General Audience Summary

Lung infection (pneumonia) is the leading cause of infectious deaths worldwide, and pneumonia occurring in patients who are receiving life support from a mechanical ventilator (ventilator-associated pneumonia) is associated with an increased risk of dying. Machine learning, where computers automatically learn patterns from data to help with making decisions, offers an opportunity to predict which patients will develop ventilator-associated pneumonia. This proposal aims to develop and validate a machine learning model to predict the development of ventilator-associated pneumonia within the next seven days, with the hopes that the early identification of patients at high risk for developing this infection could guide clinicians to intervene and improve the care and outcomes of critically ill patients.

Technical Abstract

Pneumonia is the most common cause of death due to infection worldwide. Ventilator-associated pneumonia (VAP), defined as pneumonia occurring in patients after 48 hours of mechanical ventilation, is the deadliest nosocomial infection, with mortality ranging from 25-70%. While risk factors have been identified and recommended bundles for the prevention of VAP exist, no clinical models are available to specifically predict VAP development. Predictors that have been proposed use features that are late consequences, such as development of multiple organ dysfunction as measured by SOFA trajectory. Identification of patients developing VAP early in their clinical course is important, as timely recognition could prompt earlier interventions that can improve survival.

We recently developed a machine learning approach that incorporates multivariate time series electronic health record data from critically ill patients requiring mechanical ventilation who undergo bronchoalveolar lavage for suspected pneumonia. We used this approach to associate unresolving VAP episodes with worsening clinical trajectories over the ICU stay. These published data support my hypothesis that machine learning applied to electronic health records data will generate a model that predicts subsequent development of VAP with greater accuracy than SOFA score trajectory. I will take advantage of the state-of-the art diagnostics for VAP used in our Medical Intensive Care Unit. Our approach includes bronchoalveolar lavage with quantitative cultures, multiplex bacterial and viral PCR assays, differential cell count, and amylase level (to positively diagnose aspiration pneumonitis). Patient charts (including vitals, labs, notes, images, and microbiologic data) are manually reviewed by a committee of critical care physicians, using a previously developed pneumonia diagnosis and outcomes adjudication process. Accurate diagnosis of VAP is critical to any attempt to model its occurrence.

In Aim 1, I will benchmark proposed predictive scoring systems for VAP, such as SOFA trajectory and CRP. In Aim 2, I will build a machine learning model to predict VAP occurrence in the ensuing seven days. I will furthermore set up a pipeline for this model to be deployed in the electronic health record system for prospective validation. Identifying early predictors of subsequent VAP is the first step to provide opportunities for more aggressive prevention strategies, such as prophylactic aerosolized antibiotics to patients at high risk.

Future Plans

This work will serve as preliminary data foundation for extramural funding via an R01 application. If our model performs well in prospective validation, we will design a randomized controlled trial of usual care vs model-driven care. My multidisciplinary mentorship team is experienced in ICU clinical trials and supports a full-time research team who have successfully conducted >100 trials. This proposal will also enhance my skills with machine learning models, supporting my goal of becoming an independent physician-scientist focused on this area. In the long term, I hope to use predictive machine learning models to identify high risk populations that might benefit from interventions shown in preclinical studies to prevent ventilator-associated pneumonia, such as aerosolized antibiotics or G-CSF/GM-CSF, which have been proposed for study but suffer from lack of enrichment by a high-risk group. This may help improve the care of critically ill patients with respiratory failure.

Promoting Healthcare Equity

Integrating machine learning models with ICU data can be a powerful tool for promoting healthcare equity. These models are adept at processing large volumes of complex data, enabling them to uncover subtle health trends and patterns across different patient populations. Because these models will look only at available clinical data, we can minimize features that may lead to bias. We will specifically exclude race and ethnicity as features for modeling. We acknowledge that a systematic bias is inherent in some data, such as oxygen saturation measurements obtained using pulse oximetry being dependent on skin tone. Conversely, gender is a biological variable and appropriate for inclusion in modeling. Socioeconomic status clearly has an impact on critical care outcomes and potentially on risk of development of VAP but may be hard to model without clear criteria for specific aspects of social determinants of health.

