OPEN

# Predicting Future Care Requirements Using Machine Learning for Pediatric Intensive and Routine Care Inpatients

Eduardo A. Trujillo Rivera, PhD[1]
James M. Chamberlain, MD[2]
Anita K. Patel, MD[3]
Qing Zeng-Treitler, PhD[1]
James E. Bost, PhD[4]
Julia A. Heneghan, MD[3]
Hiroki Morizono, PhD[5]
Murray M. Pollack, MD[3]

**OBJECTIVES:** Develop and compare separate prediction models for ICU and non-ICU care for hospitalized children in four future time periods (6–12, 12–18, 18–24, and 24–30 hr) and assess these models in an independent cohort and simulated children's hospital.

**DESIGN:** Predictive modeling used cohorts from the Health Facts database (Cerner Corporation, Kansas City, MO).

**SETTING:** Children hospitalized in ICUs.

**PATIENTS:** Children with greater than or equal to one ICU admission ($n = 20{,}014$) and randomly selected routine care children without ICU admission ($n = 20{,}130$) from 2009 to 2016 were used for model development and validation. An independent 2017–2018 cohort consisted of 80,089 children.

**INTERVENTIONS:** None.

**MEASUREMENT AND MAIN RESULTS:** Initially, we undersampled non-ICU patients for development and comparison of the models. We randomly assigned 64% of patients for training, 8% for validation, and 28% for testing in both clinical groups. Two additional validation cohorts were tested: a simulated children's hospitals and the 2017–2018 cohort. The main outcome was ICU care or non-ICU care in four future time periods based on physiology, therapy, and care intensity. Four independent, sequential, and fully connected neural networks were calibrated to risk of ICU care at each time period. Performance for all models in the test sample were comparable including sensitivity greater than or equal to 0.727, specificity greater than or equal to 0.885, accuracy greater than 0.850, area under the receiver operating characteristic curves greater than or equal to 0.917, and all had excellent calibration (all $R^2$s $> 0.98$). Model performance in the 2017–2018 cohort was sensitivity greater than or equal to 0.545, specificity greater than or equal to 0.972, accuracy greater than or equal to 0.921, area under the receiver operating characteristic curves greater than or equal to 0.946, and $R^2$s greater than or equal to 0.979. Performance metrics were comparable for the simulated children's hospital and for hospitals stratified by teaching status, bed numbers, and geographic location.

**CONCLUSIONS:** Machine learning models using physiology, therapy, and care intensity predicting future care needs had promising performance metrics. Notably, performance metrics were similar as the prediction time periods increased from 6–12 hours to 24–30 hours.

**KEY WORDS:** criticality index; dynamic modeling; machine learning; pediatric intensive care unit; pediatrics; severity of illness

More than 2 million children are hospitalized each year, and more than 200,000 are cared for in ICUs. Approximately 20% of these children are hospitalized in free-standing children's hospitals, in



DOI: 10.1097/CCE.0000000000000505

which 25–50% of the beds are ICU-level care (1–3). Importantly, approximately 20% of PICU patients are transferred from non-ICU care areas, often after clinical deterioration (4). Pediatric inpatients requiring transfer to ICU care are more likely to develop new morbidity and more likely to die than postoperative admissions (2).

It is often difficult for clinicians to predict which patients will respond favorably to medical interventions and which will deteriorate (5). Early identification of patients responding to therapies and those at substantial risk for clinical deterioration could allow for earlier discharge or more aggressive interventions that might alter their clinical course, reduce morbidities, and, in severe cases, prevent death. Predicting future clinical events ideally includes the timing of change based on a consideration of the current physiologic state as a measure of severity of illness, the milieu of therapies and therapeutic intensity, and assessment of the trajectory of these variables. Recently, we validated a new severity measure, the Criticality Index, based on the physiology, therapies, and therapeutic intensity, which accounts for changes in these variables over time. The Criticality Index is calibrated to the probability of receiving ICU care and demonstrated large Criticality Index differences among high-intensity ICU care, ICU care, and routine care (4, 6). Therefore, predicting changes in severity of illness for pediatric inpatients can be operationalized in a single model as predicting the care area.

A major goal of clinical outcome prediction has been to predict changes in severity of illness to identify patients who will either need ICU care, continue their current care needs, or transition out of intensive care with sufficient temporal warning to allow for clinical interventions that might alter the clinical course. This goal can be operationalized in a single model by predicting severity changes at specified, future time periods based on the Criticality Index that use the outcome of ICU or non-ICU (routine) care. There were three goals for this analysis. First, we developed and compared separate machine learning models for prediction of care location (ICU or non-ICU) for hospitalized children in future time periods of 6–12, 12–18, 18–24, and 24–30 hours. This analysis used a research database with a distribution of ICU and non-ICU patients that enhanced model development. Second, we assessed performance in an independent dataset from 2017 to 2018. Third, we assessed potential clinical applicability by assessing performance in a simulated children's hospital, determining the accuracy of predicting ICU admission, and assessing the potential influence of institutional characteristics on model performances. We focused on a 24-hour time frame divided into discrete 6-hour time periods because this time frame and organization could have substantial implications for patient safety, clinical outcomes, and resource utilization.

## MATERIALS AND METHODS

### Sample

The model development dataset was derived from the Health Facts database (Cerner Corporation, Kansas City, MO) that collects comprehensive deidentified clinical data on patient encounters from hospitals in the United States with a Cerner data use agreement. Data are date- and time-stamped including admission and demographic data, laboratory results, medication data derived from pharmacy records, diagnostic and procedure codes, vital signs, respiratory data, and hospital outcome. Cerner Corporation has established HIPAA compliance operating policies to establish deidentification of Health Facts. Not all data are available for all patients. Health Facts has been assessed as representative of the United States (7) and used in previous care assessments including the Acute Physiology and Chronic Health Evaluation score (8) and medication assessments for children in ICUs (9, 10).

Details on preparing data have been published, including data cleaning and data definitions, medications and medication classification, laboratory data, and vital signs and respiratory data (6). Medication data were determined from pharmacy records using start and discontinuation times. Drugs were categorized by Multum (North Kansas City, MO) (11). Diagnoses were categorized based on the *International Classification of Diseases* (ICD), 9th Edition and ICD, 10th Edition classifications (12, 13). The primary diagnosis was used for descriptive purposes but not for modeling because it was determined at discharge.

Inclusion criteria included age less than 22 years (14), laboratory, vital signs, and medication data and care in non-ICU care units or ICUs from January

2009 to June 2016. Exclusion criteria included hospital length of stay greater than 100 days, ICU length of stay greater than 30 days, or care in the neonatal ICU. For model building, we included all patients receiving ICU care and a randomly selected sample of patients receiving only non-ICU care, approximately equal in size to the ICU sample. Therefore, we undersampled the non-ICU patients to enhance modeling.

The hospital course was discretized into consecutive 6-hour time periods because data acquisition for non-ICU care children is relatively infrequent compared with ICU patients. Each time period was categorized into the mutually exclusive categories of ICU care or non-ICU care; we excluded time periods when the patient was in both the non-ICU and the ICU.

