Journal of the American Medical Informatics Association, 2024, **31(1)**, 98–108
https://doi.org/10.1093/jamia/ocad174
Advance access publication 30 August 2023
Research and Applications

Research and Applications

Transportability of bacterial infection prediction models for critically ill patients

Garrett Eickelberg, PhD[1], Lazaro Nelson Sanchez-Pinto, MD, MBI, FAMIA[1,2,*], Adrienne Sarah Kline, MD, PhD[1], Yuan Luo, PhD[1,*]

[1]Department of Preventive Medicine (Health & Biomedical Informatics), Feinberg School of Medicine, Chicago, IL 60611, United States, [2]Departments of Pediatrics (Critical Care), Chicago, IL 60611, United States

***Corresponding author:** L. Nelson Sanchez-Pinto, MD, MBI, FAMIA, Department of Preventive Medicine (Health & Biomedical Informatics), Feinberg School of Medicine, 750 N Lake Shore, Chicago, IL 60611 (lazaro.sanchez-pinto@northwestern.edu); Yuan Luo, PhD, Department of Preventive Medicine (Health & Biomedical Informatics), Feinberg School of Medicine, 750 N Lake Shore, Chicago, IL 60611 (yuan.luo@northwestern.edu)

Abstract

Objective: Bacterial infections (BIs) are common, costly, and potentially life-threatening in critically ill patients. Patients with suspected BIs may require empiric multidrug antibiotic regimens and therefore potentially be exposed to prolonged and unnecessary antibiotics. We previously developed a BI risk model to augment practices and help shorten the duration of unnecessary antibiotics to improve patient outcomes. Here, we have performed a transportability assessment of this BI risk model in 2 tertiary intensive care unit (ICU) settings and a community ICU setting. We additionally explored how simple multisite learning techniques impacted model transportability.

Methods: Patients suspected of having a community-acquired BI were identified in 3 datasets: Medical Information Mart for Intensive Care III (MIMIC), Northwestern Medicine Tertiary (NM-T) ICUs, and NM "community-based" ICUs. ICU encounters from MIMIC and NM-T datasets were split into 70/30 train and test sets. Models developed on training data were evaluated against the NM-T and MIMIC test sets, as well as NM community validation data.

Results: During internal validations, models achieved AUROCs of 0.78 (MIMIC) and 0.81 (NM-T) and were well calibrated. In the external community ICU validation, the NM-T model had robust transportability (AUROC 0.81) while the MIMIC model transported less favorably (AUROC 0.74), likely due to case-mix differences. Multisite learning provided no significant discrimination benefit in internal validation studies but offered more stability during transport across all evaluation datasets.

Discussion: These results suggest that our BI risk models maintain predictive utility when transported to external cohorts.

Conclusion: Our findings highlight the importance of performing external model validation on myriad clinically relevant populations prior to implementation.

Key words: critical care; external validation; antibiotic stewardship; machine learning; electronic health records.

Introduction

For patients in the intensive care unit (ICU), bacterial infections (BIs) are a substantial driver of morbidity, mortality, and cost. Repeated international point prevalence studies have found that 51%-54% of patients in the ICU have suspected or proven infections, with ICU mortality rates between 25% and 30%, more than twice that of patients without an infection.[1,2] As a result, physicians in the ICU have a low threshold for initiating empiric antibiotic therapy (EAT). They do so early and broadly, and typically de-escalate or implement targeted therapies based on information collected during follow-up.[3,4] Navigating the difficult decision space between failing to treat a serious BI against overzealous antibiotic regimens is compounded by a lack of consensus treatment guidelines regarding EAT duration and protocol. However, recent efforts have sought to address this by identifying barriers, improving diagnostics, and enhancing antibiotic stewardship as a core competency of critical care.[5–8] A negative consequence of current practice is that patients with low risk of BI are potentially exposed to prolonged and unnecessary antibiotics. Prolonged antibiotic exposure is not risk free and may result in increased antimicrobial resistance in the community as well as a myriad of antibiotic-associated adverse drug events such as gut microbiome dysbiosis and hematologic abnormalities.[9–12] Developing data-driven strategies to help providers stratify patient-level BI risk shortly after an ICU admission offers a promising avenue in antibiotic stewardship.[5,13]

We previously proposed a model to predict BI risk in patients at 24 h following ICU admission in a single-center tertiary ICU setting that achieved an area under the receiver operating curve (AUROC) of 0.8 and a negative predictive value (NPV) >93% in an internal validation cohort.[14] It is widely acknowledged that the performance of any clinical prediction model should be evaluated in different populations using equivalent information prior to clinical implementation.[15–20] For a classification model trained and tested in 2 independent populations, differences between the populations in terms of predictor and/or outcome distributions can lead to

Received: April 10, 2023. **Revised:** July 20, 2023. **Editorial Decision:** August 14, 2023. **Accepted:** August 16, 2023

variation in model class discrimination and calibration performance.[17,21–24] Although a prediction model that is both valid and consistent across external populations is desirable, there are many noteworthy examples that suggest this goal is unrealistic. Studies such as the external validation of a vendor-developed sepsis model presented in Wong et al[25] demonstrated that even models developed with ample resources can demonstrate poor transportability to new settings and generalize poorly to new patient populations. This poses issues for community hospitals and organizations with limited capacity to develop prediction models in-house who may need to rely on models developed at larger institutions or those provided by third parties.

The aim of this study was to assess the transportability of a previously established model and modeling framework[14] on 2 distinct but related cohorts. Specifically, we sought to assess whether the previously published model developed in a tertiary ICU setting had external validity in both a new tertiary ICU as well as a community ICU setting. Additionally, we sought to answer whether retraining of the model with simple multisite learning techniques (data pooling and model ensembling) using data from the 2 tertiary ICUs would improve the performance in the community ICU setting.

Methods

Datasets

Data were obtained from 2 sources, each representing distinct healthcare systems and timeframes. The first dataset was extracted from the Medical Information Mart for Intensive Care III (MIMIC-III). MIMIC-III is a freely available and deidentified dataset collected from over 40 000 patients who received care at Beth Israel Deaconess Medical Center ICU between 2001 and 2012.[26,27] The second dataset was obtained from the Northwestern Medicine (NM) Enterprise Data Warehouse (EDW). The NMEDW is a comprehensive and integrated repository of all clinical and research data sources across the NM health system. NMEDW ICU encounter information was sourced from a manually curated subset of 55 989 ICU encounters across 6 NM affiliated hospitals in Northeastern Illinois (including several community hospitals) for patients admitted between January 10, 2011 and January 1, 2020.

There exist both similarities and differences between MIMIC-III and NMEDW. Both datasets contain administrative, clinical, and physiological data for all ICU encounters. They diverge in how and where their respective data arise from. MIMIC-III predominantly comprised data collected during patients' ICU stays, while the NMEDW is more comprehensive, containing information collected across the continuum of care at NM. The data present in the NMEDW is less curated and thus required more time investment in data cleaning and transformation prior to modeling. Lastly, the NMEDW represents patients seen in varied ICUs and hospital settings. ICU encounters from NMEDW were split into 2 datasets based upon hospital type and geography, where ICU encounters from NM tertiary referral hospitals were labeled NM-T and encounters from NM-affiliated community hospitals were labeled NM-C. Encounters in NM-C represented the use case of community ICUs with interest in implementing a prediction model developed using an external source, and thus served as an external validation cohort ($NM\text{-}C_{val}$) for models developed using MIMIC and NM-T data. For clarity, the unsplit datasets were labeled $MIMIC_D$, $NM\text{-}T_D$, and $NM\text{-}C_{val}$, and the models built from training data were labeled $MIMIC_M$, $NM\text{-}T_M$, $Pooled_M$, and $Ensemble_M$ (details described below).