Our study of critically ill patients, [redacted], specifically aims to recruit patients from all races, ethnicities, and with all different kinds of social determinants of health. Consent

documents are translated into Spanish and one or more members of the research associates/coordinators are fluent in Spanish. Current enrollment in [redacted] includes 141 (20.5%) of patients who are of Hispanic ethnicity, and 135 (19.6%) of patients who are of Black/African American race.

Overall, the use of machine learning in analyzing ICU data offers a promising avenue to promote healthcare equity, ensuring that all patients, regardless of their background or circumstances, have access to high-quality, personalized, and equitable healthcare services.

A. STATEMENT OF PROBLEM AND SPECIFIC AIMS

Ventilator-associated pneumonia (VAP), defined as pneumonia occurring in patients after 48 hours of mechanical ventilation, is the deadliest nosocomial infection, with mortality ranging from 25-70%.[1] While risk factors have been identified and recommended bundles for the prevention of VAP exist,[2] no clinical models are available to specifically predict VAP development. Models that have been proposed use features that are late consequences, such as development of multiple organ dysfunction as measured by SOFA trajectory.[3] Early identification of patients who will subsequently develop VAP is important, as timely recognition of high risk patients can facilitate trials of therapies to prevent VAP, including aerosolized antibiotics,[4] granulocyte colony-stimulating factor (G-CSF),[5,6] or granulocyte-macrophage colony-stimulating factor (GM-CSF),[7] which have been proposed for study but suffer from lack of enrichment by a high-risk group. Thus, development of a model to identify which patients are likely to develop VAP early in their ICU course is urgently needed and would be highly beneficial.

Machine learning approaches improve predictive power, incorporating multivariate time series data from monitoring devices and laboratory tests collected as part of clinical care in the intensive care unit (ICU).[8–11] Any predictive model is adversely affected by inaccurate diagnosis of condition being predicted, and VAP itself can be challenging to diagnose. This proposal leverages our center's state-of-the-art use of bronchoalveolar lavage (BAL) sampling paired with quantitative cultures, differential cell count, and multiplex PCR for bacterial and viral pathogens for suspected pneumonia.[12] In addition, we also exclude aspiration pneumonitis, a VAP mimic, with use of BAL amylase levels. My proposal leverages the [redacted] (renewed until 2028), which enrolls critically ill patients requiring mechanical ventilation who undergo a BAL for suspected pneumonia.[12] During my F32, I worked to combine clinical data, multi-omics analyses of BAL samples, and expert adjudication of pneumonia cure using a rigorous multi-reviewer system.[13] To date, we have enrolled >750 patients with over >1700 BAL samples, with continued enrollment expected over the next five years. I care directly for these patients in the ICU, serve on the physician adjudication panel, and have personally performed over 200 bronchoscopies.

I recently published a machine learning approach [redacted] that incorporates multivariate time series electronic health record (EHR) data from [redacted] and used this approach to associate unresolving VAP episodes with worsening clinical trajectories over the ICU stay.[14] This scaffolding now allows me to pivot to detailed predictive modeling, and I present new preliminary data showing the 44 clinical features [redacted] can predict with moderate accuracy (area under the receiver operating curve [AUROC] of 0.72) whether a patient will develop VAP in the ensuing 7 days. As a reference comparison, the widely deployed Epic® EHR Sepsis Model has an AUROC of 0.55-0.73 at predicting sepsis.[15,16]

To address my hypothesis that a machine learning model will enhance VAP prediction, I propose the following Aims:

Aim 1. To benchmark the ability of proposed scoring systems, such as serial SOFA scores, in predicting VAP. Using our unambiguous adjudication of VAP episodes based on bronchoalveolar lavage fluid testing and a validated clinical review process,[17] I will examine the ability of different scoring systems to predict the development of VAP. We hypothesize that these scoring systems will be sensitive but not specific for VAP diagnosis.

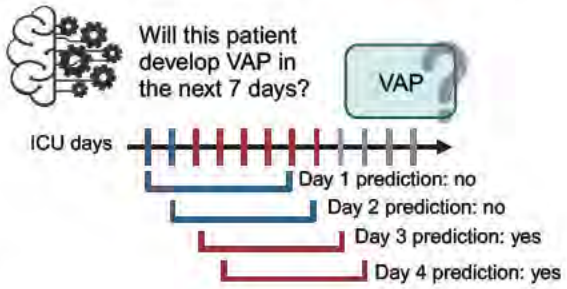


Figure 1. Schematic for a day-by-day model to predict VAP onset within the ensuing seven days.