## Independent Variables

The variables, definitions, and statistics for each variable used for modeling are shown in **Supplemental Digital Content 1** (http://links.lww.com/CCX/A736). The variables are those in the Criticality Index and consist of six routine vital signs, 30 routinely measured laboratory variables, and parenterally administered medications. The machine learning methods required laboratory and vital sign measurements in each time period, requiring imputation for missing data. Consistent with other machine learning models, we imputed laboratory results and vital signs using the last known result because, in general, physicians use the last measured values and repeat measurements when required for clinical care or when results are acquired routinely (15, 16). If during the first 6-hour time period there were missing values, these values were set to the median of the first 6-hour time periods using nine age groups (4, 6). These imputed values have been reported (4, 6). All were either in the normal range or had only minor deviations from normal. This imputation scheme is similar to other severity scores which assume normal values for unmeasured variables (2, 17), with the improvement that specific estimates for ICU patients are used rather than normal data. The imputed values were identified in the modeling (below) by setting the count equal to zero. The possibility that imputation induced a systematic biased was exploring using pairwise comparison of distributions of laboratory and vital signs with and without imputation (18, 19). No bias was evident.

## Machine Learning Methodology and Statistical Analysis

We randomly assigned 64% of patients for training, 8% for validation, and 28% for testing for the ICU and non-ICU patient groups. This distribution was chosen to maximize the test sample. Random selection was at the patient level. The training set was used for model development, and the validation set was used to fine-tune variables to avoid overfitting. The training and validation sets were combined for calibration of each of the models. The test sample and a 2017–2018 cohort were used to evaluate model performance and calibration.

Independent neural networks calibrated to risk of ICU care were developed for four future times: 6–12, 12–18, 18–24, and 24–30 hours. Therefore, predictions of ICU or non-ICU care were based on a single model for each time period. The models are sequential, and layers are fully connected. Each model had seven hidden dense layers, an output layer with one node, and logistic activation. Inputs for the models included variables of the present and immediate past time period. Our model architecture is the result of sequential efforts to maximize the Mathew Correlation Coefficient (MCC). Initially, models with one hidden layer were considered. We sequentially increased the number of internal nodes in combination with the proportion of dropout nodes. This process along with L2 norm regularization and monitoring MCC values between training and validation sets determined the final number of nodes for the first hidden layer. We stopped increasing the number of nodes when the MCC of the validation and trainings sets converged to a common value. The architecture of this hidden layer was frozen, and additional hidden layers proceeded similarly. We stopped adding hidden layers when they did not significantly increase the MCC of the training and validation sets. Overfitting was avoided by keeping the MCC of the training and validation sets at a difference of no more than 0.05 as well as maintaining the stability of the other performance metrics. Each model was independently calibrated to the respective future risk of ICU care (20). These model outputs predict future care areas of non-ICU and ICU care (4, 6).

The performance of the four models was first assessed in the test sample. Initially, the models were assessed with confusion matrices at the decision cut point of 0.5 (21–23) and areas under the receiver operating

characteristic curves (AUROC) and precision-recall curves (PRAUC) with their 95% CIs (24). The number needed to evaluate is not shown but can be calculated as 1/precision. Accuracy, precision, and negative predictive value for the test sample were assessed for sensitivities (true positive rate) and specificities (true negative rate) of 0.85, 0.90, 0.95, and 0.99 for the approximate lower boundary of the 95% CI. Prediction of true positive indicates the patient is expected to be transferred to the ICU or remain in the ICU for the outcome time period, whereas prediction of a true negative indicates the patient is expected to remain in a non-ICU care area or be transferred out of the ICU to a non-ICU care area. For those patients correctly predicted to be transferred from non-ICU to ICU care, we computed the percentage of those receiving either mechanical ventilation or vasoactive agents within 24 hours of transfer. Second, we assessed the calibration of each model over the full range of risk intervals using the differences between the observed and expected proportions of ICU outcomes within the intervals. The numbers of calibration intervals for the four models were greater than 2,900 (**Fig. 1**, Supplemental Digital Content 3, http://links.lww.com/CCX/A738). We used multiple metrics to assess calibration. We computed the regression line for the predicted proportions for comparison to the ideal and assessed the $R^2$ from the regression lines as measure of tightness around the regression lines. We also computed the differences between observed and predicted ICU proportions within each calibration interval and report the percentage of intervals with no evidence for difference. Third, we assessed the accuracy, precision, and negative predictive value for sensitivities and specificities of 0.85, 0.90, 0.95, and 0.99 for the approximate lower boundary of the 95% CIs.

Since the models were developed in a sample constructed to enhance model development but not assess "real-life" performance, we also assessed performance in an independent January 2017 to June 2018 Health Facts cohort without ICU sample enhancement in a similar manner as the test sample. We also assessed the potential clinical utility in three ways. First, we constructed a simulated children's hospital by random selection from the test sample such that 20% of the total sample were cared for in the ICU and 20% of the ICU patients were initially admitted to non-ICU care areas prior to transfer to the ICU. These population estimates were obtained from previous analyses (3, 4) and a query of Children's Hospital Association database (MM Pollack, unpublished data, 2020). In addition, we created three additional sets of randomly selected test patients with prevalences of 10%, 15%, and 30% for the ICU patients. Second, since the most valuable potential utility is the prediction of transfer from non-ICU to ICU care, we assessed the accuracy of each model in patients who changed their care areas from non-ICU to ICU care in both the test sample and independent cohort. The accuracy was assessed if any of the prediction models were correct. The first 6-hour time period after transfer into the ICU had predictions from all four models, the second 6-hour period had predictions from three models, the third 6-hour period had predictions from two models, and the fourth period had predictions from one model. We also assessed the accuracy for the first, second, third, and fourth 6-hour time periods after transfer but only when the prediction was done prior to the transfer. Finally, we assessed the influence of institutional characteristics on model performances in institutions with different characteristics including teaching, geographical region, and hospital bed size determined from the Health Facts database.

## RESULTS

There were 20,014 patients with an ICU stay and 20,130 patients cared for in non-ICU care areas only in the 2009–2016 test sample. Demographic data are in **Table 1**. Details of this sample have been previously published (4, 6). Compared with patients with ICU stays, non-ICU care patients were older (median 132.2 vs 28.0 mo; $p < 0.0001$), had shorter median hospital stays (71 vs 110 hr; $p < 0.0001$), and had a lower mortality rate (0.1% vs 3.2%; $p < 0.0001$). Most diagnostic categories differed between the two groups ($p < 0.0001$). The numbers of patients and 6-hour time periods in the ICU and non-ICU care locations in the training, validation, and testing samples for each of the prediction models are shown in **Supplemental Digital Content 2** (http://links.lww.com/CCX/A737). Overall, there were greater than 325,000 6-hour time periods for each future time period.