Cohort

Cohort selection and computational phenotype labeling were performed on all patients as detailed in our prior work.[14] Briefly, patients 16 years or older suspected of having a BI upon admission to an ICU were eligible for our study. Patients matched this phenotype if they had: one or more antibiotic doses administered in the ICU within 96 h of ICU admission and a microbiology culture sampled from a sterile site within 24 h of the first antibiotic dose. Patients who matched these cohort criteria were allocated to 1 of 3 groups based upon their BI status: serious BI, nonserious BI/no BI, and unknown BI status (Figure 1), detailed below. Due to the common occurrence of occult BIs, a direct classification of BI status via microbiology culture indication could result in false negative BI labels. To adjust for this, we considered both duration of antibacterial therapy and microbiology culture status when assigning ICU encounters to the BI status groups.

Microbiology cultures

Microbiology cultures incorporated into cohort enrollment and BI phenotyping were accepted from the following specimens: blood, joint, cerebral spinal fluid, pleural cavity, peritoneum, or bronchoalveolar lavage. Microbiology cultures were assigned a binary classification for "BI indicated" based upon the bacterial species and observed colony size. ICU encounter-level microbiology culture status was considered positive if any microbiology cultures were "BI indicated" in the 72 h following the first qualifying microbiology culture. Information for this classification was sourced from the structured MICROBIOLOGYEVENTS table in MIMIC-III and from free text microbiology reports in the NMEDW. All free text microbiology reports were analyzed using our previously published Python package, *MicrobEx*.[28] Briefly, *MicrobEx* is a rule-based text parser that was developed and externally validated for extracting BI status and bacterial species information from free text microbiology reports.[28] Two consecutive positive cultures were required for positive BI for coagulase-negative *Staphylococcus* and other common contaminate species to warrant inclusion.

Antibiotic prescriptions

All instances of prescriptions used for systemic, empiric, or targeted antibacterial usage were considered for cohort enrollment. In MIMIC-III, prescriptions tagged with Anatomical Therapeutic Chemical (ATC) code J01[29] were selected. In NMEDW, the same was identified using regular expressions, manual curation, and medical expert review.[14] In both datasets, antibiotic duration was calculated as the number of consecutive antibiotic days starting with the first antibiotic dose described above (t_0). ICU encounters were classified as "short" if consecutive antibiotic days were less than 96 h, otherwise they were considered "prolonged." Patients often received antibiotic therapy prior to ICU admission and often continue following discharge. To capture this, consecutive antibiotic duration was permitted to start up to 24 h prior to ICU admission and continue to accumulate up until hospital discharge if the medication was also administered during the patient's ICU stay. Additionally, patients who died within 24 h of their final antibiotic dose were coded as having

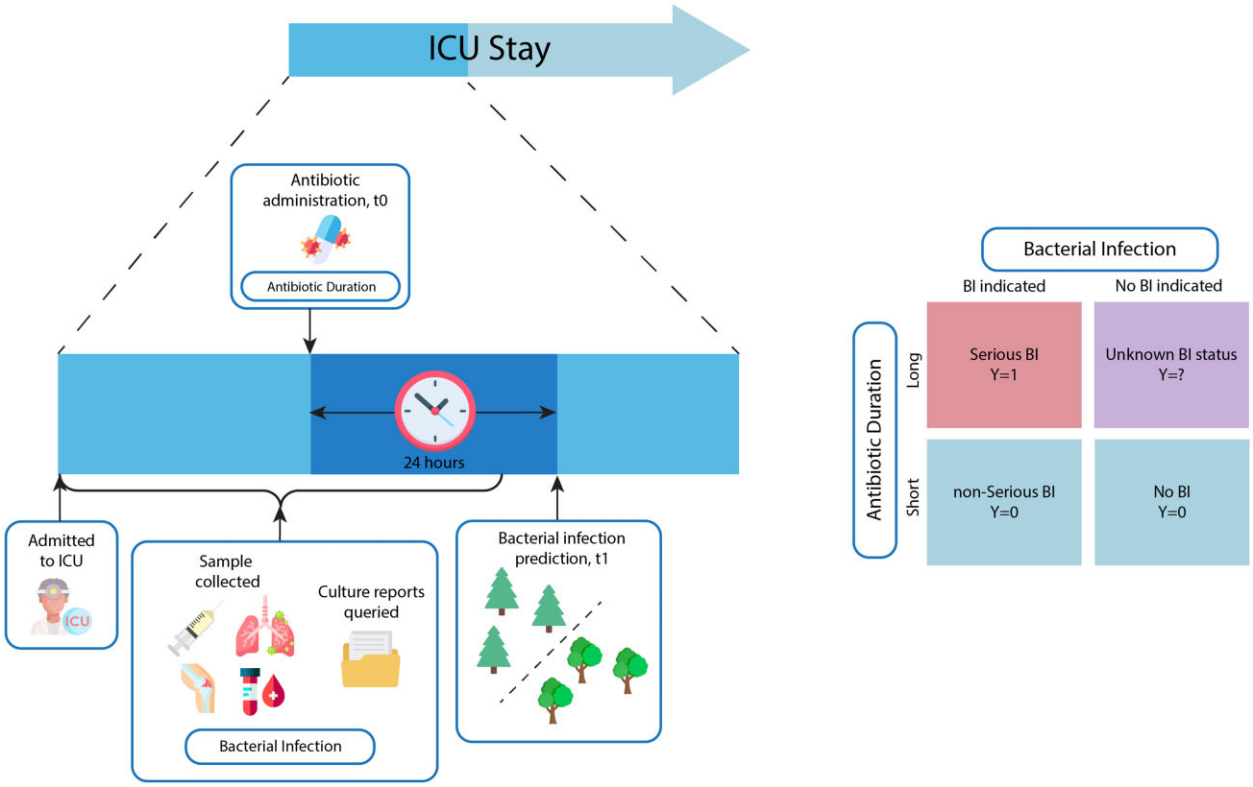

Figure 1. BI status labeling and classification. Our phenotype for BI suspicion upon ICU admission requires that: (1) an antibiotic be administered within 96 h following ICU admission and (2) a microbiology culture be drawn within 24 h of (1). Clinical data were collected for 24 h after the first ICU antibiotic (t_0) and are used to predict binary BI status at t_0+24h (LEFT). Binary infection status was categorized as a function of continuous antibiotic duration and bacterial infection status (RIGHT). BI status was classified as serious (prediction event) for patients who received a positive bacterial culture and prolonged antibiotic therapy. Patients who received short antibiotic therapy with either positive culture (nonserious BI) or negative culture (no BI) were labeled prediction nonevents.

received prolonged antibiotic therapy (n=MIMIC$_D$: 1266, NM-T$_D$: 1238, NM-C$_{val}$: $n = 484$).

Outcome

Patients with both a positive bacterial culture and prolonged antibiotic therapy were classified as serious BI status (prediction event) for the t_{0+24h} prediction timepoint (Figure 1). Patients with negative bacterial culture and a short antibiotic timeline were considered to have no BI (prediction nonevent). Patients with a positive BI culture but short antibiotic treatment duration were considered to have nonserious BIs (prediction nonevent). Finally, patients who received prolonged antibiotics without a positive bacterial culture have less clear infection statuses due to the possibility of occult infections. We follow previous study[14] to categorize these patients as unknown BI status and exclude them from modeling for this study.