Aim 2. To develop a machine learning model to predict VAP onset in the ensuing seven days. Aim 2a. Using routinely available laboratory and physiological features from the EHR from the seven days preceding VAP onset, I will build a machine learning model to predict VAP occurrence (**Figure 1**). **Aim 2b.** To build the scaffolding for prospective deployment of this VAP prediction ML model in our EHR.

Identifying early predictors of subsequent VAP is the first step to provide opportunities for more aggressive prevention strategies, such as prophylactic aerosolized antibiotics to patients at high risk. This work will serve as the foundation for a model-based interventional study, providing the publication track record and preliminary data for an R01 application to prospectively validate my clinical model in a randomized study of model-driven vs usual care.

B. BACKGROUND AND SIGNIFICANCE

Ventilator-associated pneumonia (VAP), defined as pneumonia occurring in patients after 48 hours of mechanical ventilation, is the deadliest nosocomial infection, with mortality ranging from 25-70%.[1] Most scoring systems for VAP focus on diagnosis (i.e., the Clinical Pulmonary Infection Score[18]) or outcomes; no clinical models are available to specifically predict VAP development. Thus, development of an accurate model to predict VAP at the episode level, within an ICU stay is urgently needed. Identifying which patients may go on to subsequently develop VAP in the near future can help prompt physicians to pursue early diagnostics (such as early bronchoalveolar lavage) or interventions (such as prophylactic aerosolized antibiotics) that could improve patient outcomes.

Early identification of patients at high risk for VAP offers a key opportunity for intervention. A limitation to many predictive models is the lack of guidance for how to prevent unfavorable outcomes. For VAP, studies including the recent AMIKINHAL trial,[19] have shown that a course of inhaled antibiotics can reduce the incidence of VAP episodes. These studies enrolled critically ill patients, but a model to identify patients at high risk of imminent VAP could help direct therapy to those most likely to benefit from it and limit the harms of exposure in patients unlikely to benefit.

Other potential interventions that could reduce the incidence of VAP and improve outcomes include selective digestive decontamination or selective oropharyngeal decontamination, in which topical non-absorbable antibiotics are applied to the oropharynx or gastrointestinal tract.[20] Proton pump inhibitors are often used for stress ulcer prophylaxis, but these promote bacterial growth and are associated with higher VAP rates.[21] Switching patients at high risk of VAP to an alternative agent such as sucralfate may decrease VAP incidence.[22]

Other therapies that have shown promise in preventing VAP such as GM-CSF,[7,23] but not yet demonstrated benefit in clinical outcomes, suffer from study design that focused on all patients. G-CSF[5,6] may also have potential for prevention.[5,6] Model-driven care focusing on high-risk patients can optimize the design of future studies.

Limited predictive models currently exist for VAP. There are a very limited number of models that predict VAP. Most studies focus on predicting outcomes such as mortality after VAP happens.[24,25] There are some proposed models to predict the occurrence of VAP during an ICU stay, but these are limited because they use ICD code-based VAP diagnosis and do not offer day-by-day predictions.[26,27] A recently proposed score for VAP development in children, RISVAP, includes gender, mechanical ventilation > 4 days, length in PICU>7 days, and previous colonization.[28] While there are some proposed risk scores for VAP,[29,30] these are not commonly employed in clinical practice. Thus, there is great potential utility from a machine learning model to predict VAP on a day-by-day basis.

