWS]).

AUROC and AUPRC were used to describe performance at two levels of granularity: the observation level, where each observation is treated independently (i.e., did the patient deteriorate in the next 24 hr?), and the encounter level, which describes a patient's maximum score during their encounter compared with their ultimate outcome (i.e., did they ever deteriorate during their encounter?). Due to the difference in observation and encounter level event rates between the two institutions, precision (and thus AUPRC) at hospital 1 (MM) was standardized to match the event rate at hospital 2 (HMC) to facilitate a more direct comparison. 95% CIs were computed for encounter level statistics via bootstrap with 1,000 replications to compute pivotal CIs. For observation level statistics, the bootstrap was blocked to ensure randomization both between encounters and in observations within an encounter. *p* values were computed for differences in AUROC and AUPRC by counting the fraction of bootstrapped differences in evaluation metrics < 0. Compared with its original published performance (AUROC 0.87), NEWS was observed to experience a substantial drop in performance both in (26) (0.72–0.76) and in our own analysis (0.751 at MM, 0.777 at HMC). Publicly available information surrounding the training of the EDI is scarce, but Epic's large data resources spanning many institutions likely contributes to its successful performance at both MM and HMC.

Geographic differences between the two hospitals produce significant variations in the demographic makeup, for example, MM's population contains significantly more White and fewer female patients (Table 1). Despite differences in event rates at both institutions, there was no significant difference in PICTURE's performance across any of the races at either institution (Table 3). While males were more likely to deteriorate than females, all three scores performed better in females than in males at both institutions.

Evaluation of PICTURE Performance Across Institutions

PICTURE's performance was consistent between the two institutions, with a slight increase in both observation and encounter level AUROC and AUPRC at the second institution, although the encounter level results fell well within their 95% CIs. Thus, despite the many inter-institutional differences, PICTURE was able to successfully generalize outside its home institution and account for the last-mile problem of dataset shift. While both the EDI and NEWS scores also saw a similar slight increase in performance between the two institutions, the comparison between the two hospitals for EDI and NEWS is different than in the case of PICTURE. This is because both MM and HMC represent external sites for the EDI and NEWS. Hence, the similarity between the performance of EDI and NEWS at the two hospitals does not reflect generalization from model development to external validation. Unlike PICTURE, a prior study demonstrated that NEWS