Data extraction, cleaning, and preprocessing

We follow our previous studies' in-depth descriptions and open-source code for the extraction, cleaning, and preprocessing of static and longitudinal data from the MIMIC-III database.[14] Data preparations for both NM-T$_D$ and NM-C$_{val}$ followed the same framework and are detailed herein.

Static and longitudinal predictor data were extracted from the NMEDW using structured SQL queries and data warehouse expert support. The query code was adapted from open-source code provided by the team responsible for the MIMIC-III database. All raw longitudinal and categorical variables were collected to reflect the 24-h window after the first antibiotic dose (t_0: t_{0+24}) (Table S1). Raw data were cleaned using an iterative process of data harmonization, quality assessments, and manual review with clinical domain expert input. Disparate units were addressed using conversion dictionaries. Variable density plots, missingness, and distribution parameters were compared across all 3 datasets and manually reviewed (Table S1 and Figures S3-S12). Predictors lacking values during the t_0: t_{0+24} window were considered missing and were imputed using the median values from the associated training set. If issues were identified, conservative thresholds paired with clinical expertise and reference value ranges were used to remove erroneous values. Cleaned data were then converted into median-based unit variances relative to the median and interquartile ranges of patients with prediction nonevents (eqn. 1). Finally, all continuous values within the 24-h collection window were aggregated using functions (minimum, maximum, or both), based on our previous model.[14] After 1-hot encoding categorical variables, our final feature list included 55 variables.

$$Z = \frac{X - \tilde{X}_{\left(\frac{\text{neg}}{\text{short}}\right)}}{\text{IQR}_{\left(\frac{\text{neg}}{\text{short}}\right)}}. \quad (1)$$

Modeling and statistical analyses

ICU encounters from MIMIC$_D$ and NM-T$_D$ were split 70/30 into independent train ($_{train}$) and test ($_{test}$) sets, while encounters in NM-C were set aside for model validation (NM-C$_{val}$).

Individual patients with more than 1 eligible ICU encounter were assigned to the same split. Missing values were imputed using the median values from the associated training set. $Pooled_{train}$ ($n = 4637$) and $Pooled_{test}$ ($n = 1989$) were created by pooling equal sized samples that maintained the respective BI proportions from MIMIC and NM-T.

Random Forests classifiers were trained using Python 3 and scikit-learn.[30,31] Models ($MIMIC_M$, $NM\text{-}T_M$, $Pooled_M$) were trained on $MIMIC_{train}$, $NM\text{-}T_{train}$, and $Pooled_{train}$. Model hyperparameters were selected through a 10-fold cross validation process using a binary cross entropy loss function and a consistent grid-search hyperparameters dictionary (number of trees: [25, 50, 150, 250], max features: [3, 10, 20, 'auto'], max depth: [5, 7, 10, 15], minimum samples split: [2, 5, 10], minimum samples leaf: [2, 5, 10]).

False negative BI classifications are particularly impactful. Thus, steps were taken to calibrate each model to the associated training data, and then measure the class discrimination threshold. Models were fit and calibrated to their associated training set using the *CalibratedClassifierCV* method in scikit-learn, which uses 10-fold cross-validation to estimate classifier parameters and calibrate predicted probabilities using Platt scaling.[30,32] Fit models applied to test sets from differing institutions (eg, $MIMIC_M$ on $NM\text{-}T_{test}$) were first recalibrated on the associated training set (eg, $NM\text{-}T_{train}$). High sensitivity (≥ 0.9) class discrimination thresholds specific to each model and training set were found using 10-fold cross validation. In cases where models demonstrated poor calibration on a given set of training data, a known characteristic of ensembled tree models,[33] the high-sensitivity threshold was instead determined with a ridge regression model via 10-fold cross validation. $Ensemble_M$ was a mean fusion ensemble (soft-voting) assembled from $MIMIC_M$ and $NM\text{-}T_M$ (both calibrated to the associated training set) and was chosen due to its simplicity and comparable class discrimination performance over other weighted and stacked ensembling techniques.[34] Class discrimination and prediction performance among models were measured using AUROC, F1 score, precision, recall, and NPV. Following our main use case, the external validity and transportability of the models were assessed based on class discrimination and calibration performance in the external community ICU cohort ($NM\text{-}C_{val}$). Statistical differences between AUROCs were measured using DeLong's algorithm.[35,36] Model feature importance were calculated using permutation-based methods implemented in scikit-learn based on the impact of shuffling single feature values on model performance. Model calibration was assessed using mean calibration, cox regression, and calibration curves, comparing predicted risk to observed risk.[23,37,38] Case-mix characteristics and relatedness between development cohorts ($MIMIC_D$ or $NM\text{-}T_D$) and the validation cohort ($NM\text{-}C_{val}$) were measured using the AUROC of respective membership models (transportability c-statistic) as recommended in Ref.[22] We set $\alpha = 0.005$ by default, as previously recommended for large datasets.[39] In cases with more than 10 comparisons, a Bonferroni correction was applied. Model fairness was assessed using an equal opportunity fairness definition and was calculated as described in Hardt et al.[40,41]

Results

Cohort characteristics

We identified ICU encounters in $MIMIC_D$ ($n = 19\ 633$; 37.7% of all ICU encounters), $NM\text{-}T_D$ ($n = 11\ 076$; 40.2% of all ICU encounters), $NM\text{-}C_{val}$ ($n = 4059$; 38.8% of all ICU encounters) that met our study inclusion criteria. The demographics of patients in the 3 datasets used are presented in Table 1. Table 2 shows the distribution of bacterial culture results, antibiotic therapy duration, and BI status (prediction variable) across each dataset. Notably, patients in the $MIMIC_D$ were found to have a BI prevalence of 24.8% while patients in $NM\text{-}T_D$ and $NM\text{-}C_{val}$ had BI prevalence of 44.2% and 44.6%, respectively.

Table 1. Demographics of BI positive and negative labeled patients across hospital datasets.

Variable	$MIMIC_D$	$NM\text{-}T_D$	$NM\text{-}C_{val}$
Gender—*N*, %			
Female	5340 (47%)	3112 (47%)	1241 (50%)
Male	6013 (53%)	3514 (53%)	1244 (50%)
Age in years (SD)	65.3±17.0	64.1±17.1	66.7±17.8
Race and ethnicity—*N*, %			
Black/non-Hispanic	1294 (11%)	782 (12%)	151 (6%)
White/non-Hispanic	8218 (73%)	4586 (69%)	2040 (82%)
Hispanic	468 (4%)	606 (9%)	166 (7%)
Other	1373 (12%)	652 (10%)	128 (5%)

NM-T, NM tertiary referral hospitals; NM-C, NM community hospitals; MIMIC, MIMIC-III.

Model evaluation

In Table 3, we present the model evaluation results for the models in the tertiary ICU settings (MIMIC and NM-T). On both test sets ($MIMIC_{test}$ and $NM\text{-}T_{test}$), the models trained and tested on their respective training cohorts (eg, $MIMIC_M$ on $MIMIC_{test}$) had significantly ($P < .002$) higher AUROC than models trained on external development cohorts (eg, $MIMIC_M$ on $NM\text{-}T_{test}$). Relatedness between $MIMIC_{train}$ and $NM\text{-}T_{train}$ measured through the membership model AUROC (case-mix c-statistic) was 0.97, suggesting large differences in case-mix characteristics between both development cohorts.

