# Supplement

**Supplement 1: Epic Deterioration Index (EDI) and NEWS**

To provide a point of comparison, both institutions use the Epic Deterioration Index (EDI), a proprietary model developed by the Epic Systems Corporation. Similar to PICTURE, it uses structured clinical data to predict deterioration outcomes such as death, ICU transfer, and resuscitation. Its use in predicting deterioration in COVID-19 patients has been evaluated at MM (1). In contrast to PICTURE, which is calculated each time a new data point is available, the EDI is calculated every 15-20 minutes. However, many details surrounding its training and validation have not been shared publicly.

The National Early Warning Score (NEWS) is a second index used to detect patients at an increased risk of deterioration including death, cardiac arrest, and ICU transfer (2, 3). As a scoring system, its basis is openly available and was calculated as described in Smith et. al. It was calculated retrospectively each time a new data point was updated, in a similar fashion to PICTURE.

Of the three, NEWS is the simplest model, being an aggregate rule-based model and using the smallest feature set. It is primarily concerned with vital signs, including respiratory rate, oxygen saturation, supplemental oxygen, temperature, systolic blood pressure, heart rate, and mental status. EDI adds slightly more complexity, being based on logistic regression and adding additional variables such as age, cardiac rhythm, and laboratory values (hematocrit, white blood cell count, potassium, sodium, pH, platelets, and blood urea nitrogen) to those found in NEWS (1). PICTURE is a larger XGBoost model and uses the widest array of features, a full list of which is noted in Supplemental Table 6. While there are advantages to using simpler models such as NEWS – most notably the ability to calculate scores by hand – these have been found to come at the cost of lower PPV and sensitivity (4, 5) and lack the ability to be customized or retrained in new contexts or situations (6). Similarly, linear scores like the EDI) can also fail to fully capture the complexity of clinical care and physiology and fall short of deeper machine-learning methods (7).

# Supplement 2: Institutional Review Board Approval

This study was approved with a waiver of informed consent for secondary data use by the respective 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.”

**Supplement 3: Initial model construction and internal validation**

The model uses common structured data from the EHR as its input, including vital signs, laboratory values, and demographics. Prediction targets were defined as death, ICU transfer or accommodation, mechanical ventilation, or cardiac arrest within 24 hours. It was trained and tuned on 131,546 encounters from a large academic medical system (MM) encompassing the years 2014 to 2018 and validated on a hold-out test set of 33,472 encounters from 2019 and an additional 637 COVID-19 positive encounters from 2020. If a patient with multiple hospital encounters appeared in the test set, any encounter prior to 2019 was removed to prevent overlap between the training and test sets. Inclusion criteria were patients 18 years of age or older who were hospitalized as an inpatient or other observation status, and patients who experienced ICU transfers from locations other than the general wards (e.g. operating rooms) were removed to exclude planned ICU transfers. The data was forward-filled such that each observation represented the most recently updated value at each timepoint, and remaining missing values (prior to the first observed value for each variable) were imputed using a Bayesian iterative imputer designed to prevent the model from learning patterns in missingness (1). These missingness patterns represent an important component of dataset shift, as different institutions (or even different units within the same institution) may have different clinical criteria which affect how often and which types of labs or vital signs are collected (2, 3). Learning these patterns may artificially boost performance and reduce generalizability (1). In addition, missingness patterns can encode clinician behavior (e.g., presence of a blood lactate laboratory value indicates clinicians’ suspicion of impending deterioration and vice versa), leading to a model that relies on late indicators of deterioration instead of patient physiology and provides less novel alarms. Classification was conducted using XGBoost v. 0.90, an open-source implementation of a gradient-boosting tree framework, using a binary cross-entropy objective function with a tree depth of three nodes and a learning rate of 0.05. Model explanations at the prediction level were determined via Shapley value (4, 5). A full list of features, as well as their median, IQR, and missingness rate, are presented in the supplement (Supplement 6). Further details on the model construction and internal validation, are available in our previous publication (6).This study was developed following TRIPOD reporting guidelines, with the checklist available as a separate supplement (7).

**Supplement 4: Data preprocessing**

EHR systems are designed to account for the great deal of flexibility that occurs with different clinical workflows. Even though each hospital uses the same EHR vendor (Epic Systems, Inc.), the actual implementation of the health record varies significantly between institutions, as each hospital system can customize their data elements to meet their specific patient care needs. This is often referred to as the “curly braces problem” in health informatics. During the initial construction of the PICTURE model, great care was used when selecting the specific flowsheet names and laboratory identifiers that make up PICTURE’s input features. Because of these customized data elements, the fields selected at the first institution (MM) needed to be mapped to those at the second (HMC).