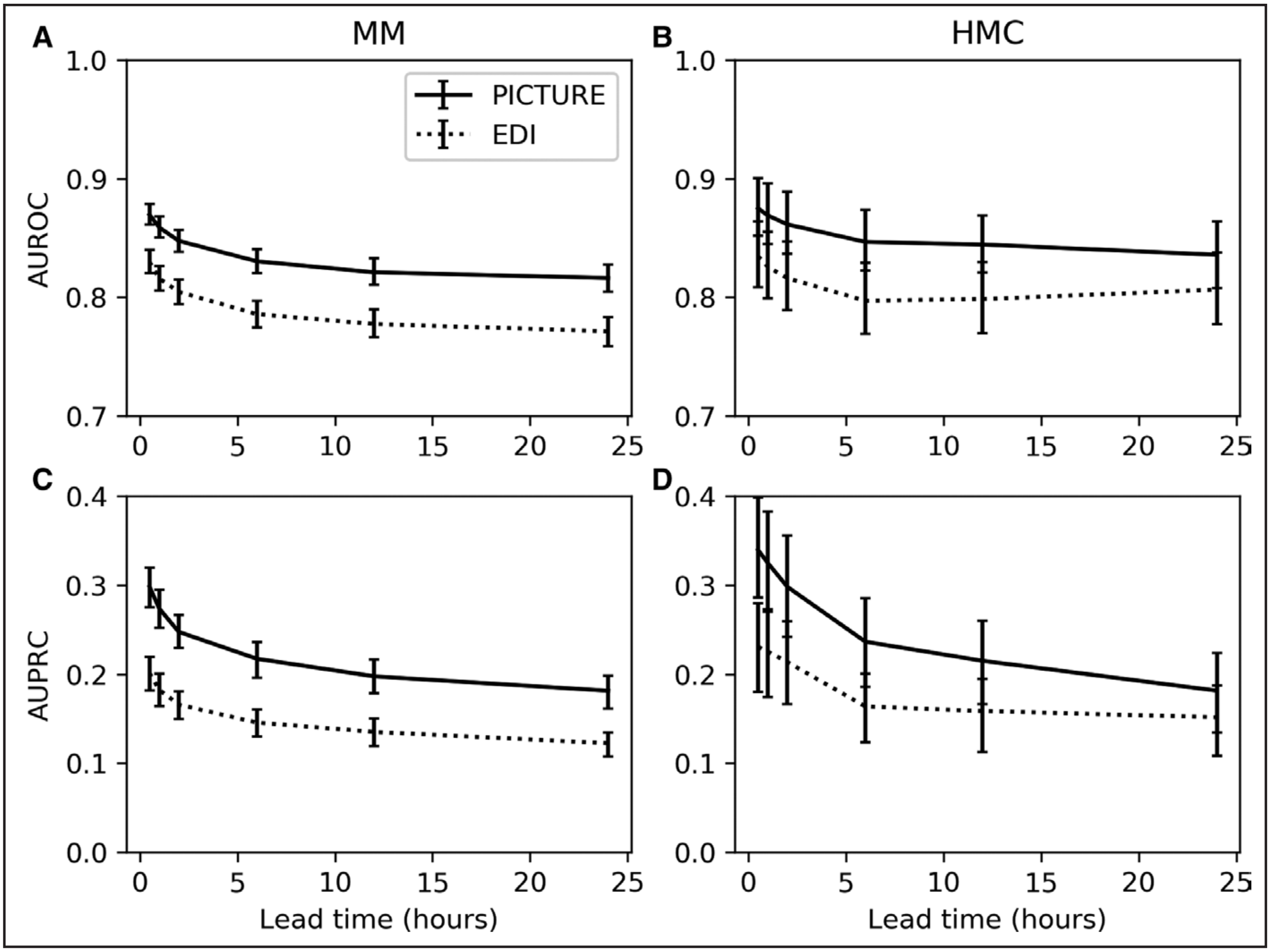

Figure 2. Lead time simulation. Area under the receiver operating characteristic curve (AUROC) and area under the precision-recall curve (AUPRC) were evaluated for Predicting Intensive Care Transfers and other UnfoReseen Events (PICTURE) and Epic Deterioration Index (EDI) by calculating the maximum prediction score prior to *x* hr before the deterioration event, with *x* ranging from 0.5 to 24 hr. Twenty-four hr was selected as the limit since, during model training, only observations less than 24 hr in advance of the deterioration were labeled as positive. AUPRC is again adjusted to the event rate of 2.5% to match the second hospital (Hurley Medical Center [HMC]). *Error bars* representing 95% CIs are reported using the 1,000-replicate bootstrap described previously. **A**, AUROC at hospital 1 (Michigan Medicine [MM]) for PICTURE and EDI scores. **B**, AUROC at hospital 2 (HMC). **C**, AUPRC at MM. **D**, AUPRC at HMC.

sustained an 11–15% drop in AUROC (0.87 vs 0.72–0.76) when moving from its original dataset to two external test sets (AUPRC was not reported), which is roughly consistent with the MM and HMC AUROCs (0.751 and 0.777) reported here (26). In other words, NEWS has already experienced a drop in performance. There is little information publicly available surrounding the training of the EDI, but Epic's large data resources spanning many institutions likely contributes to its successful generalization at both MM and HMC.

We attribute PICTURE's ability to generalize to two factors: a carefully designed multiple imputation mechanism which disguises missingness patterns and our focus on physiologic features (in contrast to indicators of clinician behavior). Both of these guard against changing patterns in patient care guidelines, which can change both between institutions and in time and alter performance (23). Additionally, the EDI was trained on a multicenter cohort, and NEWS similarly constructed from existing EWSs at multiple hospitals in the U.K.'s National Health System (29, 30). In all comparisons, however, PICTURE significantly ($p < 0.001$) outperformed both tools. PICTURE also performed consistently across varying lead times. This lead time gives clinicians an increased opportunity to act on the alert before the patient deteriorates.