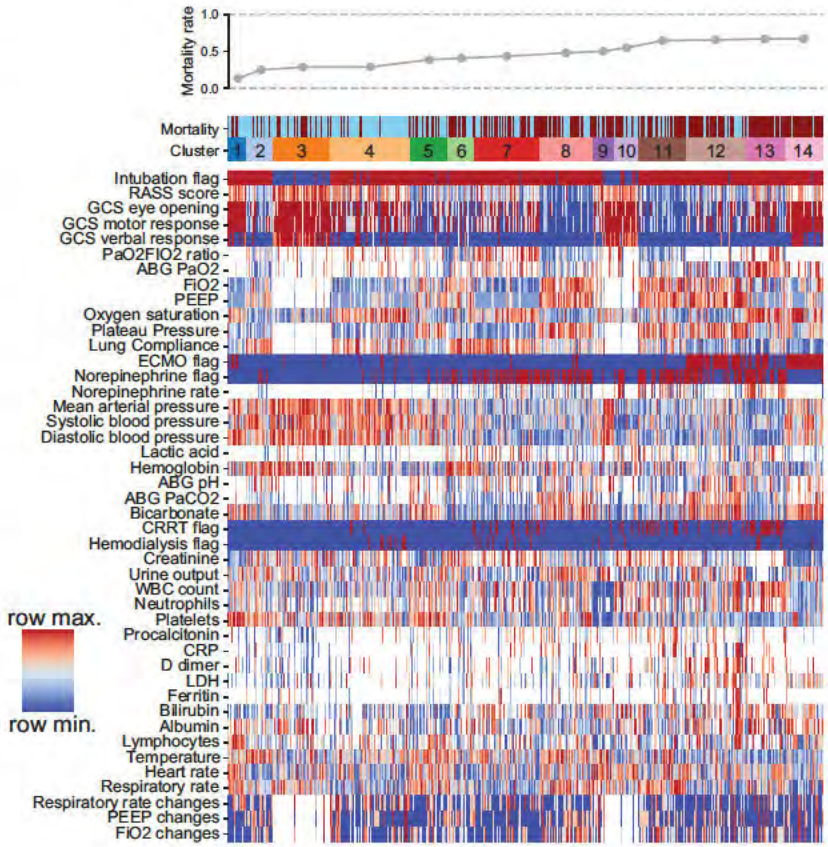


Figure 2. [redacted] groups ICU patient-days into clusters representing discrete clinical states. Heatmap of 44 clinical parameters from 12,495 ICU patient-days grouped into clinical states, ordered by increasing mortality.

Machine learning is a powerful tool for integrating detailed information from complex biomedical datasets, allowing for improved diagnostics and outcome

prediction across medical conditions.[8–10] Importantly, early recognition of clinical trends can allow for timeline interventions that can improve patient care.[31–33] Machine learning approaches can incorporate the multivariate time series data from monitoring devices and laboratory tests available in the ICU.[11] Previous studies have shown that clinicians underperform models at a variety of prediction tasks.[34,35]

Building a per-day scaffold reveals recognizable clinical states: [redacted] we clustered clinical data discretized by patient-day using a machine learning method [redacted] Inspired by the practice of daily physician rounds, we gathered 44 clinical features (including vital signs, laboratory values, ventilator parameters, dialysis status, and others, **Figure 2**). We aggregated these clinical features from 585 patients enrolled [redacted] study and performed hierarchical clustering on 12,495 patient-days.[36] Distinct patterns emerged, with higher-mortality clusters showing more organ failure, findings consistent with previously published scoring systems.[37–39] These day-by-day features and clinical states can be overlaid with a patient timeline to identify predictors of downward trajectory. We examined the [redacted] clinical states through which patients passed before and during VAP episodes. We compared these trajectories by summing up transitions between clinical states and found a statistically significant divergence as early as three days into a defined episode between patients who had cured vs not cured VAP, as well as a downward trajectory the day preceding VAP onset.[14]

Innovation:

Focusing on per-day timepoints allows for examination of intermediate events: The structure of the electronic health record (EHR) is not optimized for research.[40] Even when significant cross-system heterogeneity is removed and data are maximally pre-processed, such as with the Medical Information Mart for Intensive Care (MIMIC),[41] challenges remain in the capture of temporal information.[42] [redacted] overcomes many of these challenges by aggregating data on a per-day basis, allowing a granular evaluation of the unfolded ICU stay.

The interpretation of the dynamic clinical course represents a formidable data science challenge: Analyzing individual VAP episodes rather than an entire ICU stay is critical to identifying predictive factors. It is often difficult to gather and aggregate clinical features in a detailed longitudinal format. Using our [redacted] day-level scaffolding, we can understand data in their correct and detailed clinical context.