The performances of all four models predicting the future care location were similar in the test sample (**Table 2*A***). At a decision threshold of 0.5, the sensitivity for the 6–12-hour time period was 0.797 and decreased

## TABLE 1.
## Population Characteristics

| Characteristics | All | ICU | Non-ICU | Transfers Non-ICU to ICU | Transfers ICU to Non-ICU | Simulated Children's Hospital | 2017–2018 Pediatric Inpatients |
|---|---|---|---|---|---|---|---|
| *n* | 40,144 | 20,014 | 20,130 | 6,181 | 15,336 | 7,054 | 80,089 |
| Age (mo), median (25–75th percentile) | 96 (16–201) | 28 (0–188) | 132 (53–209) | 109 (8–210) | 26 (0–186) | 120 (37–207) | 107 (11–215) |
| Female, *n* (%) | 19,599 (48.8) | 8,913 (44.5) | 10,686 (53.1) | 2,851 (46.1) | 6,804 (44.4) | 3,600 (51.0) | 43,512 (54.3) |
| Race, *n* (%) | | | | | | | |
| Black | 9,354 (23.3) | 5,256 (26.3) | 4,098 (20.4) | 1,659 (26.8) | 4,032 (26.3) | 1,531 (21.7) | 17,565 (21.9) |
| Caucasian | 19,048 (47.5) | 10,335 (51.6) | 8,713 (43.3) | 3,247 (52.5) | 7,942 (51.8) | 3,195 (45.3) | 46,139 (57.6) |
| Other–unknown | 11,742 (29.3) | 4,423 (22.1) | 7,319 (36.4) | 1,275 (20.6) | 3,362 (21.9) | 2,328 (33.0) | 16,385 (20.5) |
| Hospital LOS (hr), median (25–75th percentile) | 87 (47–174) | 110 (55–237) | 71 (44–125) | 121 (65–236) | 118 (60–247) | 74 (46–140) | 59 (41–96) |
| ICU LOS (hr), median (25–75th percentile) | 0 (0–74) | 75 (33–169) | 0 (0–0) | 65 (30–138) | 73 (33–171) | 0 (0–0) | 0 (0–0) |
| Hospital mortality, *n* (%) | 657 (1.6) | 631(3.2) | 26 (0.1) | 91 (1.5) | 384 (2.5) | 57 (0.8) | 238 (0.3) |
| Positive pressure ventilation, *n* (%)[a] | 6,131 (15.3) | 5,313 (26.6) | 818 (4.1) | 1,281 (20.7) | 4,037 (26.3) | 604 (8.6) | 1,128 (1.41) |
| Diagnostic group, *n* (%) | | | | | | | |
| Respiratory | 3,614 (13.8) | 1,563 (14.3) | 2,051 (13.5) | 529 (15.3) | 1,220 (14.3) | 691 (13.6) | 10,572 (13.2) |
| Endocrine, nutritional, metabolic, and immune disorders | 3,158 (12.1) | 1,624 (14.8) | 1,534 (10.1) | 325 (9.4) | 1,255 (14.7) | 584 (11.5) | 8,810 (11.0) |
| Gastrointestinal | 2,556 (9.8) | 503 (4.6) | 2,053 (13.5) | 215 (6.2) | 403 (4.7) | 591 (11.6) | 7,608 (9.5) |
| Infectious and parasitic | 2,377 (9.1) | 962 (8.8) | 1,415 (9.3) | 368 (10.7) | 759 (8.9) | 458 (9.0) | 6,327 (7.9) |
| Injury and poisoning | 2,261 (8.6) | 1,390 (12.7) | 871 (5.7) | 491 (14.2) | 1,070 (12.5) | 349 (6.6) | 5,766 (7.2) |
| Neurologic | 1,856 (7.1) | 788 (7.2) | 1,068 (7.0) | 293 (8.5) | 634 (7.4) | 345 (6.8) | 7,208 (9.0) |
| Neoplasms | 1,636 (6.3) | 231 (2.1) | 1,405 (9.2) | 137 (4.0) | 179 (2.1) | 439 (8.6) | 8,249 (10.3) |
| Circulatory | 1,196 (4.6) | 735 (6.7) | 461(3.0) | 264 (7.6) | 548 (6.4) | 197 (3.9) | 3,043 (3.8) |
| Not otherwise specified/other[b] | 7,527 (18.6) | 3,147 (15.7) | 4,380 (21.8) | 834 (13.5) | 2,480 (16.2) | 1,440 (20.4) | 17,700 (22.1) |

LOS = length of stay.

[a]Criteria for positive pressure were continuous positive airway pressure, positive end-expiratory pressure, peak inspiratory pressure.

[b]Other = genital-urinary, musculoskeletal, mental, skin and subcutaneous tissue, and hematology disorders.

The population ($n = 40,144$) includes all patients used in any of the four models and the special samples. The demographics for the ICU patients ($n = 20,014$) and non-ICU patients ($n = 20,130$) are shown in the second and third columns. ICU patients had a stay at any time during their hospitalization. ICU and non-ICU patients were statistically different ($p < 0.001$) except for respiratory system conditions ($p = 0.059$), infectious and parasitic diseases ($p = 0.176$), and nervous system conditions ($p = 0.567$). Patients in the two transfer groups (non-ICU to ICU: $n = 6,181$; ICU to non-ICU: $n = 15,336$) and the simulated children's hospital sample were not mutually exclusive and were not compared statistically. The transfer groups pertains to analyses reported in **Table 4** (Supplemental Digital Content 5, http://links.lww.com/CCX/A740). The simulated children's hospital sample ($n = 7,054$) was randomly selected such that 20% of patients were ICU patients and 20% of the ICU patients were initially admitted to non-ICU areas prior to transfer to the ICU. The 2017–2018 pediatric inpatients dataset ($n = 80,089$) is an independent Health Facts cohort without ICU patient enhancement with 5,561 ICU admissions (6.94%).

## TABLE 2.

**Performance Metrics for the Test Sample, Independent 2017–2018 Cohort, and Simulated Children's Hospital and at a Decision Threshold of 0.5 for Prediction of ICU Care During the Future Time Period**