Table 3 additionally summarizes the results of 2 multisite learning approaches designed to improve overall model generalizability of the models in the tertiary ICU setting. We compared a soft-voting ensemble ($Ensemble_M$) of models ($MIMIC_M$ and $NM\text{-}T_M$), each calibrated to the evaluation site, to a calibrated model trained on pooled training data from each cohort. The AUROC generated by $Pooled_M$ and $Ensemble_M$ were each significantly different from those produced by $NM\text{-}T_M$ on $NM\text{-}T_{test}$ and $MIMIC_M$ on $MIMIC_{test}$ with a Bonferroni-adjusted $P < .002$. Similarly, the difference between the AUROC of $Ensemble_M$ and $Pooled_M$ was significant on $NM\text{-}T_{test}$ ($P = 2\times10^{-4}$) but not on $MIMIC_{test}$ ($P = .037$). The recall values observed for $Pooled_M$ were notably lower than the desired recall of 0.9 on both $MIMIC_{test}$ and $NM\text{-}T_{test}$ due to poor calibration (see below).

Table 4 and Figure 2 summarize model performances in the community cohort ($NM\text{-}C_{val}$) where $NM\text{-}T_M$ showed better or indistinguishable discrimination performance than the other models. Compared to other models, $MIMIC_M$ had a significantly lower AUROC ($P < .002$) and achieved lower precision and recall at a comparable classification threshold. For the multisite models, the AUROC for $Pooled_M$ was significantly different from $NM\text{-}T_M$, however no difference was observed between $NM\text{-}T_M$ and $Ensemble_M$, or between $Ensemble_M$ and $Pooled_M$. Case-mix characteristics between development ($MIMIC_D$ and $NM\text{-}T_D$) and $NM\text{-}C_{val}$ cohorts

Table 2. Cohort stratified by BI status and hospital datasets.

Microbiology culture	Antibiotic duration	BI status classification	MIMIC *N* (% cohort; % ICU)[a]	NM-T *N* (% cohort; % ICU)[a]	NM-C *N* (% cohort; % ICU)[a]
Positive	Prolonged	Positive	2829 (14.4%; 5.4%)	2926 (26.4%; 10.6%)	1109 (27.3%; 10.6%)
Negative	Short	Negative	6988 (35.6%; 13.4%)	2786 (25.1%; 10.1%)	987 (24.3%; 9.4%)
Positive	Short	Negative	1536 (7.8%; 2.9%)	914 (8.3%; 3.3%)	389 (9.6%; 3.7%)
Negative	Prolonged	Unknown	8280 (42.2%; 15.9%)	4450 (40.2%; 16.2%)	1574 (38.8%; 15.0%)

[a] Percentages are listed as percentage relative to patients meeting cohort criteria, and relative to all adult ICU encounters. Patients meeting cohort criteria represented 36.77%, 29.00%, and 36.34% of all adult ICU encounters in MIMIC, NM-T, and NM-C, respectively.
NM-T, NM tertiary referral hospitals; NM-C, NM community hospitals; MIMIC, MIMIC-III.

Table 3. $MIMIC_M$, $NM\text{-}T_M$, $Ensemble_M$, and $Pooled_M$ classification performance.

Model[a]	Evaluation set	Evaluation set BI (%)	AUROC	F1	NPV	Precision	Recall	High sensitivity threshold
$MIMIC_M$	$MIMIC_{test}$	24.8	0.782	0.502	0.924	0.351	0.884	0.131
$NM\text{-}T_M$	$MIMIC_{test}$	24.8	0.694	0.440	0.900	0.291	0.909	0.145
$Pooled_M$	$MIMIC_{test}$	24.8	0.774	0.538	0.878	0.436	0.703	0.131
$Ensemble_M$	$MIMIC_{test}$	24.8	0.767	0.458	0.937	0.303	0.942	0.131
$NM\text{-}T_M$	$NM\text{-}T_{test}$	44.3	0.810	0.715	0.867	0.594	0.898	0.267
$MIMIC_M$	$NM\text{-}T_{test}$	44.3	0.722	0.657	0.808	0.521	0.891	0.274
$Pooled_M$	$NM\text{-}T_{test}$	44.3	0.788	0.696	0.773	0.662	0.734	0.267
$Ensemble_M$	$NM\text{-}T_{test}$	44.3	0.798	0.695	0.879	0.556	0.926	0.267

[a] All models are calibrated to the respective training set (eg, $MIMIC_{train}$) for a given testing set (eg, $MIMIC_{test}$).
NM-T, NM tertiary referral hospitals; NM-C, NM community hospitals; MIMIC, MIMIC-III; $Pooled_M$, equal sized samples from MIMIC and NM-T concatenated together; $Ensemble_M$, soft-voting ensemble of $NM\text{-}T_M$ and $MIMIC_M$.

Table 4. Modeling classification discrimination and performance on $NM\text{-}C_{val}$.

Model	AUROC	F1	NPV	Precision	Recall	High sensitivity threshold	Case-mix C-statistic
$MIMIC_M$	0.741	0.671	0.835	0.529	0.915	0.274	0.98
$NM\text{-}T_M$	0.807	0.712	0.877	0.582	0.919	0.267	0.82
$Pooled_M$	0.795	0.711	0.795	0.644	0.794	0.267	0.87
$Ensemble_M$	0.798	0.697	0.896	0.552	0.945	0.267	N/A

$NM\text{-}C_{val}$ BI prevalence: 44.6%. All models calibrated on $NM\text{-}T_{train}$. NM-T, NM tertiary referral hospitals; NM-C, NM community hospitals; MIMIC, MIMIC-III; $Pooled_M$, equal sized samples from MIMIC and NM-T concatenated together; $Ensemble_M$, soft-voting ensemble of $NM\text{-}T_M$ and $MIMIC_M$; Case-mix C-Statistic, C-statistic from membership model for model's development data and $NM\text{-}C_{val}$.

appear to be highly distinct based on the C-statistics presented in Table 4 and cohort BI status distributions presented in Table 2. Higher C-statistic values also appear to correspond with lower AUROC values in $NM\text{-}C_{val}$, suggesting that model discrimination performance is affected by the case-mix variation in our cohorts. Finally, a visual summary of the best performing models across all evaluation sets is presented in Figure 3. When comparing each model's performance across all evaluation sets, the range of AUC values for multisite learning models were lower (more stable) compared to single institution models.

Figure S15 presents a baseline characterization of $NM\text{-}T_M$, $MIMIC_M$, and $Ensemble_M$ predicted probabilities in the $NM\text{-}C_{Val}$ unknown BI status cohort (prolonged antibiotics and negative microbiologic culture). Overall, the 3 models agreed on 79.8% of classifications, and 85.4% of negative classifications made by $MIMIC_M$ were agreed upon by $NM\text{-}T_M$. Table S5 presents a chi-square contingency table analysis comparing the all-cause in-hospital mortality, evaluated up to 28 days following t_0, versus predicted BI status. Across $MIMIC_{Test}$, $NM\text{-}T_{Test}$, and $NM\text{-}C_{Val}$, we were able to reject the null hypothesis that negative and positive BI risk prediction had equivalent mortality rates at $P < .005$.

Predictor effects

Figure 4 displays the relative variable importance for each model in $NM\text{-}C_{val}$. Maximum temperature was consistently found to have ≥85% relative importance in all models. Having a blood culture performed, leukocytes present in urine and norepinephrine delivered were highly important in some but not all models. These categorical variables also had relatively high differences in distribution among sites, suggesting site-specific predictor effects (Figures S3-S12). Blood urea nitrogen (BUN), heartrate, white blood cell count (WBC), ratio of arterial oxygen partial pressure to fractional inspired oxygen (PaO_2:FiO_2), respiration rate, and systolic blood pressure (SBP) were found to be moderately important among all models. These continuous variables, along with Maximum temperature, had relatively minor differences in distributions between the sites (Figures S3-S12). Figure S13 displays the relative variable importance for the $NM\text{-}T_{train}$/$MIMIC_{train}$ versus $NM\text{-}C_{val}$ membership models. Minimum Glasgow coma score and having blood bands measured were most important to differentiate $NM\text{-}T_{train}$ versus $NM\text{-}C_{val}$, while PaO_2:FiO_2, having a urine culture performed, and receiving external ventilation were most important to differentiate $MIMIC_{train}$ versus $NM\text{-}C_{val}$.