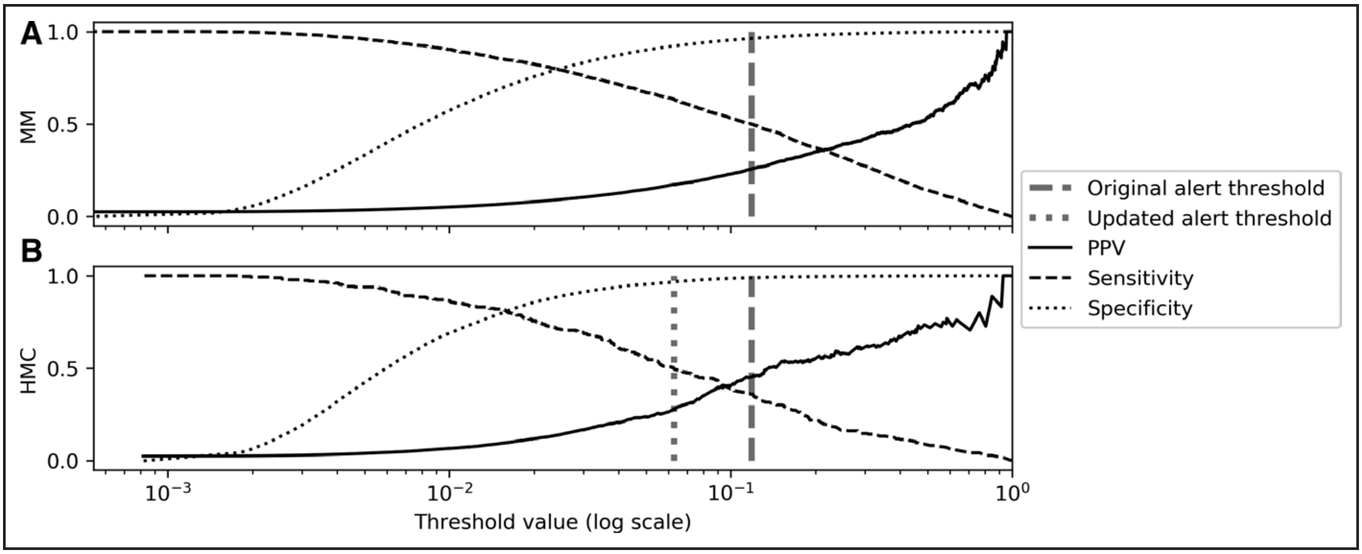

Figure 3. Positive predictive value (PPV), sensitivity, and specificity change with alert threshold. PPV, sensitivity, and specificity are plotted at varying thresholds. A candidate threshold was selected at a sensitivity of 0.5 using data from hospital 1 (**A**, *dark dashed line*), and then applied to data from hospital 2 (**B**, *dark dashed line*). Note that PPV at hospital 1 is adjusted to reflect the event rate at the second institution. A second candidate alert threshold (*light dotted line*) was chosen using the same procedure on data from hospital 2 to indicate the possible desirability of choosing separate thresholds to better fit clinical care in the different environments. HMC = Hurley Medical Center, MM = Michigan Medicine.

Alert thresholds are another critical piece of model implementation, as they allow control over the number and quality of alerts generated and impact clinician perception of the model. For example, if the threshold is set too low, alerts will be generated too frequently and PPV will decrease, contributing to alert fatigue. In the other direction, a threshold set too high will miss patients (low sensitivity). However, dataset shift—and most notably, changes in event rate—can impact the performance of a set threshold. Customizing the alert threshold between institutions can be used to ensure the model is fitting the needs of clinical practice at each location. Thus, it is important to validate not only aggregate model performance (e.g., AUROC and AUPRC) but also individual alert thresholds.

Limitations

While we were successfully able to demonstrate PICTURE's generalizability at a second institution, validation across further hospitals would ensure portability across a wider variety of clinical settings. Second, this study uses retrospective data from both hospitals. The model is currently implemented in real-time at MM, and prospective data are being collected for further validation. A follow-up study is underway to evaluate methods for alert implementation, including threshold selection and notification delivery.