| Metrics | 6–12 hr | 12–18 hr | 18–24 hr | 24–30 hr |
|---|---|---|---|---|
| A. Test sample 2009–2016[a] | | | | |
| AUROC | 0.917 (0.916–0.917) | 0.917 (0.916–0.917) | 0.919 (0.918–0.919) | 0.920 (0.919–0.920) |
| AUPRC | 0.867 (0.865–0.869) | 0.867 (0.865–0.869) | 0.866 (0.864–0.868) | 0.864 (0.862–0.866) |
| Sensitivity | 0.797 (0.795–0.799) | 0.780 (0.777–0.782) | 0.749 (0.746–0.751) | 0.727 (0.724–0.73) |
| Specificity | 0.885 (0.883–0.886) | 0.896 (0.895–0.898) | 0.907 (0.906–0.909) | 0.916 (0.915–0.918) |
| Precision (PPV) | 0.783 (0.781–0.786) | 0.797 (0.795–0.799) | 0.809 (0.806–0.811) | 0.820 (0.817–0.822) |
| Negative prediction value | 0.893 (0.892–0.894) | 0.886 (0.885–0.888) | 0.874 (0.872–0.875) | 0.865 (0.864–0.867) |
| Accuracy[b] | 0.855 (0.853–0.856) | 0.856 (0.855–0.857) | 0.853 (0.852–0.854) | 0.851 (0.850–0.853) |
| F1 score[c] | 0.790 (0.788–0.792) | 0.788 (0.787–0.790) | 0.778 (0.776–0.779) | 0.771 (0.769–0.772) |
| False discovery rate | 0.217 (0.214–0.219) | 0.203 (0.201–0.205) | 0.191 (0.189–0.194) | 0.180 (0.178–0.183) |
| B. Independent cohort January 2017 to June 2018[a] | | | | |
| AUROC | 0.948 (0.947–0.948) | 0.948 (0.948–0.949) | 0.946 (0.945–0.946) | 0.947 (0.946–0.947) |
| AUPRC | 0.726 (0.725–0.728) | 0.742 (0.740–0.743) | 0.736 (0.734–0.737) | 0.741 (0.739–0.742) |
| Sensitivity | 0.590 (0.589–0.591) | 0.545 (0.544–0.546) | 0.570 (0.569–0.571) | 0.571 (0.569–0.572) |
| Specificity | 0.972 (0.972–0.972) | 0.977 (0.977–0.977) | 0.973 (0.973–0.973) | 0.973 (0.973–0.973) |
| Precision (PPV) | 0.748 (0.747–0.749) | 0.771 (0.770–0.772) | 0.757 (0.756–0.758) | 0.760 (0.759–0.762) |
| Negative prediction value | 0.944 (0.944–0.944) | 0.937 (0.937–0.937) | 0.939 (0.939–0.939) | 0.938 (0.938–0.939) |
| Accuracy[b] | 0.925 (0.925–0.925) | 0.923 (0.922–0.923) | 0.922 (0.921–0.922) | 0.921 (0.921–0.921) |
| F1 score[c] | 0.660 (0.659–0.661) | 0.639 (0.638–0.640) | 0.650 (0.649–0.651) | 0.652 (0.651–0.653) |
| False discovery rate | 0.252 (0.251–0.253) | 0.229 (0.228–0.230) | 0.243 (0.242–0.244) | 0.240 (0.238–0.241) |
| C. Simulated children's hospital[a] | | | | |
| AUROC | 0.971 (0.970–0.971) | 0.969 (0.968–0.969) | 0.966 (0.965–0.966) | 0.967 (0.966–0.967) |
| AUPRC | 0.857 (0.853–0.862) | 0.843 (0.838–0.847) | 0.827 (0.823–0.832) | 0.816 (0.811–0.821) |
| Sensitivity | 0.833 (0.829–0.837) | 0.829 (0.825–0.834) | 0.797 (0.792–0.802) | 0.791 (0.786–0.796) |
| Specificity | 0.952 (0.951–0.953) | 0.951 (0.950–0.952) | 0.956 (0.955–0.957) | 0.956 (0.955–0.957) |
| Precision (PPV) | 0.740 (0.735–0.745) | 0.724 (0.719–0.729) | 0.728 (0.723–0.733) | 0.719 (0.713–0.724) |
| Negative prediction value | 0.972 (0.971–0.973) | 0.973 (0.972–0.974) | 0.969 (0.969–0.970) | 0.970 (0.969–0.971) |
| Accuracy[b] | 0.935 (0.934–0.937) | 0.934 (0.933–0.935) | 0.935 (0.934–0.936) | 0.936 (0.934–0.937) |
| F1 score[c] | 0.784 (0.780–0.787) | 0.773 (0.770–0.777) | 0.761 (0.757–0.765) | 0.753 (0.749–0.757) |
| False discovery rate | 0.260 (0.255–0.265) | 0.276 (0.271–0.281) | 0.272 (0.267–0.277) | 0.281 (0.276–0.287) |

AUPRC = area under the precision-recall curve, AUROC = area under the receiver operating characteristic curve, PPV = positive predictive value.

[a]Data in parenthesis are the 95% CIs.

[b]Accuracy = (true positives + true negatives)/(positives + negatives).

[c]The F1 score is a measure of accuracy with a maximum score of 1. It is the harmonic mean of precision and sensitivity.

The population demographics are shown in Table 1. The following components of the confusion matrix are not shown since they can be computed from other data: false-negative rate = (1−sensitivity), false-positive rate = (1−specificity), false omission rate = (1−negative predictive value), and number needed to evaluate = 1/precision.

with lengthening prediction time intervals to 0.727 for the 24–30-hour time period. Specificity for the 6–12-hour time period was 0.885 and increased to 0.917 for the 24–30-hour time period. Accuracy was greater than 0.85 for all time periods. AUROCs were all 0.92 (Fig. 1, Supplemental Digital Content 3, http://links.lww.com/CCX/A738). The PRAUCs for the four time periods were similar (**Fig. 2**, Supplemental Digital Content 3, http://links.lww.com/CCX/A738) with the PRAUCs ranging from 0.864 to 0.867. The calibration plots (**Fig. 3**, Supplemental Digital Content 3, http://links.lww.com/CCX/A738) show the observed and expected proportions of patient in each of the risk intervals were closely matched. The regression lines shown in each plot have very small constants (range, –0.02 to –0.03), the slopes are very close to identity (range, 1.05–1.06), and the $R^2$s from the regressions between the observed and expected proportions were all greater than 0.98. The percent of the calibration intervals with the 95% CIs crossing zero ranged from 93.96% to 95.49%. There was a small tendency in all models to underpredict ICU care in the lower risk ranges consistent with care of stable patients in the ICU receiving primarily monitoring and a smaller tendency to overpredict in the middle and upper ranges consistent with some sicker patients being cared for in non-ICU care areas.

The accuracy, precision, and negative predictive value for the whole test sample were assessed for sensitivities and specificities of 0.85, 0.90, 0.95, and 0.99 (**Table 1**, Supplemental Digital Content 3, http://links.lww.com/CCX/A738). Overall, the performance metrics did not significantly decrease as the prediction time interval increased. The precisions decreased as the sensitivity increased from greater than 0.73 for a sensitivity of 0.85 to greater than 0.47 for a sensitivity of 0.99. Accuracies for the sensitivities in ascending order for the four prediction models were greater than 0.84, greater than 0.82, greater than 0.77, and greater than 0.61. The assessment of negative predictive value and accuracy for specificities of 0.85, 0.90, 0.95, and 0.99 was similar with negative predictive values of greater than 0.91, greater than 0.87, greater than 0.82, and greater than 0.73 for the prediction models. All accuracies were greater than 0.76.

The performance of the models was also assessed in the independent 2017–2018 Health Facts dataset. Demographic data are shown in Table 1, the number of time periods are shown in Supplemental Digital Content 2 (http://links.lww.com/CCX/A737), and performance data are shown in **Table 2*B***. Overall, the performance of the models slightly decreased compared with the test sample for AUPRC (0.867–0.726), sensitivity (0.797–0.590), precision (0.783–0.748), F1 score (0.790–0.660), and false discovery rate (0.217–0.252) and increased for accuracy (0.855–0.925), AUROC (0.917–0.948), specificity (0.885–0.972), and negative predictive value (0.893–0.944). The calibration plots are shown in **Figure 1**. The regression lines have very small constants (range, –0.01 to –0.02), the slopes are close to identity (all 0.94), and the $R^2$s all are 0.98. The AUPRCs are shown in **Figure 2** with all values greater than 0.725. The AUROC curves are all greater than 0.945 (**Fig. 1**, Supplemental Digital Content 4, http://links.lww.com/CCX/A739).