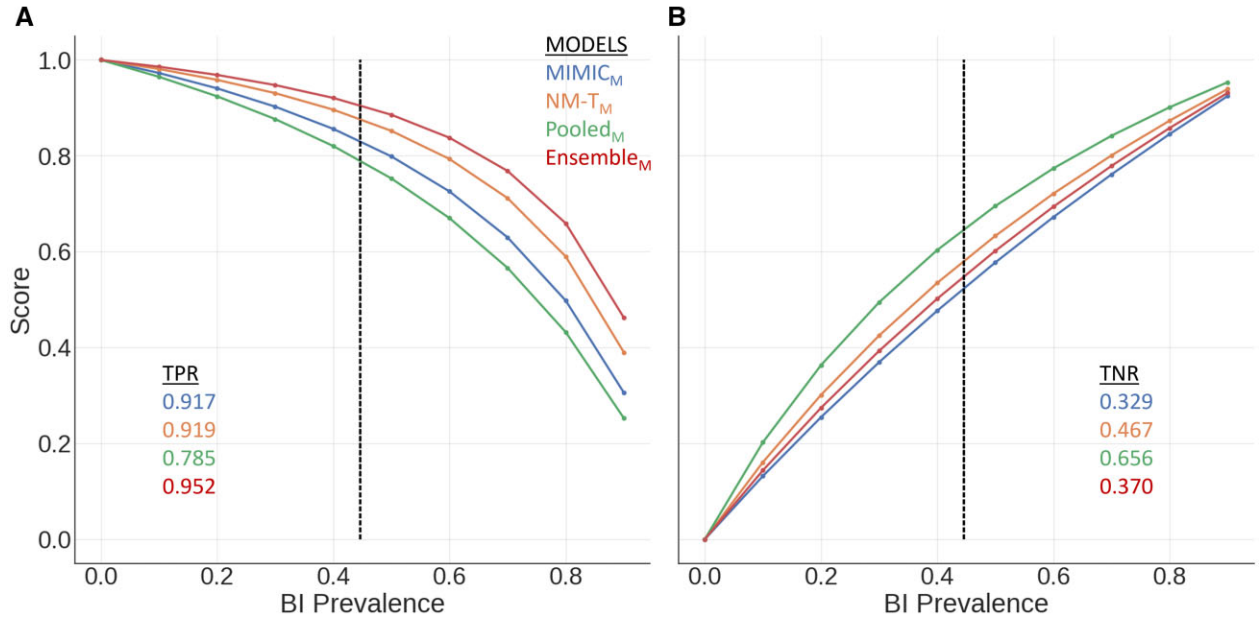

Figure 2. NM-Cval A. Receiver operating characteristic curves (ROC). B Precision recall curves (PRC). The ROC generated by NM-T_M was the highest of any model on NM-Cval and was significantly different than PooledM and MIMICM but not significantly different than EnsembleM at adjusted $P \leq .002$ via DeLong's test. However, there was no difference observed between the ROC of EnsembleM and PooledM. Finally, MIMICM's ROC was significantly different from all other models at $P \leq .002$. All significant differences observed in ROC were additionally observed in PRC.

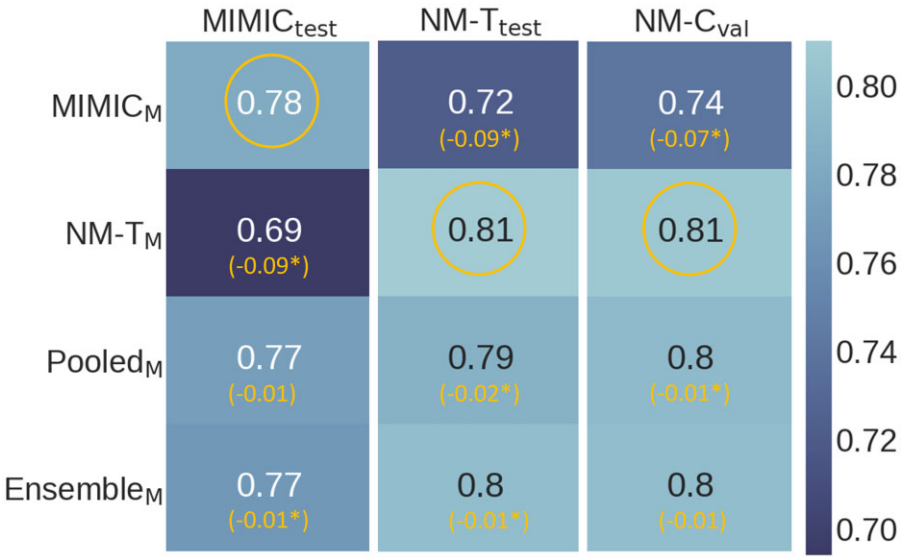

Figure 3. AUROC heatmap between models and evaluation sites. Gold rings indicate the best-performing model on each evaluation cohort while gold numbers present the AUROC delta relative to the gold ring model. *Denotes significant difference from gold ring model using DeLong test at $P < .002$. Although NM-T_M and MIMICM performed best in each of the individual evaluation sets, the difference between their highest and lowest AUROC across all evaluation cohorts was larger (0.012, 0.06, respectively) compared to PooledM and EnsembleM (0.03 and 0.03, respectively).

Model calibration

Overall, average BI predictions for each model closely matched the BI prevalence in NM-T_{test} and NM-C_{val}, however standard deviations were large across all models and datasets, especially for Pooled$_M$ (Table S2). The reliability diagram for patients in NM-C_{val} (Figure 5) and calibration statistics presented in Table S3 suggest NM-T_M, MIMIC$_M$, and Ensemble$_M$ achieve comparable and acceptable calibration on NM-C_{val}. Pooled$_M$ demonstrated poor calibration on patients in NM-C_{val} across all calibration statistics (Table S3) and all BI prevalence patient bins (Figure 5), likely contributing to the mismatch between observed recall and desired recall in Table 4. Table S4 presents results from the NM-C_{val} fairness analysis, where equality of opportunity and AUROC differences observed between advantaged and disadvantaged groups are less than 0.11 (NM-T_M) and 0.09 (MIMIC$_M$).