CONCLUSIONS

Moving a predictive model to a new clinical environment outside that which it was trained is fraught with challenges. Key differences were observed between the initial academic tertiary care center and a second large community hospital, including changes in the distribution of laboratory values and vital signs, frequency of deterioration, and changes in demographic makeup. Despite these differences, PICTURE was able to consistently predict deterioration events and outperform existing metrics at both institutions, and supports its suitability as an early-warning reminder tool to predict deterioration in general ward patients across different clinical settings. However, despite this initial success, it remains critically important to continuously monitor EWSs both throughout and long after implementation to account for ongoing dataset shifts in an evolving healthcare system.

[1] The Max Harry Weil Institute of Critical Care Research & Innovation, University of Michigan, Ann Arbor, MI.

[2] Department of Emergency Medicine, University of Michigan Medical School, Ann Arbor, MI.

[3] Information Technology, Hurley Medical Center, Flint, MI.

[4] Department of Emergency Medicine, Hurley Medical Center, Flint, MI.

[5] Department of Internal Medicine, University of Michigan Medical School, Ann Arbor, MI.

6 Medicine Service, LTC Charles S. Kettles VA Medical Center, Ann Arbor, MI.

7 Department of Learning Health Sciences, University of Michigan Medical School, Ann Arbor, MI.

8 Precision Health, University of Michigan, Ann Arbor, MI.

9 Department of Biomedical Engineering, University of Michigan, Ann Arbor, MI.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's website (http://journals.lww.com/ccmjournal).

Mr. Cummings's, Mr. Blackmer's, Dr. Farzaneh's, Ms. Cao's, and Dr. Ansari's institutions received funding from the Michigan Institute for Data Science and Airstrip Technologies. Mr. Cummings, Mr. Blackmer, Dr. Farzaneh, Ms. Cao, and Drs. Gillies, Medlin, Ward, and Ansari have disclosed that multiple patents have been filed for this work and invention disclosures have been submitted with the Office of Technology Transfer, University of Michigan, Ann Arbor, and that Airstrip Technologies has a license option for Predicting Intensive Care Transfers and other UnfoReseen Events from the University of Michigan. Mr. Motyka disclosed that he is currently employed by at Strata Oncology (Precision Oncology). Dr. Gillies disclosed that he is employed at Regeneron Pharmaceuticals. Dr. Admon's institution received funding from the National Heart, Lung, and Blood Institute. Drs. Admon and Sjoding received support for article research from the National Institutes of Health (NIH). Dr. Singh's institution received funding from Blue Cross Blue Shield of Michigan and Teva Pharmaceuticals; he received funding from Flatiron Health. Dr. Sjoding's institution received funding from the NIH. The remaining authors have disclosed that they do not have any potential conflicts of interest.