The accuracy, precision, and negative predictive value for the independent cohort were also assessed for sensitivities and specificities of 0.85, and 0.95 for the lower boundary of the 95% CI (**Table 3**). Overall, performance metrics were stable as the prediction time interval increased. For a sensitivity of 0.95, accuracies for the four models varied from 0.780 to 0.800, precisions ranged from 0.362 to 0.377, specificities ranged from 0.755 to 0.779, and negative predictive values ranged from 0.990 to 0.991. For a specificity of 0.95 for the four models, accuracies for the four models varied from 0.921 to 0.924, precisions ranged from 0.676 to 0.684, sensitivities ranged from 0.722 to 0.733, and negative predictive values ranged from 0.958 to 0.962. The number needed to evaluate (= 1/precision) was always less than three patients for a sensitivity and specificity of 0.95 (data for sensitivities and specificities of 0.85, 0.90, 0.95, and 0.99 are shown in **Table 1**, Supplemental Digital Content 4, http://links.lww.com/CCX/A739).

Potential clinical utility was first assessed in a simulated children's hospital sample with varying prevalences of ICU patients. Demographic data are shown in Table 1, and performance data are shown in **Table 2*C*** for the sample composed of 20% ICU patients. Overall for the sample with 20% ICU patients, there were improvements in AUROC, sensitivity, specificity, negative predictive value, and accuracy and small decreases in AUPRC, precision, and false discovery rate. Changing the percent of ICU patients (**Table 1**, Supplemental Digital Content 5, http://links.lww.com/CCX/A740) changed the performance metrics by

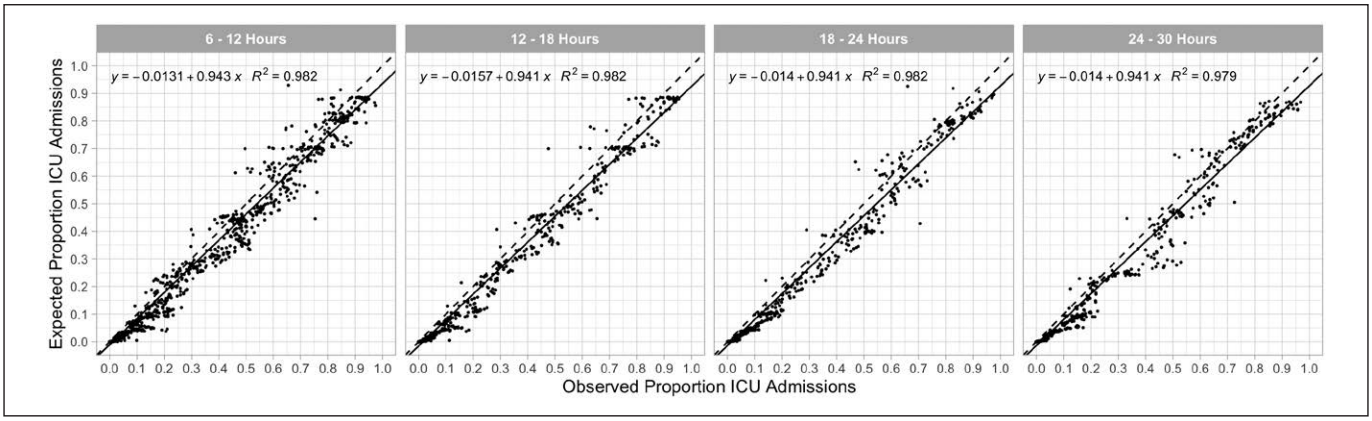


**Figure 1.** Model calibration for the 2017–2018 independent cohort. The *y*-axis shows the expected proportion of ICU care areas for the time periods based on the risk intervals, and the *x*-axis shows the observed proportion of ICU care areas in the time periods. The line of identity is the *dashed line*. The limits of the risk intervals were initially defined by 1,100 equally spaced empirical quantiles. Some intervals were combined to have a larger number of risk values but ensuring 99% of the intervals were smaller than 0.007 resulting in all intervals containing at least 1,770 time periods and 1,063, 1,026, 1,034, and 1,028 risk intervals in the four models. Within each interval, we compute the average expected risk of ICU admission and the observed risk of ICU admission. The *circles* indicate the observed and the expected proportions of ICU 6-hr time periods over ascending Criticality Index intervals. 94.26%, 89.01%, 90.81%, and 90.64% of the risk intervals within each plot have a Cohen's h value less than 0.2, implying there are mostly small effect size differences between the observed and expected proportions. A linear regression is reported in each panel with their respective $R^2$, and the fitted mean is represented with the *solid line*. Calibration was accomplished for each model by using their output for B-spline polynomials as covariates in a linear logistic regression with outcome ICU/regular care for the respective future time. This calibration method is similar to the Platt scaling method for support vector machines (20).

small amounts with all AUROCs greater than 0.95, all PRAUCs greater than 0.80, all sensitivities greater than 0.8, all specificities greater than 0.93, and all accuracies greater than 0.90. Second, we assessed the accuracies of transfers from non-ICU to ICU care if any of the models correctly predicted the transfer, in both the test sample and the 2017–2018 cohort (**Table 2**, Supplemental Digital Content 5, http://links.lww.com/CCX/A740). The percentages for correct predictions were similar in both samples. For example, for sensitivities of 0.95, the percentage of correct ICU care predictions in the first 6-hour period was 92.7% and 88.7% for the test sample and 2017–2018 cohort and ranged from 88.7% to 95.8% for the four models in the two samples. The timing of correct prediction by one or more predictors in the first four 6-hour time periods after ICU admission is shown in **Table 3** (Supplemental Digital Content 5, http://links.lww.com/CCX/A740). For a 0.95 sensitivity, all accuracies were greater than or equal to 0.85 in the test sample and greater than or

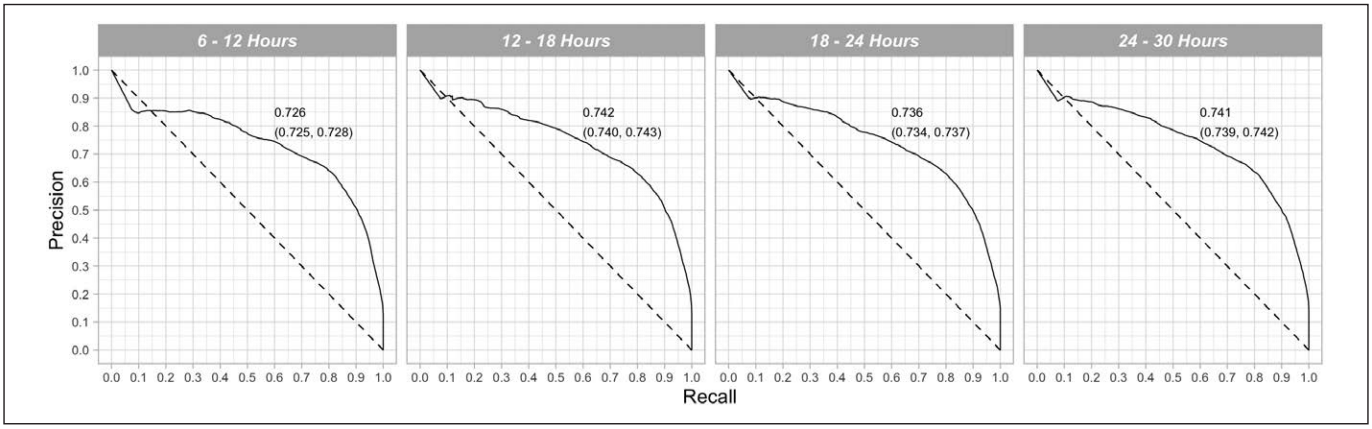


**Figure 2.** Precision-recall curves for the 2017–2018 independent cohort classifying care as routine or ICU for the respective future time periods. The area under the precision-recall curves and 95% CIs are included in each panel. The areas were computed with integral approximations, and the CIs were computed using a logit method (24).

equal to 0.74 in the 2017–2018 cohort. The best performing models in the 2017–2018 models for correct prediction for ICU transfer for the first 6-hour ICU time period were the 12–18-hour model (89.6%) and the 6–12-hour model (84.8%). Finally, the performance metrics assessed by the hospital characteristics of bed size, geographic region, and teaching status are shown in Table 3 (Supplemental Digital Content 5, http://links.lww.com/CCX/A740). The maximum reductions compared with the test sample (Table 2*A*) were less than 11%, whereas most were equivalent or better. Hospitals greater than 500 beds, those in the northeast, and teaching hospitals had the lowest performance metrics, and there was sometimes a small decrease in performance as the prediction time period increased.