As we were not able to directly access the EHR at HMC, we relied on a team of informaticists, clinicians, and subject matter experts at both institutions to coordinate this mapping. We provided a list of the fields we used at our own institution, as well as descriptions of the desired content. This took the form of two lists, one for each of the two primary data types (flowsheet rows and laboratory values) which were sent to the HMC team for mapping. Both teams reviewed the final mappings for accuracy.

After this mapping, the hospital-specific element identifiers were used to extract data from our cohort. The data was then deidentified and encrypted to protect PHI and returned to the investigators at the first institution for model application and evaluation.

After extraction, data from both the first (MM) and second (HMC) institutions were processed in parallel in order to ensure that identical steps were applied to each cohort. The data consisted of patients from both institutions with an admission date between January 1, 2015 and November 23, 2020. Since the model had been trained on data up until December 31, 2018, these patients were removed from the MM dataset so that the evaluation occurred only on unseen test data. We further refined the cohort to only include 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 in order to exclude planned transfers.

After cohort selection, we applied our target definitions from the previous study to the data. Observations within 24 hours of a deterioration event were regarded as positives, and all others as negatives. Observations occurring 30 minutes before the first event and later were removed to avoid making predictions using data that was unlikely to be available for analysis in real-time; this is consistent with previously described approaches (8). The target outcomes were death, ICU transfer or accommodation, and new need for mechanical ventilation. At both institutions, the calculation of these targets was analogous: mortality data was taken from the patient summary table at both institutions, ICU transfer or accommodation from the Admissions/Discharge/Transfer (ADT) table, and mechanical ventilation use was derived from the ventilator associated flowsheet documentation. In contrast to Cummings 2021, cardiac arrest and the administration of IV fluids were not included due to data availability. The final target was constructed as a composite outcome, where observations 24 hours in advance of the first event (regardless of type of deterioration) were marked as positive and all subsequent observations were removed. Thus the prediction task was binarized and does not distinguish between event types.

Missing values were forward-filled by result time on an encounter-level basis such that each observation contained the most up-to-date data points available at any given time. Remaining missing values (e.g. if a given lab had not yet been drawn) were imputed using the same iterative imputation mechanism developed in Gillies 2020 and trained in Cummings 2021 in an effort to disguise missingness patterns from the model.

For this evaluation, only observations made while the patient was physically located on a general floor were considered. This is for two reasons: first, it more closely matches the realities of prospective clinical implementation, in which the clinicians on these units would be our primary target for alerts. Secondly, a substantially higher proportion of observations occur outside of general wards at the second institution which makes direct comparison more difficult. Since our target use-case is on the general wards, restricting the evaluation to these units facilitates a more direct and relevant comparison.

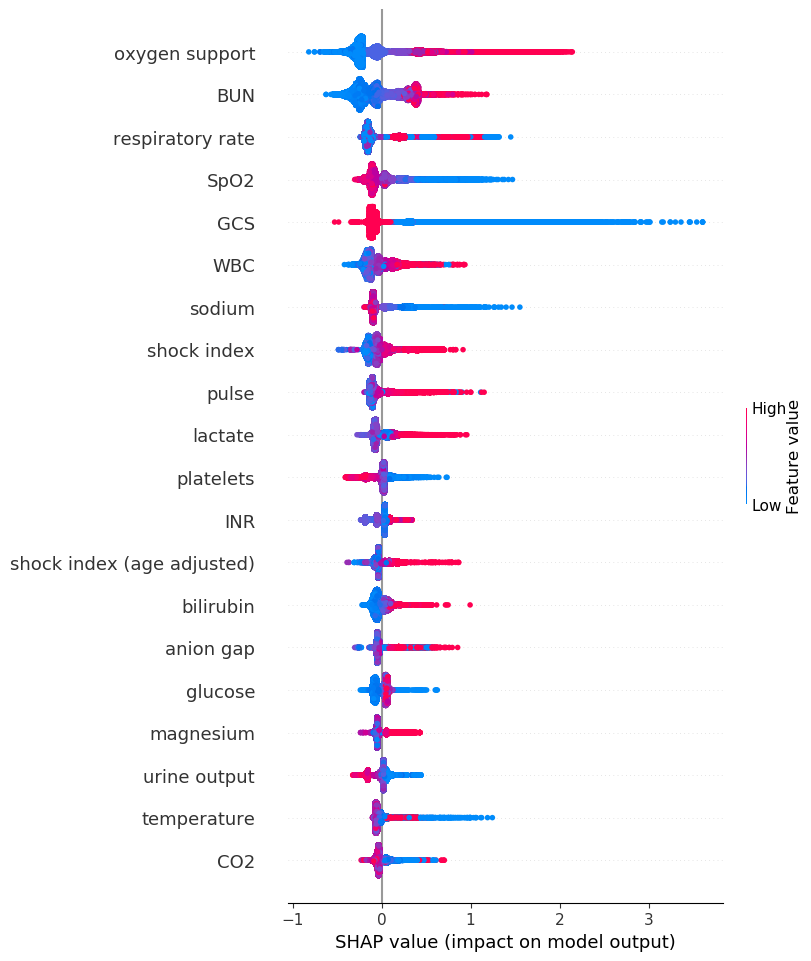
**Supplement 5: Shapley values for prediction explanation**