Data sharing: As I have done in the past, publishing both a dataset of deidentified data[43] full code repository,[44] and data browser,[45] I will distribute developed code upon publication of manuscripts to benefit future researchers.

A multidisciplinary team effort: I have assembled a team of experts from multiple domains, in order to collaborate on this proposed task. **Primary Mentor** [redacted] An internationally recognized expert in pneumonia clinical and critical care outcomes research, [redacted] been my primary mentor for the T32, LRP, and F32. He has a full-time research team in the Medical ICU to enroll patients, collect samples, and gather clinical data. He has mentored over thirty clinical fellows and early career faculty. [redacted] Dr. [redacted] is an accomplished physician scientist with a focus on the role of regulatory T cells during pneumonia. Dr. [redacted] has worked closely with me throughout fellowship, including for weekly [redacted] meetings; he is senior author on several of my published papers including my most recent paper in the *JCI*. He has mentored a K08 and Parker B. Francis awardee and numerous NRSA pre- and post-doctoral trainees. **Co-Mentor**: **Dr.** [redacted] He is an expert in computational phenotyping and machine learning analysis applied to structured and unstructured EHR data. He will offer significant insight and training on advanced computational analyses to further improve our machine learning models. He has mentored over twenty-three graduate students, postdoctoral fellows, and junior faculty. **Dr.** [redacted] We have worked closely together on my recent [redacted] publication; he will continue to provide overarching career guidance, divisional support, and training guidance. He has mentored over eighteen clinical PCCM fellows and early career faculty. **Dr.** [redacted] she has built and deployed EHR tools and will provide guidance during my model's background deployment and prospective validation. **Dr.** [redacted] is a [redacted] and expert in deep learning. We have collaborated on other deep learning projects to leverage sequential

models to analyze day-by-day changes in ICU patients. **Dr.** [redacted] focused on machine learning methodology and will provide guidance on the analysis of these computationally intensive methods. The entire research advisory committee will meet every six months immediately following my presentation at our Research-in-Progress conference. [redacted] is the Director of [redacted] I have worked closely with his team for the last three years in building the SCRIPT datasets. I will work with his team to build the pipeline for prospective model deployment.

Name	Role	Expertise
[redacted]	[redacted]	Critical care and pneumonia outcomes
[redacted]		Critical care and lung immunology
[redacted]		Machine learning
[redacted]		Mentorship, lung pathobiology
[redacted]		EHR implementation
[redacted]		Deep learning models
[redacted]		Machine learning Biostatistician
[redacted]		Director of [redacted]

C. WORK ACCOMPLISHED AND PRELIMINARY DATA

[redacted]**CU BAL sampling practices**: Physicians in our ICU routinely obtain BAL samples to diagnose or exclude pneumonia in patients with respiratory failure. We have previously shown that they appropriately discontinue or narrow antibiotics in response to this clinical information.[46] Our ICU has a team of trained respiratory therapists who perform non-bronchoscopic BAL in addition to physician-performed bronchoscopic BAL. This approach allows physicians to obtain serial samples in patients, and more than half of [redacted]enrollees had serial sampling. Residual BAL fluid from clinical samples is available for research purposes. These practices provide us with a rich and unique biorepository, giving us a direct view into the alveolar microenvironment, as opposed to studies focusing on blood biomarkers.

Annotation of pneumonia episodes using a validated multi-reviewer process: I am a member of [redacted]panel of critical care physicians who review each chart and provide standardized adjudication of pneumonia episode and outcomes.[13] In intubated patients with severe pneumonia, we define individual bacterial pneumonia episodes as preceded by 48 hours off antibiotics accompanied by a change in clinical status (new fever, leukocytosis, shock, worsening ventilator parameters, new imaging abnormalities), and a bronchoalveolar lavage (BAL) with positive culture, PCR, or neutrophilia.[14,47] We define day 1 of a pneumonia episode as the date of the diagnostic BAL procedure. For bacterial episodes, a positive BAL is defined as culture positivity, PCR-positivity on Biofire Multiplex Pneumonia Pathogen Panel,[13] or >50% BAL neutrophils with a high clinical suspicion of pneumonia as evidenced by systemic symptoms of fever, leukocytosis, shock, and abnormal chest imaging. Patient charts are reviewed independently by two reviewers, who are blinded to each other's responses, and discrepancies are sent to a third reviewer, who is also blinded to the previous two reviewers. Discrepancies between three separate reviewers are discussed at weekly committee meetings.