Discussion

We previously developed and validated a variety of tools and frameworks to help identify patients at low risk of BI who are likely to benefit from discontinuing EAT within 24 h of

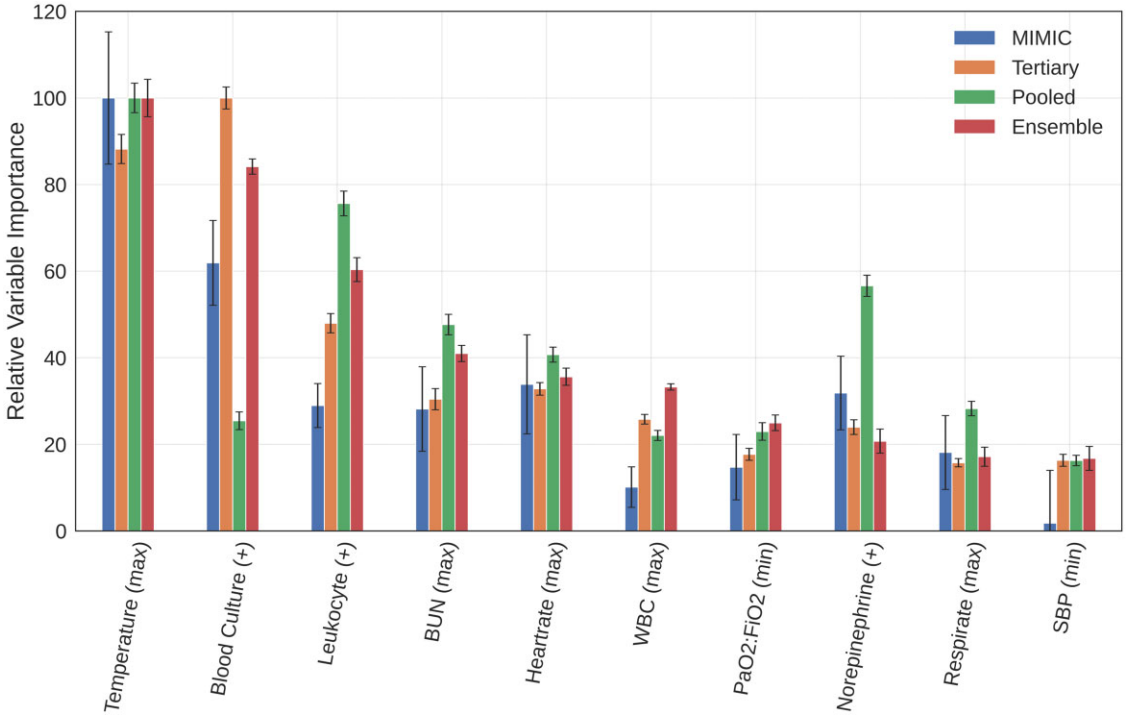

Figure 4. Relative variable importance across models for top 10 important variables. All models are calibrated and permuted on NM-Ttrain. Variable importance values for each model were scaled relative to each model's most important variable. pCO_2, carbon dioxide partial pressure; blood culture, indication for microbiology culture performed on blood sample; leukocyte, indication for leukocytes in urine; BUN, blood urea nitrogen; WBC, white blood cell count; PaO_2:FiO_2, ratio of arterial oxygen partial pressure to fractional inspired oxygen; norepinephrine, indication for having received norepinephrine; respirate, max respiration rate; SBP, systolic blood pressure.

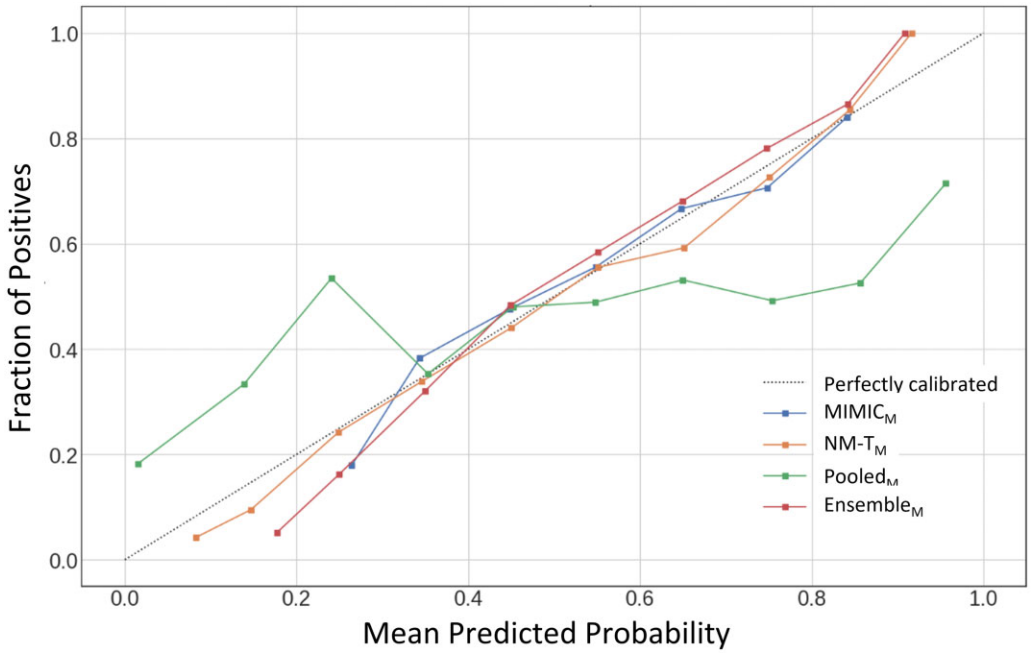

Figure 5. Model calibration plot for NM-C_{val}. All models were calibrated or recalibrated to NM-Ttrain using Platt scaling. MIMICM (Platt scaled), NM-T_M and EnsembleM all achieved adequate calibration despite some deviations from perfect calibration on the lowest and highest BI fractions. Our Pooled RandomForests model demonstrated poor model calibration across a range of BI fractions and was resistant to Platt scaling.

initiation.[14,28] In the current study, we carried out an in-depth transportability assessment of the previously developed BI prediction model architecture on 2 distinct cohorts—tertiary and community ICUs. We additionally explored how model transportability was affected by employing multisite learning (data pooling and model ensembling) with data from each development cohort. In line with a variety of model development and validation recommendations, we placed particular emphasis on: (1) analyzing/correcting for both model discrimination and calibration, (2) examining model performance across different populations, (3) reporting similarity between development and validation cohort, and (4) testing strategies to improve model transportability.[15,18,20,22] Our main findings are as follows: A BI model developed in a historical tertiary ICU (MIMIC) transported adequately to an unaffiliated community ICU (NM-C) with a highly different case-mix, whereas a BI model developed in an affiliated tertiary ICU setting (NM-T) with more similar case-mix and extract-transform-load (ETL) processes transported well. Additionally, learning on information from both tertiary ICU settings (MIMIC, NM-T) offered no significant improvement in model discrimination in the community setting, however multisite models offered more stable transportability across all evaluation datasets. The results from this study demonstrate that while the architecture of a clinical prediction model (ie, electronic phenotype, predictor input, etc.) may be transportable between different sites, the models themselves may not translate with the same performance. Furthermore, model transportability can be affected by numerous factors relating to the case-mix profiles, predictor effects, and ETL processes and thus should be addressed on a case-by-case basis. Our results highlight the importance of performing external model validation on a variety of clinically relevant populations prior to model implementation.

Model performance variation in a new population can be influenced both by the model fit parameters and population case-mix (eg, distribution of predictor variables and setting characteristics).[18,20,21,24,25,42] In the external validation, we observed a relatively strong agreement in variable importance among models. We also observed that model performance was higher in an external validation cohort when the c-statistic comparing case-mix differences between development and validation cohort was lower. These results suggest that case-mix differences between cohorts are a plausible source for the model performance discrepancies observed in the external validation cohort (NM-C_{val}). One potential driver of case-mix variation between cohorts is likely to come from differences in BI prevalence. The higher prevalence of BI in the NM cohorts compared to MIMIC cohort suggest that patients in NM cohorts had either a higher baseline risk for BI or that clinicians at NM sites had a higher threshold of BI suspicion needed to trigger microbiology cultures and empiric antibiotics. This latter interpretation is aligned with the microbiology culture and antibiotic prescribing practice changes expected from recent antibiotic stewardship efforts given that the NM cohorts represent a more contemporary population and practice pattern than the MIMIC cohort.[8,10,43–45] Additionally, differences in BI prevalence between MIMIC and NM cohorts could also be impacted by upstream factors relating to data warehousing, such as availability of information on antibiotic prescriptions and or microbiology cultures performed outside of the ICU. Furthermore, relatively low differences were observed between model performances across demographic subgroups, suggesting cross-site model performance is not driven by equal opportunity disparities. Finally, high levels of semantic and syntactic variability have been shown to exist between data derived from different EHR systems, but the relative effects of these differences on the resulting cohorts remain unclear.[46,47]