One key feature of the PICTURE model is its ability to provide explanations alongside each prediction. This is accomplished through the use of Shapley values, which assign an influence score to each feature (9). This allows the most important feature, as well as their approximated contribution, to be forwarded to the clinician along with the alert. Given that PICTURE is targeting a wide range of deterioration causes, it provides much needed contextual information that the clinician can use to respond to the alert.

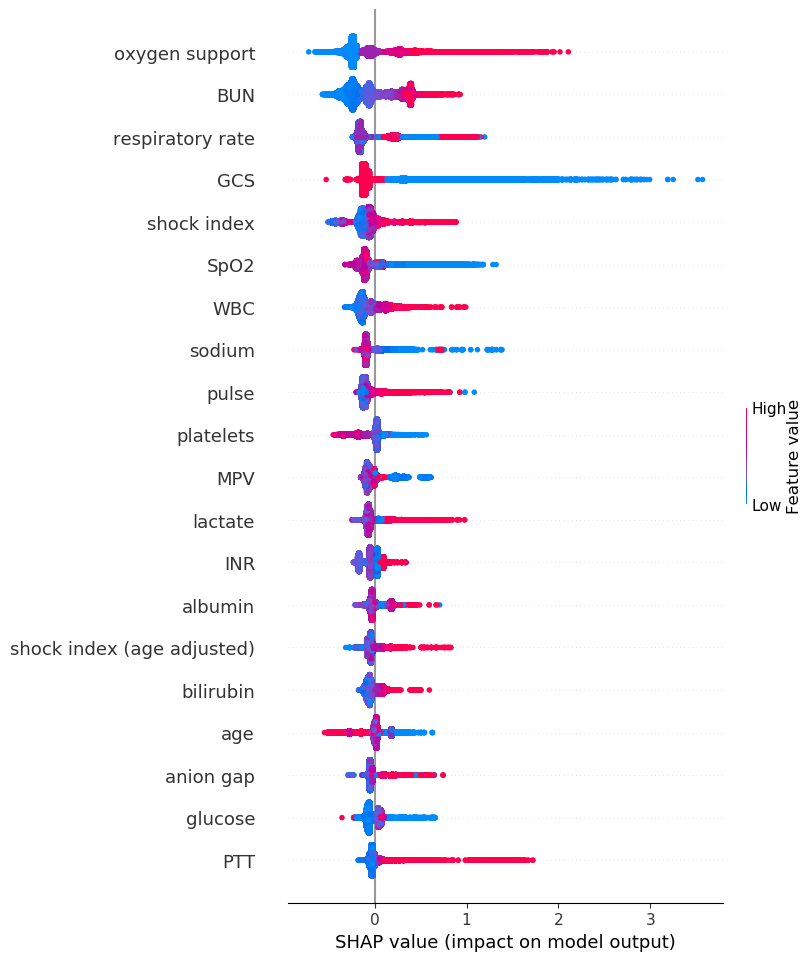
At each prediction, each feature is assigned an estimated contribution score using Shapley analysis. This results in a matrix of identical shape to the input matrix, but where feature values are replaced by their corresponding importance. In addition to its utility on an individual prediction level - giving clinicians a starting point when responding to the alert - these are also useful in aggregate to display the degree to which the PICTURE model has learned relevant physiology.

Figure S5.1 below depicts the 20 features that were found to be the most useful overall, determined by the mean absolute value across all predictions. Feature value (e.g. low or high oxygen support for the topmost feature) is indicated by color, while the importance is described by position on the x-axis. Highly positive values indicate the feature (e.g. high oxygen support pushed the model toward predicting deterioration), while negative values indicate the inverse. Many of the most important features are consistent between the two institutions, but there are differences present in how they are ranked. The directions of the relationship between a variable's magnitude and contribution score were also consistent between the two institutions - e.g. higher amounts of oxygen support pushed the model towards predicting deterioration in both cases.