## DISCUSSION

Identification of patients' future care needs as ICU or non-ICU care is an estimate of changing severity of illness and may identify patients who will have increased, decreased, or stable care requirements. The Criticality Index which demonstrates large differences among high-intensity ICU care, ICU care, and non-ICU care is an appropriate framework to predict changes in severity of illness as reflected in care needs. This analysis focused on neural network models predicting future care needs in time periods ranging from 6–12 hours to 24–30 hours and evaluated their potential clinical applicability in a simulated children's hospital and in an independent cohort without ICU patient enhancement. In the independent 2017–2018 cohort with a decision cut point of 0.5, all models predicting the need for ICU care had an AUROC greater than 0.945, AUPRC greater than 0.72, and accuracy greater than 0.92, and all had excellent calibration. Notably, the performances in the different prediction time periods were very similar with only small decrements in performance as the prediction time increased for some performance metrics. Altering the decision cut points changed the performance metrics, and this is illustrated for sensitivities and specificities of 0.85, 0.90, 0.95, and 0.99. The stability of model performance across time could be explained by the relative infrequency of changes in care area. We evaluated this possibility by computing accuracies for transfers to the ICU, and the accuracy of predictions of ICU care was greater than 88% with a sensitivity of 0.95. In addition, the positive predictive value of these patients needing vasoactive agent infusions or mechanical ventilation if correctly predicted was 37–38%, at least as good as the performance of the Pediatric Early Warning Score (PEWS) paired with clinical assessment (25). We assessed potential "real-world" performance both in a simulated children's hospital sample and an independent cohort without an enhanced ICU sample from 2017 to 2018, and the performances were comparable with the test sample including the precision indicating a number needed to evaluate of less than 2. Additionally, hospital characteristics had only minor influences on the performance metrics. These results indicate the methodology is appropriate for validation and optimization in a clinical environment.

Risk scores are evolving from those generally directed at identifying patients with high risk of death to those that predict clinical deterioration (26–29). Relatively simple models, such as the PEWS, predominantly use vital signs to derive immediately actionable information (30, 31). Although in widespread use, they generally require large "numbers needed to evaluate" (i.e., high false-positive rates) to achieve reasonable sensitivity and did not improved hospital outcome when tested in a large effectiveness study (32). In the 2017–2018 independent cohort, the number needed to evaluate for a sensitivity of 0.95 was less than or equal to three patients for all models.

Accurate predictions could provide major benefits in assisting clinician decision-making (33–35). We operationalized improving or deteriorating severity of illness as changes in care area, enabling the prediction of ICU or non-ICU care within the same model. Although none of the current risk assessment or prediction methods have significantly enhanced the ability of bedside caregivers to recognize early patterns of deterioration (36–38), the methodology described in this article has the potential to identify patients who will require future transfer to ICU care, potentially altering the clinical trajectory and improving hospital outcomes, patients who will be ready to transition to non-ICU care from ICU care, and those with stable care needs. However, these predictions have different clinical utilities. An alert that a patient may need ICU care usually results in an immediate clinical assessment, often by a rapid response team. Patients predicted to transition out of intensive

## TABLE 3.

**Analysis of the 2017–2018 Cohort for Accuracy, Precision, and Negative Predictive Value at Sensitivities and Specificities of 0.85, and 0.95, at the Lower Boundary of the 95% CI**

| Sensitivities[a] | | | | | |
|---|---|---|---|---|---|
| Prediction Time Period, hr | Sensitivity[b] | Accuracy | Precision[d] | Specificity[c] | Negative Predictive Value |
| 6–12 | 0.851 (0.850–0.852) | 0.906 (0.905–0.906) | 0.580 (0.579–0.582) | 0.913 (0.913–0.914) | 0.978 (0.977–0.978) |
| 6–12 | 0.951 (0.950–0.951) | 0.800 (0.800–0.801) | 0.377 (0.376–0.378) | 0.779 (0.779–0.780) | 0.991 (0.991–0.991) |
| 12–18 | 0.851 (0.850–0.852) | 0.904 (0.904–0.904) | 0.580 (0.579–0.582) | 0.912 (0.911–0.912) | 0.977 (0.977–0.977) |
| 12–18 | 0.951 (0.950–0.952) | 0.794 (0.793–0.794) | 0.374 (0.373–0.375) | 0.771 (0.771–0.772) | 0.991 (0.991–0.991) |
| 18–24 | 0.851 (0.850–0.852) | 0.900 (0.900–0.900) | 0.573 (0.572–0.574) | 0.907 (0.907–0.907) | 0.977 (0.976–0.977) |
| 18–24 | 0.951 (0.950–0.951) | 0.780 (0.779–0.780) | 0.362 (0.361–0.363) | 0.755 (0.754–0.755) | 0.991 (0.990–0.991) |
| 24–30 | 0.851 (0.850–0.852) | 0.899 (0.899–0.899) | 0.575 (0.574–0.576) | 0.906 (0.906–0.907) | 0.976 (0.976–0.976) |
| 24–30 | 0.951 (0.950–0.951) | 0.785 (0.784–0.785) | 0.371 (0.370–0.372) | 0.760 (0.759–0.760) | 0.990 (0.990–0.991) |
| **Specificitiesa** | | | | | |
| Prediction Time Period–hr | Specificityc | Accuracy | Precisiond | Sensitivityb | Negative Predictive Value |
| 6–12 | 0.853 (0.853–0.853) | 0.861 (0.861–0.862) | 0.469 (0.468–0.470) | 0.921 (0.920–0.921) | 0.987 (0.987–0.987) |
| 6–12 | 0.950 (0.950–0.951) | 0.924 (0.923–0.924) | 0.676 (0.674–0.677) | 0.736 (0.734–0.737) | 0.962 (0.962–0.963) |
| 12–18 | 0.850 (0.850–0.851) | 0.987 (0.987–0.987) | 0.469 (0.468–0.470) | 0.920 (0.919–0.921) | 0.987 (0.987–0.987) |
| 12–18 | 0.950 (0.950–0.951) | 0.922 (0.921–0.922) | 0.677 (0.676–0.679) | 0.726 (0.724–0.727) | 0.960 (0.960–0.960) |
| 18–24 | 0.851 (0.850–0.851) | 0.858 (0.858–0.859) | 0.472 (0.471–0.473) | 0.912 (0.911–0.913) | 0.985 (0.985–0.985) |
| 18–24 | 0.950 (0.950–0.951) | 0.921 (0.921–0.921) | 0.680 (0.679–0.682) | 0.722 (0.720–0.723) | 0.959 (0.959–0.959) |
| 24–30 | 0.851 (0.850–0.851) | 0.859 (0.858–0.859) | 0.477 (0.476–0.478) | 0.915 (0.914–0.916) | 0.985 (0.985–0.985) |
| 24–30 | 0.950 (0.950–0.951) | 0.921 (0.920–0.921) | 0.684 (0.682–0.685) | 0.722 (0.720–0.723) | 0.958 (0.958–0.958) |

[a]Lower bound of the 95% CI.