While there are no other prediction models that seek to specifically identify patients at low risk of BI, our classification results are in line with previously published models designed for similar prediction tasks. For instance, previous studies reported AUROCs ranging from 0.78 to 0.85 for predicting sepsis and septic shock in the emergency department and ICU setting during the 8-24 h prior to diagnosis[48,49] and AUROCs in the 0.65 to 0.80 range for predicting mortality in septic patients.[50–52] The most important variables in these studies (SBP, BUN, respiratory rate, and temperature) were also among the top 10 most important variables in all 4 of our models. Next, several related publications have reported improvements in model performance after employing multisite learning techniques such as federated and transfer learning.[34,48,53] Wardi et al[48] presented a pretrained deep learning model for sepsis detection that observed significant performance increases after using transfer learning to task specific fine-tuning. While a similar approach could have offered model performance improvements, we chose to use a Random Forests model with simple multisite learning techniques to balance model performance with model simplicity, parsimony, and usability. We found that models trained using 2 simple multisite learning methods (data pooling and ensembling) had indistinguishable or reduced discrimination performance compared to models trained on data from a single institution with a similar case-mix. However, when looking at model discrimination across all evaluation datasets, our multisite learning models offered more stable transportability than single institution models. These mirror the findings of Reps et al,[34] who found that across 5 datasets and 21 outcomes, weighted fusion ensembles produced more stable class discrimination when transported to new databases compared to single database models. Encouragingly, our results suggest that for some multisite learning tasks, model ensembling can offer similar performance to centralized data pooling while also avoiding complicated data sharing processes. These results warrant further study to compare the performance dynamics of data pooling and model ensembling.

Our study has several limitations. First, the observational design of our study required us to use a computational phenotype to infer patient information such as BI status and antibiotic days. Furthermore, due to the free-text nature of microbiology culture notes, we used a previously validated software package that we developed to infer the BI status of patients, but it is possible that patients may have been misclassified in some cases.[28] We addressed both limitations by employing extensive manual case review through the data extraction, preprocessing, and modeling phases of our study for each dataset. A manual review was carried out on 10 false negative patients for $Ensemble_M$ and NM-T_M (Supplementary Data S1 and S2). In this chart review, we identified that in 7 out of the 10 cases, urinary tract infections were the sole infection identified and were often a secondary issue in the encounter. These results highlight the challenge associated with dichotomizing the results from nuanced microbiology reports, where significant variability exists in reporting and interpretation.[54,55]

The intended use-case of our model is to output a data-driven metric that critical care providers can consider when making antibiotic de-escalation treatment decisions starting 1 day after starting antibiotics empirically. In the current paradigm of treating suspected BIs in the ICU, there exist many cases where clinical gestalt and existing methods of characterizing BI risk guide effective antibiotic treatment practices. However, there also exist numerous opportunities where clinical guidelines and existing practices provide less clear guidance on how to proceed. It is for these such cases that our model provides the most utility by reducing the number of unnecessary or inappropriate antibiotic days given to patients who are at low risk of BI. Patients with unclear BI status are arguably those who stand to benefit from our model predictions the most. Figure S15 and Table S5 present a preliminary, but limited, characterization of model performance in patients in this population. To better assess the clinical effectiveness and safety of the model, a cluster-randomized clinical trial, randomizing at the provider or unit level would be warranted.

Conclusion

We evaluated the external validity and transportability of a previously established BI risk prediction model developed in a tertiary ICU setting in both a new tertiary ICU and a community ICU setting. Additionally, we examined whether utilizing simple multisite learning techniques with data from the 2 tertiary ICUs improved model performance in the community ICU setting. Overall, our results suggest that our BI risk models maintain predictive utility when transported to external cohorts. Echoing published guidelines, we recommend that institutions seeking to implement an externally developed prediction model: (1) chose model(s) developed on data with similar case-mix and predictor effects and (2) evaluate and recalibrate the chosen model(s) in the cohort(s) where the model(s) will be used prior to implementation. Furthermore, while models developed with multisite learning have the potential to improve class discrimination and performance stability, these improvements are not guaranteed and should therefore be evaluated on a case-by-case basis.

Acknowledgments

This research was made possible by the ample support provided by Anna Pawlowski, Prasanth Nannapaneni, and Daniel Schneider in the NMEDW Information Technology team.

Author contributions

All authors contributed to study design and funding acquisition. G.E. performed all software development, manuscript writing, data wrangling, data analysis, and model development. A.S.K. assisted in statistical analysis, figure generation, manuscript edits. L.N.S.-P. and Y.L. are co-corresponding authors and were involved in data curation, project administration, and manuscript edits.

Supplementary material

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

Funding

This work was supported by the National Institutes of Health (grant numbers U01TR003528 and R01LM013337), the National Library of Medicine (grant number 5T32LM01220304), and the National Institute of Child Health & Human Development (grant number R01HD105939).

Conflicts of interest

None declared.

Data availability

The benchmark MIMIC-III dataset that supports the findings of this study are available from the official website: https://physionet.org/content/mimiciii/1.4/. The EHR data associated with Northwestern Medicine contain protected health information and are not able to be shared.

Code availability

Python code used to train and evaluate the models in this study can be assessed at GitHub (https://github.com/geickelb/BI_Model_ExValidation).

References

1. Vincent J-L, Rello J, Marshall J, et al.; EPIC II Group of Investigators. International study of the prevalence and outcomes of infection in intensive care units. *JAMA* 2009; 302(21):2323-2329.
2. Vincent J-L, Sakr Y, Singer M, et al.; EPIC III Investigators. Prevalence and outcomes of infection among patients in intensive care units in 2017. *JAMA* 2020;323(15):1478-1487.
3. Goff DA, File TM. The risk of prescribing antibiotics "just-in-case" there is infection. *Semin Colon Rectal Surg*. 2018;29(1): 44-48.
4. Evans L, Rhodes A, Alhazzani W, et al. Surviving sepsis campaign: international guidelines for management of sepsis and septic shock 2021. *Crit Care Med*. 2021;49(11):e1063-e1143.
5. Wunderink RG, Srinivasan A, Barie PS, et al. Antibiotic stewardship in the intensive care unit. An Official American Thoracic Society Workshop Report in collaboration with the AACN, CHEST, CDC, and SCCM. *Ann Am Thorac Soc*. 2020;17(5):531-540.
6. Core Elements of Hospital Antibiotic Stewardship Programs. *Antibiotic Use*. CDC; 2019.
7. Campion M, Scully G. Antibiotic use in the intensive care unit: optimization and de-escalation. *J Intensive Care Med*. 2018;33(12):647-655.
8. Luyt C-E, Bréchot N, Trouillet J-L, Chastre J. Antibiotic stewardship in the intensive care unit. *Crit Care*. 2014;18(5):480.
9. Tamma PD, Avdic E, Li DX, Dzintars K, Cosgrove SE. Association of adverse events with antibiotic use in hospitalized patients. *JAMA Intern Med*. 2017;177(9):1308-1315.
10. Claridge JA, Pang P, Leukhardt WH, Golob JF, Carter JW, Fadlalla AM. Critical analysis of empiric antibiotic utilization: establishing benchmarks. *Surg Infect (Larchmt)*. 2010;11(2):125-131.
11. Francino MP. Antibiotics and the human gut microbiome: dysbioses and accumulation of resistances. *Front Microbiol*. 2015;6:1543.
12. Thomas Z, Bandali F, Sankaranarayanan J, Reardon T, Olsen KM; Critical Care Pharmacotherapy Trials Network. A multicenter