***Figure S5.1: Shapley summary plot.*** *Features were ranked according to overall importance (mean of the absolute value) and the top 20 features at each institution displayed below.* ***Panel A*** *represents the primary institution (MM) and* ***Panel B*** *represents the second (HMC). The relative value of each variable (e.g. the amount of oxygen support) is indicated by color, while its importance is indicated by the location on the x-axis.*

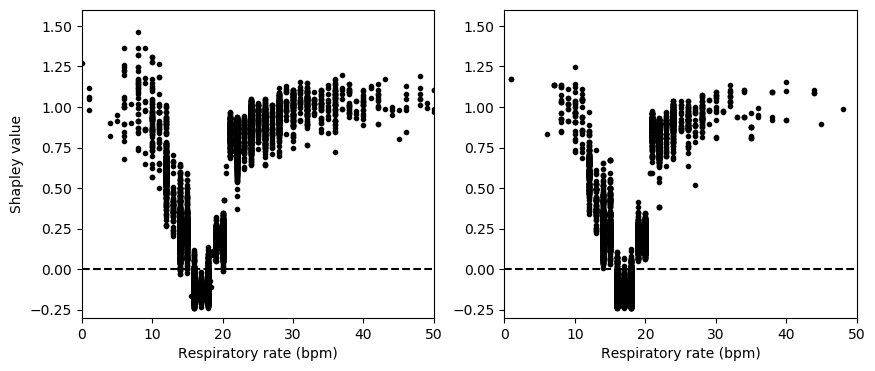


**A**



**B**

Many of these value/importance relationships are nonlinear - for example, respiratory rate. Figure S5.2 demonstrates the relationship between Shapley value and feature value in more detail. The model has learned that respiratory rates in the 15-20 bpm range are associated with lower risk (negative Shapley values), while values toward the extremes push the model toward predicting deterioration. These relationships were preserved despite the model being moved from one institution to another.



**A**

**B**

**Figure S5.2: Shapley values for respiratory rate.** Shapley values, estimates of feature contribution, are displayed against feature value for an example variable. Respiratory rate was used as an example as it is the highest-ranked feature that demonstrated a clearly nonlinear relationship (Figure 5.1). **Panel A** displays the results at Hospital 1 (MM), while **Panel B** represents the same analysis for Hospital 2 (HMC).