[b]Sensitivity = true positive rate = ICU care time periods.

[c]Specificity = true negative rate = non-ICU care time periods.

[d]Precision = positive predictive value = true positives (cared for in the ICU)/(true positive + false positive). Number needed to evaluate = 1/precision.

The identification of true positives (ICU care time periods) is most relevant to identifying those patients expected to transfer to the ICU from non-ICU care areas or remain in the ICU. The identification of true negatives (non-ICU care time periods) is most relevant to identifying those patients not expected to transfer to the ICU or transfer from the ICU to non-ICU care areas. The data shown are the estimates and 95% CIs.

Data for sensitivities and specificities of 0.85, 0.90, 0.95, and 0.99 are shown in Table 1 (Supplemental Digital Content 4, http://links.lww.com/CCX/A739).

care, however, do not need immediate evaluation, the transfer is often influenced by administrative and organization factors, and therefore, models are expected to have lesser performance.

These models, based on the Criticality Index, integrate past and current physiologic data, therapeutic data, and therapeutic intensity. This is conceptually consistent with historically important ICU severity

advances (39–41). If the performance is validated with real-life data and if the methodology has sufficient face validity for providers, it could improve clinical decision-making by supplementing the limitations of cognitive processing and reducing medical errors (42–46). Medical errors are often based in heuristics and are more likely to occur in high-pressure, high-stakes decisions, particularly when dealing with incomplete information, such as assessing a deteriorating patient or determining the need for ICU care (47–49).

This study has several limitations. First, the database did not contain the full spectrum of data available in the electronic health record (EHR), and therefore, these results might be further optimized. Second, potential clinical applicability needs to be confirmed using real-time EHRs and, when possible, models specific to individual hospitals. Our assessment of clinical applicability using a simulated children's hospital and independent cohort justifies optimism for successful clinical application. Third, we used time periods of 6 hours; shorter time periods might allow better predictive models. Fourth, although machine learning methods have the advantage of measuring intrinsically complicated interactions, the deep neural network models are not transparent, making the clinical importance of individual or sets of variables difficult to ascertain (50, 51).

## CONCLUSIONS

Machine learning models, based on laboratory, vital sign, and medication data predicting future care needs of 6–12, 12–18, 18–24, and 24–30 hours based on the Criticality Index, had promising performance metrics. The performances in all time periods were similar without a significant drop-off as the prediction time period increased, and we demonstrated the models for different times were not simply predicting lack of change since they were able to predict care area changes. This conceptual framework and modeling method are applicable to assessing future care needs represented by care areas, including early detection of major changes in care needs and potentially identifying patients who would benefit from early clinical interventions.

*1 George Washington University School of Medicine and Health Sciences, Washington, DC.*

*2 Department of Pediatrics, Division of Emergency Medicine, Children's National Hospital and George Washington University School of Medicine and Health Sciences, Washington, DC.*

*3 Department of Pediatrics, Division of Critical Care Medicine, Children's National Hospital and George Washington University School of Medicine and Health Sciences, Washington, DC.*

*4 Children's National Hospital and George Washington University School of Medicine and Health Sciences, Washington, DC.*

*5 Children's National Research Institute, George Washington University School of Medicine and Health Sciences, Washington, DC.*

*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/ccejournal).*

*All authors made a substantial contribution to the concept and design of the work, reviewed and critically revised the work for intellectual content, approved the final version, and agreed to be accountable for all aspects of the work in accordance with the International Committee of Medical Journal Editors guidelines. Drs. Trujillo Rivera, Patel, and Morizono were the primary contributors to data preparation. Analysis was under the leadership of Dr. Trujillo Rivera. The first draft of the article was by Drs. Trujillo Rivera and Pollack.*

*Supported, in part, by philanthropic support from Mallinckrodt LLC and award numbers UL1TR001876 from the National Institutes of Health (NIH) National Center for Advancing Translational Sciences and KL2TR001877 from the NIH National Center for Advancing Translational Sciences (to Dr. Patel).*

*Current affiliation for Dr. Heneghan: Department of Pediatrics, Division of Critical Care Medicine, University of Minnesota Masonic Children's Hospital, Minneapolis, MN.*

*The authors have disclosed that they do not have any conflicts of interest.*

*For information regarding this article, E-mail: mpollack@childrensnational.org*

*Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the National Center for Advancing Translational Sciences or the National Institutes of Health.*

*The data extraction was done in structured query language and R with custom code. The data preparation and exploration, model development and evaluation, generation of tables, plots, and results were done in R with custom code. Code for specific tasks is available upon request.*

## REFERENCES

1. Leyenaar JK, Ralston SL, Shieh MS, et al: Epidemiology of pediatric hospitalizations at general hospitals and freestanding children's hospitals in the United States. *J Hosp Med* 2016; 11:743–749
2. Pollack MM, Holubkov R, Funai T, et al; Eunice Kennedy Shriver National Institute of Child Health and Human Development Collaborative Pediatric Critical Care Research Network: Simultaneous prediction of new morbidity, mortality, and

survival without new morbidity from pediatric intensive care: A new paradigm for outcomes assessment. *Crit Care Med* 2015; 43:1699–1709