Any predictive model is adversely affected by inaccurate diagnosis of condition being predicted, and VAP itself can be challenging to diagnose. There exists a wide array of definitions used for VAP amongst different studies, which can make different scoring systems challenging to compare.[48–50] This proposal leverages our center's state-of-the-art use of bronchoalveolar lavage (BAL) sampling paired with quantitative cultures, differential cell count, and multiplex PCR for bacterial and viral pathogens for suspected pneumonia to form a robust framework for expert adjudication.[12] In addition, we also exclude aspiration pneumonitis, a VAP mimic, with use of BAL amylase levels. This rich layer of clinical adjudication provides a cleanly annotated dataset to perform modeling with, and the models developed with this robust methodology to mitigate adjudication bias will enjoy high confidence in predicting VAP onset.

Our preliminary *XGBoost* model using 44 *CarpeDiem* clinical features can predict VAP onset in the next seven days with AUROC >0.7: *XGBoost*[51] is a decision tree-based machine learning method used in clinical predictions[52–54] where it consistently outperforms traditional methods such as logistic regression.[55,56] It supports missing values common in EHR data by learning branch directions for missing values during training and automatically discovers interactions between features.[57] Our preliminary model using *XGBoost* on the 44 laboratory and physiological features from our first version of *CarpeDiem* has an area under the receiver operating characteristics (AUROC) curve of 0.72 at VAP onset in the next seven days, outperforming SOFA score (**Figure 3**). As a reference, the widely deployed Epic® EHR Sepsis Model has an AUROC of 0.55-0.73 at predicting sepsis.[15,16] Important features in model predictions include ventilator parameters such as PEEP and Plateau Pressure, as well as inflammatory markers such as absolute neutrophils from a CBC's differential.