**Supplement 6: Feature list.** Along with all features used in the model, the median, IQR, and missingness rate at each institution is presented. Median, IQR, and missingness rate are calculated on the encounter-level to avoid biasing the calculation towards patients with more frequently-updated results.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Numeric features** | **MM** | | | **HMC** | | |
| **Median** | **IQR** | **Missingness** | **Median** | **IQR** | **Missingness** |
| Age | 61.2 | 47.0-71.7 | 0.0% | 57.6 | 44.0-70.0 | 0.0% |
| Albumin | 3.7 | 3.3-4.1 | 27.4% | 3.7 | 3.3-4.1 | 40.5% |
| Anion gap | 11.8 | 10.3-13.3 | 0.5% | 11.1 | 9.6-12.9 | 0.5% |
| Bicarbonate | 26 | 23.9-29.0 | 50.9% | 25.4 | 22.1-29.7 | 96.4% |
| Bilirubin | 0.6 | 0.4-0.9 | 28.6% | 0.5 | 0.3-0.7 | 40.7% |
| BUN | 16 | 12-24 | 0.6% | 15 | 10-22 | 0.6% |
| Calcium | 8.8 | 8.4-9.1 | 0.4% | 8.6 | 8.2-9.0 | 0.4% |
| Chloride | 105 | 102-107 | 0.4% | 105 | 102-108 | 0.4% |
| CO2 | 26 | 24-28 | 0.5% | 27 | 25-29 | 0.5% |
| Creatinine | 0.9 | 0.7-1.1 | 0.4% | 0.8 | 0.6-1.1 | 0.4% |
| Diastolic | 67 | 60-74 | 0.0% | 69 | 62-77 | 0.0% |
| GCS | 15 | 15-15 | 0.2% | 15 | 15-15 | 0.0% |
| Glucose | 108 | 93-1.6 | 0.3% | 103 | 90-129 | 0.4% |
| Height | 67 | 64-70 | 11% | 66 | 64-70 | 4.6% |
| Hematocrit | 35.0 | 29.9-39.9 | 0.4% | 36.1 | 31.7-40-2 | 0.3% |
| Hemoglobin | 11.3 | 9.5-12.9 | 0.4% | 11.7 | 10.1-13.1 | 0.3% |
| INR | 1.1 | 1.0-1.2 | 41.8% | 1.1 | 1.0-1.2 | 47.4% |
| Lactate | 1.3 | 1.0-1.9 | 51.0% | 1.4 | 1.0-1.9 | 0.77% |
| Magnesium | 1.8 | 1.7-2.0 | 26.1% | 1.9 | 1.7-2.0 | 26.4% |
| MAP | 87 | 79-95 | 0.0% | 89 | 82-97 | 0.0% |
| MCHC | 32.5 | 31.6-33.4 | 0.5% | 32.3 | 31.3-33.2 | 0.4% |
| MCH | 29.7 | 28.1-31.2 | 0.5% | 29.6 | 27.9-31.2 | 0.4% |
| MPV | 10.1 | 9.5-10.8 | 1.3% | 8.3 | 7.8-8.9 | 0.4% |
| Oxygen supplement | 0 | 0-2 | 0.0% | 0 | 0-2 | 0.0% |
| Phosphorus | 3.5 | 3.0-4.0 | 41.2% | 3.4 | 2.9-4.0 | 27.5% |
| Platelets | 214 | 162-275 | 0.5% | 236 | 184-300 | 0.4% |
| Potassium | 4.2 | 3.9-4.4 | 0.4% | 4.1 | 3.8-4.4 | 0.4% |
| Protein | 6.0 | 5.4-6.5 | 28.5% | 3.8 | 3.0-4.5 | 99.4% |
| PT | 10.9 | 10.4-12.0 | 42.5% | 14.1 | 13.3-15.4 | 47.3% |
| PTT | 25.2 | 23.4-28.0 | 52.0% | 30.1 | 27.2-34.3 | 50.7% |
| Pulse pressure | 57 | 49-68 | 0.0% | 57 | 49-68 | 0.0% |
| Pulse | 78 | 70-88 | 0.0% | 78 | 69-88 | 0.0% |
| RBC | 3.9 | 3.3-4.4 | 0.4% | 4.0 | 3.5-4.5 | 0.4% |
| RDW | 14.1 | 13.1-15.9 | 0.5% | 14.6 | 13.6-16.1 | 0.4% |
| Resp. rate | 17 | 16-18 | 0.0% | 18 | 18-18 | 0.0% |
| Shock index | 0.63 | 0.54-0.73 | 0.0% | 0.61 | 0.53-0.70 | 0.0% |
| Shock index (age adjusted) | 36.2 | 28.3-44.6 | 0.1% | 34.2 | 26.1-42.5 | 0.0% |
| Sodium | 139 | 137-140 | 0.4% | 139 | 137-141 | 0.4% |
| Systolic | 126 | 114-138 | 0.0% | 128 | 117-141 | 0.0% |
| Temperature | 36.7 | 36.6-36.9 | 0.0% | 36.7 | 36.6-36.8 | 0.0% |
| Urine output | 250 | 150-400 | 14.4% | 250 | 3-400 | 31.9% |
| Weight | 179 | 148-216 | 1.5% | 176.4 | 145-217 | 1.9% |
| WBC | 8.0 | 6.0-10.5 | 0.4% | 7.6 | 5.8-10.2 | 0.4% |
| SpO2 | 94 | 92-95 | 0.0% | 11.1 | 9.6-12.9 | 0.5% |
| **Categorical features** | **MM** | | | **HMC** | | |
| **Proportion** | **N Positive** | **Missingness** | **Proportion** | **N Positive** | **Missingness** |
| Is Female | 49.5% | 29648 | 0.0% | 56.3% | 20787 | 0.0% |
| Is Asian | 2.0% | 1223 | 0.0% | 0.1% | 72 | 0.0% |
| Is Black | 12.8% | 7646 | 0.0% | 46.7% | 17226 | 0.0% |
| Is White | 80.0% | 47909 | 0.0% | 49.7% | 18376 | 0.0% |
| Is Other | 5.2% | .3085 | 0.0% | 3.4% | 1273 | 0.0% |
| O2 Device | 100.0% | 59862 | 0.0% | 99.9% | 36944 | 0.0% |

**Supplement 7: Alignment of PICTURE to the EDI**

As PICTURE and the EDI generate scores at different frequencies, the two scores had to first be aligned in time. We elected to give EDI any potential advantage in this alignment process by keeping all EDI scores generated by Epic on our cohort, and mapped the most recently-available PICTURE score onto each EDI value (a backward merge). This means that, in the case where PICTURE and the EDI scores did not occur simultaneously, the PICTURE scores would be older than their EDI counterpart and thus could not be using any future information as compared to the EDI. This procedure is described in detail in (8). On average, EDI scores are calculated every half-hour. After alignment, our MM cohort decreased from 59,863 to 44,202 encounters, and our HMC cohort decreased from 36,947 to 11,083 encounters. Note that the EDI score was activated in 2018 for both institutions, and patients before this time do not have EDI scores. PICTURE’s performance on the full cohort is presented in Supplemental Table 8.