3. Pelletier JH, Rakkar J, Au AK, et al: Trends in US pediatric hospital admissions in 2020 compared with the decade before the COVID-19 pandemic. *JAMA Netw Open* 2021; 4:e2037227
4. Rivera EAT, Patel AK, Zeng-Treitler Q, et al: Severity trajectories of pediatric inpatients using the criticality index. *Pediatr Crit Care Med* 2021; 22:e19–e32
5. Klein Klouwenberg PMC, Spitoni C, van der Poll T, et al; MARS consortium: Predicting the clinical trajectory in critically ill patients with sepsis: A cohort study. *Crit Care* 2019; 23:408
6. Rivera EAT, Patel AK, Chamberlain JM, et al: Criticality: A new concept of severity of illness for hospitalized children. *Pediatr Crit Care Med* 2021; 22:e33–e43
7. DeShazo JP, Hoffman MA: A comparison of a multistate inpatient EHR database to the HCUP nationwide inpatient sample. *BMC Health Serv Res* 2015; 15:384
8. Bryant C, Johnson A, Henson K, et al: Apache outcomes acriss venues predicing inpatient mortality using electronic medical record data. *Critical Care Medicine* 2018; 46:8
9. Heneghan JA, Trujillo Rivera EA, Zeng-Treitler Q, et al: Medications for children receiving intensive care: A national sample. *Pediatr Crit Care Med* 2020; 21:e679–e685
10. Patel AK, Trujillo-Rivera E, Faruqe F, et al: Sedation, analgesia, and neuromuscular blockade: An assessment of practices from 2009 to 2016 in a national sample of 66,443 pediatric patients cared for in the ICU. *Pediatr Crit Care Med* 2020; 21:e599–e609
11. Fung KW, Kapusnik-Uner J, Cunningham J, et al: Comparison of three commercial knowledge bases for detection of drug-drug interactions in clinical decision support. *J Am Med Inform Assoc* 2017; 24:806–812
12. Centers for Disease Control and Prevention: *ICD-9-CM Official Guidelines for Coding and Reporting*. Atlanta, GA, Centers for Medicare & Medicaid Services, 2011
13. Centers for Medicare and Medicaid Services: *ICD-10-CM Official Guidelines for Coding and Reporting FY 2018*. Baltimore, MD, 2018
14. Hardin AP, Hackell JM; Committee On Practice And Ambulatory Medicine: Age limit of pediatrics. *Pediatrics* 2017; 140:e20172151
15. Ma J, Lee DKK, Perkins ME, et al: Using the shapes of clinical data trajectories to predict mortality in ICUs. *Crit Care Explor* 2019; 1:e0010
16. Mohamadlou H, Panchavati S, Calvert J, et al: Multicenter validation of a machine-learning algorithm for 48-h all-cause mortality prediction. *Health Informatics J* 2020; 26:1912–1925
17. Leteurtre S, Martinot A, Duhamel A, et al: Validation of the paediatric logistic organ dysfunction (PELOD) score: Prospective, observational, multicentre study. *Lancet* 2003; 362:192–197
18. Kowarik A, Templ M: Imputation with the R Package VIM. *J Stat Softw* 2016; 74:1–1
19. Zhang Z: Missing data exploration: Highlighting graphical presentation of missing pattern. *Ann Transl Med* 2015; 3:356
20. Platt J: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers* 1999; 10:61–74
21. Powers DM: Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation.*J Mach Learn Technol*. 2011; 2:37–63
22. Saito T, Rehmsmeier M: The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* 2015; 10:e0118432
23. Tharwat A: Classification assessment methods. *Appl Comput Inform* 2020; 17:168–192
24. Boyd K, Eng KH, Page CD: Area under the precision-recall curve: Point estimates and confidence intervals. *In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases: 2013*. Berlin, Heidelberg (Switzerland), Springer, 2013: 451–466
25. Bonafide CP, Localio AR, Roberts KE, et al: Impact of rapid response system implementation on critical deterioration events in children. *JAMA Pediatr* 2014; 168:25–33
26. Pollack MM, Holubkov R, Funai T, et al: The pediatric risk of mortality score: Update 2015. *Pediatr Crit Care Med* 2016; 17:2–9
27. Straney L, Clements A, Parslow RC, et al; ANZICS Paediatric Study Group and the Paediatric Intensive Care Audit Network: Paediatric index of mortality 3: An updated model for predicting mortality in pediatric intensive care*. *Pediatr Crit Care Med* 2013; 14:673–681
28. Dean NP, Fenix JB, Spaeder M, et al: Evaluation of a pediatric early warning score across different subspecialty patients. *Pediatr Crit Care Med* 2017; 18:655–660
29. Rothman MJ, Tepas JJ 3rd, Nowalk AJ, et al: Development and validation of a continuously age-adjusted measure of patient condition for hospitalized children using the electronic medical record. *J Biomed Inform* 2017; 66:180–193
30. Lambert V, Matthews A, MacDonell R, et al: Paediatric early warning systems for detecting and responding to clinical deterioration in children: A systematic review. *BMJ Open* 2017; 7:e014497
31. Trubey R, Huang C, Lugg-Widger FV, et al: Validity and effectiveness of paediatric early warning systems and track and trigger tools for identifying and reducing clinical deterioration in hospitalised children: A systematic review. *BMJ Open* 2019; 9:e022105
32. Parshuram CS, Dryden-Palmer K, Farrell C, et al; Canadian Critical Care Trials Group and the EPOCH Investigators: Effect of a pediatric early warning system on all-cause mortality in hospitalized pediatric patients: The EPOCH randomized clinical trial. *JAMA* 2018; 319:1002–1012
33. Rajkomar A, Oren E, Chen K, et al: Scalable and accurate deep learning with electronic health records. *npj Digital Medicine* 2018; 1:1–10
34. Escobar GJ, Greene JD, Scheirer P, et al: Risk-adjusting hospital inpatient mortality using automated inpatient, outpatient, and laboratory databases. *Med Care* 2008; 46:232–239
35. Croskerry P, Singhal G, Mamede S: Cognitive debiasing 2: Impediments to and strategies for change. *BMJ Qual Saf* 2013; 22(Suppl 2):ii65–ii72

36. Hayes MM, Chatterjee S, Schwartzstein RM: Critical thinking in critical care: Five strategies to improve teaching and learning in the intensive care unit. *Ann Am Thorac Soc* 2017; 14:569–575
37. Tallentire VR, Smith SE, Skinner J, et al: Exploring patterns of error in acute care using framework analysis. *BMC Med Educ* 2015; 15:3
38. Saposnik G, Redelmeier D, Ruff CC, et al: Cognitive biases associated with medical decisions: A systematic review. *BMC Med Inform Decis Mak* 2016; 16:138
39. Cullen DJ, Civetta JM, Briggs BA, et al: Therapeutic intervention scoring system: A method for quantitative comparison of patient care. *Crit Care Med* 1974; 2:57–60
40. Pollack MM, Ruttimann UE, Getson PR: Pediatric risk of mortality (PRISM) score. *Crit Care Med* 1988; 16:1110–1116
41. Knaus WA, Wagner DP, Draper EA, et al: The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. *Chest* 1991; 100:1619–1636
42. Kahneman D: *Thinking, Fast and Slow.* New York, NY, Farrar, Straus and Giroux, 2001
43. Balakrishnan K, Arjmand EM: The impact of cognitive and implicit bias on patient safety and quality. *Otolaryngol Clin North Am* 2019; 52:35–46
44. Itri JN, Patel SH: Heuristics and cognitive error in medical imaging. *AJR Am J Roentgenol* 2018; 210:1097–1105
45. Stiegler MP, Tung A: Cognitive processes in anesthesiology decision making. *Anesthesiology* 2014; 120:204–217
46. Lee CS, Kadom N, Nagy P: Reducing errors from cognitive biases through quality improvement projects. *J Am Coll Radiol* 2017; 14:852–853
47. Tversky A, Kahneman D: Judgment under uncertainty: Heuristics and biases. *Science* 1974; 185:1124–1131
48. Croskerry P: From mindless to mindful practice–cognitive bias and clinical decision making. *N Engl J Med* 2013; 368:2445–2448
49. Croskerry P, Singhal G, Mamede S: Cognitive debiasing 1: Origins of bias and theory of debiasing. *BMJ Qual Saf* 2013, 22(Suppl 2):ii58–ii64
50. Handelman GS, Kok HK, Chandra RV, et al: Peering into the black box of artificial intelligence: Evaluation metrics of machine learning methods. *AJR Am J Roentgenol* 2019; 212:38–43
51. Holzinger A, Langs G, Denk H, et al: Causability and explainability of artificial intelligence in medicine. *Wiley Interdiscip Rev Data Min Knowl Discov* 2019; 9:e1312