evaluation of prolonged empiric antibiotic therapy in adult ICUs in the United States. *Crit Care Med*. 2015;43(12):2527-2534.
13. Zimmerman JJ. Society of critical care medicine presidential address – 47th Annual Congress, February 2018, San Antonio, Texas. *Crit Care Med*. 2018;46(6):839-842.
14. Eickelberg G, Sanchez-Pinto LN, Luo Y. Predictive modeling of bacterial infections and antibiotic therapy needs in critically ill adults. *J Biomed Inform*. 2020;109:103540.
15. Moons KGM, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med*. 2015;162(1):W1-W73.
16. Klann JG, Estiri H, Weber GM, et al.; Consortium for Clinical Characterization of COVID-19 by EHR (4CE) (CONSORTIA AUTHOR). Validation of an internationally derived patient severity phenotype to support COVID-19 analytics from electronic health record data. *J Am Med Inform Assoc*. 2021;28(7):1411-1420.
17. Collins GS, de Groot JA, Dutton S, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Med Res Methodol*. 2014;14(1):40.
18. Ramspek CL, Jager KJ, Dekker FW, Zoccali C, van Diepen M. External validation of prognostic models: what, why, how, when and where? *Clin Kidney J*. 2021;14(1):49-58.
19. Sanchez-Pinto N, Stroup E, Pendergrast T, Pinto N, Luo Y. Derivation and validation of novel phenotypes of multiple organ dysfunction syndrome in critically ill children. *JAMA Netw Open*. 2020;3(8):e209271.
20. Riley RD, Ensor J, Snell KIE, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *BMJ* 2016;353:i3140.
21. Debray TP, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KG. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J Clin Epidemiol*. 2015;68(3):279-289.
22. Luo Y, Wunderink RG, Lloyd-Jones D. Proactive vs reactive machine learning in health care: lessons from the COVID-19 pandemic. *JAMA* 2022;327(7):623-624.
23. Van Calster B, McLernon DJ, van Smeden M, et al.; Topic Group 'Evaluating Diagnostic Tests and Prediction Models' of the STRATOS Initiative. Calibration: the Achilles heel of predictive analytics. *BMC Med*. 2019;17(1):230.
24. Vergouwe Y, Moons KG, Steyerberg EW. External validity of risk models: use of benchmark values to disentangle a case-mix effect from incorrect coefficients. *Am J Epidemiol*. 2010;172(8):971-980.
25. Wong A, Otles E, Donnelly JP, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern Med*. 2021;181(8):1065-1070.
26. Johnson AEW, Pollard TJ. *The MIMIC-III Clinical Database*. version 1.4. PhysioNet; 2016. https://doi.org/10.13026/C2XW26.
27. Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. *Sci Data*. 2016;3(1):160035.
28. Eickelberg G, Luo Y, Sanchez-Pinto LN. Development and validation of MicrobEx: an open-source package for microbiology culture concept extraction. *JAMIA Open* 2022;5(2):ooac026.
29. Methodology WCCfDS. *ATC Classification Index with DDDs*. World Health Organization; 2019.
30. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825-2830.
31. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5-32.
32. Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv Large Margin Classif*. 2000;10:3-6.
33. Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. In: *Proceedings of the 22nd International Conference on Machine Learning*. Association for Computing Machinery; 2005:625-632.
34. Reps JM, Williams RD, Schuemie MJ, Ryan PB, Rijnbeek PR. Learning patient-level prediction models across multiple healthcare databases: evaluation of ensembles for increasing model transportability. *BMC Med Inform Decis Mak*. 2022; 22(1):142.
35. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988; 44(3):837-845.
36. Sun X, Xu W. Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Process Lett*. 2014;21(11):1389-1393.
37. Huang Y, Li W, Macheret F, Gabriel RA, Ohno-Machado L. A tutorial on calibration measurements and calibration models for clinical prediction models. *J Am Med Inform Assoc*. 2020;27(4):621-633.
38. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J Clin Epidemiol*. 2016;74:167-176.
39. Ioannidis JPA. The proposal to lower P value thresholds to.005. *JAMA* 2018;319(14):1429-1430.
40. Zafar MB, Valera I, Rodriguez MG, Gummadi KP. Fairness beyond disparate treatment & disparate impact: learning classification without disparate mistreatment. In: Proceedings of the 26th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee; 2017:1171-1180.
41. Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning. In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. Curran Associates Inc.; 2016:3323-3331.
42. Nieboer D, van der Ploeg T, Steyerberg EW. Assessing discriminative performance at external validation of clinical prediction models. *PLoS One* 2016;11(2):e0148820.
43. Khilnani GC, Zirpe K, Hadda V, et al. Guidelines for antibiotic prescription in intensive care unit. *Indian J Crit Care Med*. 2019;23(Suppl 1):S1-S63.
44. Singh N, Yu VL. Rational empiric antibiotic prescription in the ICU. *Chest* 2000;117(5):1496-1499.
45. Dellinger RP, Levy MM, Rhodes A, et al.; Surviving Sepsis Campaign Guidelines Committee including the Pediatric Subgroup. Surviving sepsis campaign: international guidelines for management of severe sepsis and septic shock: 2012. *Crit Care Med*. 2013;41(2):580-637.
46. Fu S, Wen A, Schaeferle GM, Wilson PM, Demuth G, Ruan X, et al. Assessment of data quality variability across two EHR systems through a case study of post-surgical complications. *AMIA Annu Symp Proc*. 2022;2022:196-205.
47. Paxton C, Niculescu-Mizil A, Saria S. Developing predictive models using electronic medical records: challenges and pitfalls. *AMIA Annu Symp Proc*. 2013;2013:1109-1115.
48. Wardi G, Carlile M, Holder A, Shashikumar S, Hayden SR, Nemati S. Predicting progression to septic shock in the emergency department using an externally generalizable machine-learning algorithm. *Ann Emerg Med*. 2021;77(4):395-406.
49. Nemati S, Holder A, Razmi F, Stanley MD, Clifford GD, Buchman TG. An interpretable machine learning model for accurate prediction of sepsis in the ICU. *Crit Care Med*. 2018;46(4):547-553.
50. Ding M, Luo Y. Unsupervised phenotyping of sepsis using nonnegative matrix factorization of temporal trends from a multivariate panel of physiological measurements. *BMC Med Inform Decis Mak*. 2021;21(Suppl 5):1-15.
51. Shin J, Li Y, Luo Y, editors. Early prediction of mortality in critical care setting in sepsis patients using structured

features and unstructured clinical notes. In: *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE; 2021.
52. Wang H, Li Y, Naidech A, Luo Y. Comparison between machine learning methods for mortality prediction for sepsis patients with different social determinants. *BMC Med Inform Decis Mak*. 2022;22(Suppl 2):1-13.
53. Corey KM, Lorenzi E, Balu S, Sendak M, editors. Model ensembling vs data pooling: alternative ways to merge hospital information across sites. In: *Machine Learning for Healthcare*. University of Michigan; 2019.
54. Ashley EA, Dance DAB, Turner P. Grading antimicrobial susceptibility data quality: room for improvement. *Lancet Infect Dis*. 2018;18(6):603-604.
55. Turner P, Fox-Lewis A, Shrestha P, et al. Microbiology Investigation Criteria for Reporting Objectively (MICRO): a framework for the reporting and interpretation of clinical microbiology data. *BMC Med*. 2019;17(1):70.