For information regarding this article, E-mail: cummingb@med.umich.edu

REFERENCES

1. Allen J, Currey J, Jones D, et al: Development and validation of the medical emergency team-risk prediction model for clinical deterioration in acute hospital patients, at time of an emergency admission. *Crit Care Med* 2022; 50:1588–1598
2. Saab A, Abi Khalil C, Jammal M, et al: Early prediction of all-cause clinical deterioration in general wards patients: Development and validation of a biomarker-based machine learning model derived from rapid response team activations. *J Patient Saf* 2022; 18:578–586
3. Reardon PM, Seely AJE, Fernando SM, et al: Can early warning systems enhance detection of high risk patients by rapid response teams? *J Intensive Care Med* 2021; 36:542–549
4. Fernandes M, Mendes R, Vieira SM, et al: Predicting intensive care unit admission among patients presenting to the emergency department using machine learning and natural language processing. *PLoS One* 2020; 15:e0229331
5. Churpek MM, Yuen TC, Winslow C, et al: Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. *Crit Care Med* 2016; 44:368–374
6. Kipnis P, Turk BJ, Wulf DA, et al: Development and validation of an electronic medical record-based alert score for detection of inpatient deterioration outside the ICU. *J Biomed Inform* 2016; 64:10–19
7. Desautels T, Calvert J, Hoffman J, et al: Using transfer learning for improved mortality prediction in a data-scarce hospital setting. *Biomed Inform Insights* 2017; 9:1178222617712994
8. Desautels T, Das R, Calvert J, et al: Prediction of early unplanned intensive care unit readmission in a UK tertiary care hospital: A cross-sectional machine learning approach. *BMJ Open* 2017; 7:e017199
9. Alvarez CA, Clark CA, Zhang S, et al: Predicting out of intensive care unit cardiopulmonary arrest or death using electronic medical record data. *BMC Med Inform Decis Mak* 2013; 13:28
10. Green M, Lander H, Snyder A, et al: Comparison of the between the flags calling criteria to the MEWS, NEWS and the electronic Cardiac Arrest Risk Triage (eCART) score for the identification of deteriorating ward patients. *Resuscitation* 2018; 123:86–91
11. Escobar GJ, LaGuardia JC, Turk BJ, et al: Early detection of impending physiologic deterioration among patients who are not in intensive care: Development of predictive models using data from an automated electronic medical record. *J Hosp Med* 2012; 7:388–395
12. Churpek MM, Yuen TC, Park SY, et al: Using electronic health record data to develop and validate a prediction model for adverse outcomes in the wards*. *Crit Care Med* 2014; 42:841–848
13. Churpek MM, Yuen TC, Winslow C, et al: Multicenter development and validation of a risk stratification tool for ward patients. *Am J Respir Crit Care Med* 2014; 190:649–655
14. Chen L, Ogundele O, Clermont G, et al: Dynamic and personalized risk forecast in step-down units. Implications for monitoring paradigms. *Ann Am Thorac Soc* 2017; 14:384–391
15. Cabitza F, Campagner A, Balsano C: Bridging the "last mile" gap between AI implementation and operation: "Data awareness" that matters. *Ann Translat Med* 2020; 8:501–501
16. Coiera E: The last mile: Where artificial intelligence meets reality. *J Med Internet Res* 2019; 21:e16323
17. Habib AR, Lin AL, Grant RW: The epic sepsis model falls short—the importance of external validation. *JAMA Int Med* 2021; 181:1040–1041
18. Finlayson SG, Subbaswamy A, Singh K, et al: The clinician and dataset shift in artificial intelligence. *N Engl J Med* 2021; 385:283–286
19. Subbaswamy A, Saria S: From development to deployment: Dataset shift, causality, and shift-stable models in health AI. *Biostatistics* 2020; 21:345–352
20. Ramspek CL, Jager KJ, Dekker FW, et al: External validation of prognostic models: What, why, how, when and where? *Clin Kidney J* 2021; 14:49–58
21. Siontis GCM, Tzoulaki I, Castaldi PJ, et al: External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *J Clin Epidemiol* 2015; 68:25–34
22. Cummings BC, Ansari S, Motyka JR, et al: Predicting intensive care transfers and other unforeseen events: Analytic model validation study and comparison to existing methods. *JMIR Med Informat* 2021; 9:e25066

23. Gillies CE, Taylor DF, Cummings BC, et al: Demonstrating the consequences of learning missingness patterns in early warning systems for preventative health care: A novel simulation and solution. *J Biomed Inform* 2020; 110:103528
24. Fleishon HB, Itri JN, Boland GW, et al: Academic medical centers and community hospitals integration: Trends and strategies. *J Am Coll Radiol* 2017; 14:45–51
25. Hagberg A, Swart P, Schult D: Exploring network structure, dynamics, and function using NetworkX. Proceedings of the 7th Python in Science Conference. Pasadena, CA, 2008, pp 11–15
26. Linnen DT, Escobar GJ, Hu X, et al: Statistical modeling and aggregate-weighted scoring systems in prediction of mortality and ICU transfer: A systematic review. *J Hosp Med* 2019; 14:161–169
27. McDonald CJ, Huff SM, Suico JG, et al: LOINC, a universal standard for identifying laboratory observations: A 5-year update. *Clin Chem* 2003; 49:624–633
28. Hripcsak G, Duke JD, Shah NH, et al: Observational Health Data Sciences and Informatics (OHDSI): Opportunities for observational researchers. *Stud Health Technol Inform* 2015; 216:574–578
29. Singh K, Valley TS, Tang S, et al: Evaluating a widely implemented proprietary deterioration index model among hospitalized patients with COVID-19. *Ann Am Thorac Soc* 2021; 18:1129–1137
30. Royal College of Physicians: National Early Warning Score (NEWS): Standardising the Assessment of Acute Illness Severity in the NHS. London, United Kingdom, Royal College of Physicians, 2012