***Supplement 8: Evaluation of PICTURE and NEWS performance at native frequency and full cohort.*** *AUROC, AUPRC, and event rate were calculated for both PICTURE and NEWS at the two institutions, at two levels of granularity. The observation-level performance compares each individual prediction with whether or not a deterioration event occurred within the next 24 hours. The encounter-level performance compares the maximum score during a patient’s general ward stay with whether or not they ever deteriorated while in a general ward. An event happening within 24 hours after discharge from a general ward is still counted as a positive. \* indicates significance vs. NEWS performance via difference in AUC, p < 0.001.*

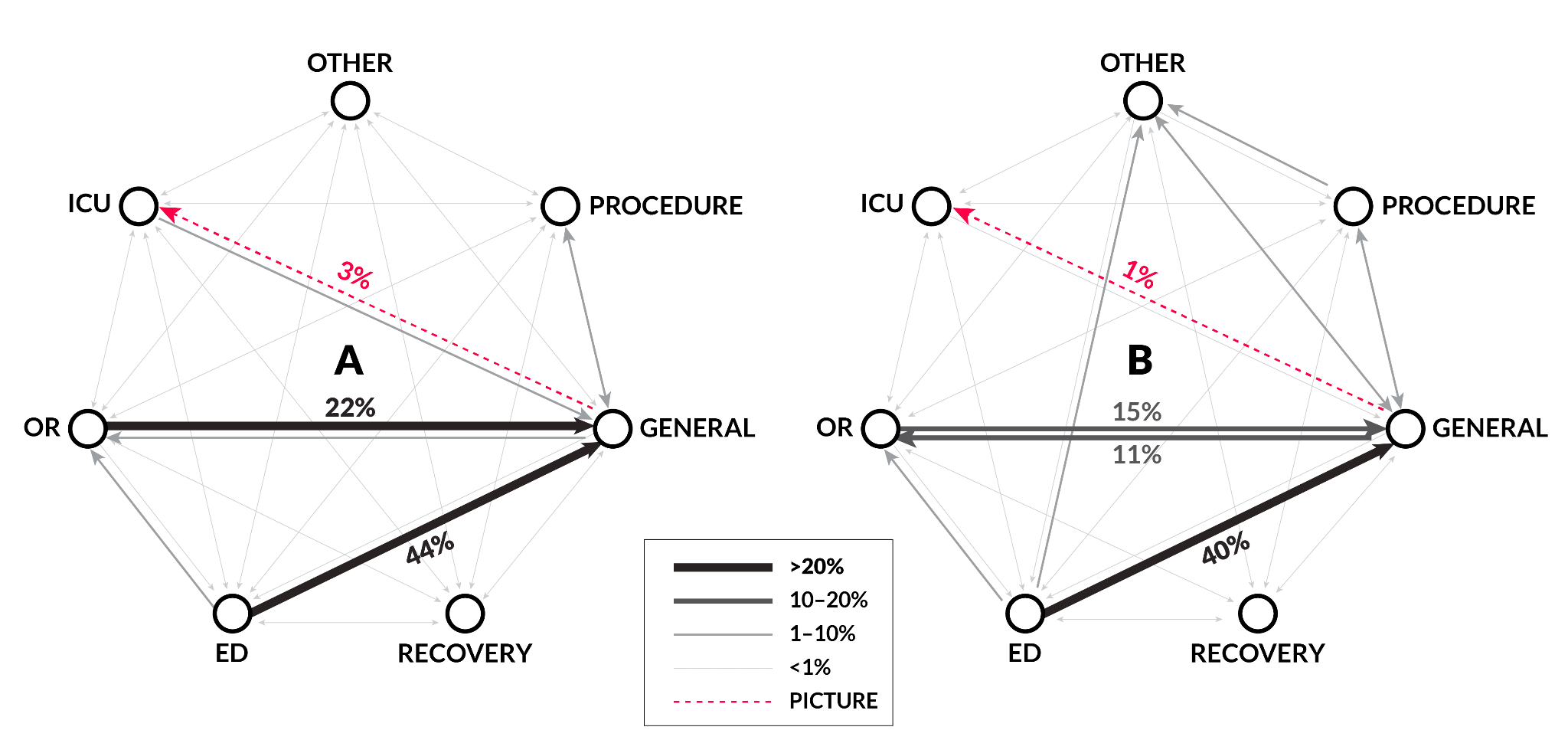
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Granularity** | **Metric** | **PICTURE** | | **NEWS** | |
| **MM** | **HMC** | **MM** | **HMC** |
| **Observation** | **AUROC**  (95% CI) | 0.827\*  (0.825-0.828) | 0.827\*  (0.823-0.831) | 0.771  (0.770-0.773) | 0.773  (0.769-0.778) |
| **AUPRC**  (95% CI) | 0.101\*  (0.099-0.103) | 0.120\*  (0.118-0.121) | 0.064  (0.063-0.066) | 0.092  (0.086-0.097) |
| **Event Rate** | 1.1% | 0.9% | 1.1% | 0.9% |
| **Encounter** | **AUROC**  (95% CI) | 0.876\*  (0.867-0.884) | 0.851\*  (0.837-0.865) | 0.826  (0.817-0.836) | 0.805  (0.790-0.821) |
| **AUPRC**  (95% CI) | 0.321\*  (0.301-0.339) | 0.336\*  (0.306-0.370) | 0.186  (0.172-0.201) | 0.210  (0.183-0.236) |
| **Event Rate** | 4.2% | 2.5% | 4.2% | 2.5% |

**Supplement 9: Performance on individual targets**

As was described previously in Section 2, a composite outcome consisting of death, ICU transfer, or a new need for mechanical ventilation was used to train, validate, and externally test PICTURE. If a patient met any of these outcomes within 24 hours of a given prediction, that observation is labeled as a positive; otherwise it is assigned to the negative class. Predictions occurring 30 minutes before the first event and after are removed. In the table below, the composite outcome is broken down into individual components – that is, death, ICU transfer, or mechanical ventilation – and the performance is assessed on the encounter-level. This means we are comparing patients whose first deterioration was from the individual component (e.g. death) to patients who never deteriorated. Patients who met multiple outcomes (e.g. transferred to the ICU, then died) are categorized under the first deterioration criteria they met, since after this point predictions are no longer made. Consequently, ICU transfer was the most common outcome met, as it typically precedes both mechanical ventilation and death. 95% confidence intervals are computed via 1000-replicate bootstrap. In contrast to Table 3 in the main manuscript, AUPRC is not adjusted for event rate differences, to better reflect differences in each individual outcome component. Due to the extremely low frequency of mechanical ventilation (that is, mechanical ventilation occurring before an ICU transfer) at both institutions (e.g. fewer than 10 individual patients met this criteria at Hospital 2), both AUROC and AUPRC estimates are highly variable for this subset.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Metric** | **Analytic** | **Hospital 1 (MM)** | | | | **Hospital 2 (HMC)** | | | |
| **Composite** | **Death** | **ICU transfer** | **Mech. ventilation** | **Composite** | **Death** | **ICU transfer** | **Mech. ventilation** |
| **Event Rate** | | 4.5% | 0.8% | 3.7% | 0.05% | 2.5% | 0.7% | 1.8% | 0.07% |
| **AUROC**  **(95% CI)** | **PICTURE** | 0.870  (0.861-0.878) | 0.967 (0.959-0.976) | 0.850  (0.840-0861) | 0.700  (0.587-0.830) | 0.875  (0.851-0.902) | 0.933  (0.900-0.976) | 0.851  (0.822-0.884) | 0.926  (0.868-1.00) |
| **EDI** | 0.830  (0.821-0.840) | 0.938 (0.925-0.953) | 0.807  (0.796-0.819) | 0.681  (0.559-0.807) | 0.835  (0.808-0.863) | 0.922  (0.891-0.960) | 0.801  (0.769-0.837) | 0.846  (0.757-0.947) |
| **NEWS** | 0.817  (0.806-0.827) | 0.930  (0.918-0.944) | 0.794  (0.793-0.804) | 0.616  (0.479-0.766) | 0.819  (0.792-0.851) | 0.867  (0.824-0.914) | 0.800  (0.766-0.835) | 0.851  (0.707-1.00) |
| **AUPRC**  **(95% CI)** | **PICTURE** | 0.403  (0.381-0.424) | 0.472  (0.423-0.526) | 0.273  (0.251-0.293) | 0.004  (0.00-0.007) | 0.340  (0.283-0.395) | 0.348  (0.233-0.460) | 0.193  (0.140-0.239) | 0.013  (0.00-0.022) |
| **EDI** | 0.293  (0.273-0.313) | 0.344  (0.296-0.396) | 0.179  (0.163-0.193) | 0.003  (0.00-0.004) | 0.231  (0.182-0.278) | 0.278  (0.174-0.388) | 0.102  (0.072-0.128) | 0.005  (0.00-0.009) |
| **NEWS** | 0.260  (0.242-0.277) | 0.205  (0.168-0.240) | 0.175  (0.160-0.191) | 0.001  (0.00-0.002) | 0.233  (0.181-0.283) | 0.151  (0.062-0.221) | 0.151  (0.100-0.194) | 0.036  (0.00-0.068) |

***Supplement 10: Patient transfers as a directed network graph.*** *A directed network was constructed such that nodes represent a type of hospital unit (e.g. ED, OR, procedure room, etc.) and directed edges represent transfers from one unit to another. Edge weights are proportional to the frequency of the transfer, e.g. ED to general floor transfers were more common than the opposite. This frequency is relative to the number of transfers (e.g. an edge weight of 44% in Panel A represents 44% of all transfers were from the ED to the general floors). Graphs were constructed using the ADT table. The “Other” category includes labor and delivery and other specialized units.* ***A:*** *Patient transfers at the initial institution (MM)* ***B:*** *Patient transfers at the second institution (HMC).*



**A: MM**

**B: HMC**

***Supplement 11: Threshold performance simulation.*** Alert thresholds should be calibrated to their environment. When moving between hospitals, a threshold constructed on the initial hospital can perform very differently when moved outside that institution, especially with respect to changing event rates. To demonstrate this, an example threshold was constructed for each analytic (PICTURE, EDI, and NEWS) corresponding to a sensitivity of 0.5 at the initial institution (MM). That same threshold was then applied directly to data from the second institution (HMC). In all three cases, the threshold does not behave as expected – PPV is higher (indicating fewer false alerts), but this comes at the cost of a lower sensitivity. In a clinical scenario, this would mean that all three models would miss more deteriorations than originally expected given the same threshold. When the threshold is recalibrated at the new institution (i.e. a new threshold is chosen with a sensitivity of 0.5 at the second hospital), the PPV generally returns closer to what’s expected at the same level of sensitivity. It should also be noted that PICTURE resulted in almost half the number of false alerts as measured by WDR when compared to the EDI using thresholds aligned to a sensitivity of 0.5.

\* PPV and workup-to-detection ratio (WDR) are adjusted to the event rate of the second institution (2.5%) to facilitate direct comparison.8

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Analytic** | **Hospital** | **Threshold** | **Sens.** | **Spec.** | **PPV\*** | **WDR\*** | **Event rate** |
| **PICTURE** | **MM** | 0.118 | 0.500 | 0.962 | 0.257 | 3.89 | 4.5% |
| **HMC** | 0.118 | 0.360 | 0.989 | 0.452 | 2.21 | 2.5% |
| **HMC**  Recalibrated | 0.062 | 0.500 | 0.967 | 0.279 | 3.58 | 2.5% |
| **EDI** | **MM** | 59.8 | 0.500 | 0.925 | 0.146 | 6.85 | 4.5% |
| **HMC** | 59.8 | 0.277 | 0.982 | 0.279 | 3.58 | 2.5% |
| **HMC**  Recalibrated | 47.8 | 0.503 | 0.931 | 0.160 | 6.25 | 2.5% |
| **NEWS** | **MM** | 10 | 0.523 | 0.903 | 0.122 | 8.20 | 4.5% |
| **HMC** | 10 | 0.371 | 0.965 | 0.212 | 4.72 | 2.5% |
| **HMC**  Recalibrated | 9 | 0.529 | 0.926 | 0.155 | 6.45 | 2.5% |

Supplemental References

1. 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

2. Royal College of Physicians: National Early Warning Score (NEWS): Standardising the assessment of acute illness severity in the NHS. *Royal College of Physicians, London, 2012*

3. Smith GB, Prytherch DR, Meredith P, et al.: The ability of the National Early Warning Score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death. *Resuscitation* 2013; 84:465–470

4. 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

5. 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

6. 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

7. 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

8. 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 Medical Informatics* 2021; 9:e25066

9. Lundberg SM, Erion G, Chen H, et al.: From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* 2020; 2:56–67