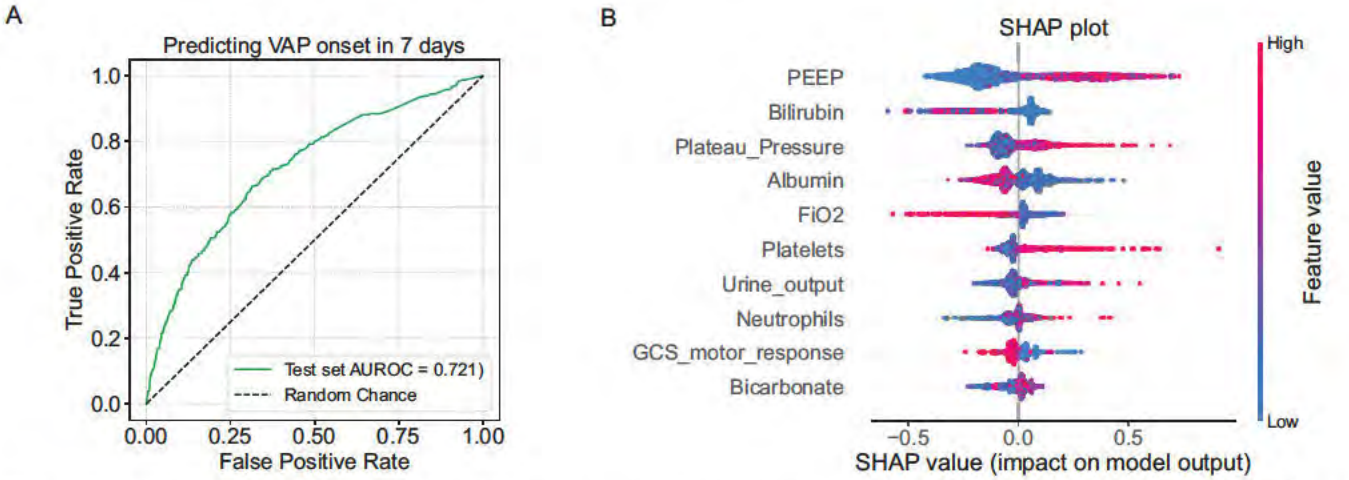


Figure 3. (A) Our preliminary model using 44 laboratory and physiological features from the first version of [redacted] in an XGBoost model can predict VAP development within 7 days with an area under the receiver operating characteristics (AUROC) curve of 0.7. **(B)** SHAP plot, top 10 values. In these plots, each feature is a row, with the color indicating the feature's value (e.g., red for higher values and blue for lower values). The position of the bar on the plot represents the feature's Shapley value, with features on the right side contributing positively to the prediction of getting VAP and those on the left side contributing to the prediction of not getting VAP. The length and direction of the bars illustrate the strength and direction of influence that each feature has on the model's decisions.

D. RESEARCH DESIGN AND METHODS

Aim 1. To benchmark the ability of proposed scoring systems such as serial SOFA scores in predicting VAP development. Most scoring systems for VAP focus on diagnosis (i.e., the Clinical Pulmonary Infection Score)[58] or outcomes[24] and rely on syndromic rather than microbiologic definitions.[59] In our cohort, we have recorded unambiguous adjudications of VAP episodes based on BAL fluid testing and a validated clinical review process.[13] I will benchmark ICU scoring systems' abilities to predict VAP in this more accurate dataset with physician-adjudicated VAP episodes.

Experimental Design: I will leverage [redacted] a prospective single-center study started in 2018 and renewed through 2028. A panel of critical care physicians (including me) reviews all [redacted] cases and annotates pneumonia episodes based on a rigorous multi-reviewer system and predetermined criteria.[13] I will compare previously proposed scoring systems for VAP prediction such as SOFA score trajectory[56] and VAP risk scores.[30] Previous work showed that while neither initial SOFA score nor 48-hour SOFA score predicted the development of VAP, the change in SOFA score from 0 to 48-hours was associated with VAP development during hospitalization. We will examine this pattern in our dataset. Another small study suggested that CRP trend in the first six days of mechancial ventilation is associated with VAP development,[60] and we will examine this in our cohort.

Study population: [redacted] enrolls patients in the medical ICU who require mechanical ventilation and undergo a BAL for suspected pneumonia. I will analyze the clinical data and adjudicated VAP episodes from the 689 patients enrolled during the first five years of SCRIPT with completed adjudications, in whom 397 VAP episodes occurred. Basic demographics and outcomes are presented in **Table 1**.

Table 1. Demographic and outcome information for the patient cohort, grouped by whether or not the patient developed a VAP.

	Overall	Did not have VAP	Had VAP
Number	689	245	444
Age, median [Q1,Q3]	62.0 [51.0,71.0]	62.0 [52.0,70.0]	62.0 [50.0,72.0]
Ethnicity, n (%)			
Hispanic or Latino	141 (20.5)	63 (25.7)	78 (17.6)
Not Hispanic or Latino	516 (74.9)	169 (69.0)	347 (78.2)
Unknown or Not Reported	32 (4.6)	13 (5.3)	19 (4.3)
Gender, n (%)			
Female	279 (40.5)	74 (30.2)	205 (46.2)
Male	410 (59.5)	171 (69.8)	239 (53.8)
Race, n (%)			
American Indian or Alaska Native	4 (0.6)	2 (0.8)	2 (0.5)
Asian	20 (2.9)	9 (3.7)	11 (2.5)
Asian Indian	1 (0.1)	1 (0.4)	
Black or African American	135 (19.6)	49 (20.0)	86 (19.4)
Native Hawaiian or Other Pacific Islander	1 (0.1)	1 (0.4)	
Unknown or Not Reported	121 (17.6)	49 (20.0)	72 (16.2)
White	407 (59.1)	134 (54.7)	273 (61.5)
Smoking status, n (%)			
Former	174 (25.3)	56 (22.9)	118 (26.6)
No	282 (40.9)	100 (40.8)	182 (41.0)
Unknown Smoking Status	172 (25.0)	76 (31.0)	96 (21.6)
Yes	61 (8.9)	13 (5.3)	48 (10.8)
BMI, median [Q1,Q3]	28.5 [24.2,33.8]	29.1 [25.4,34.1]	28.3 [23.5,33.6]
Immunocompromised flag, n (%)	481 (69.8)	181 (73.9)	300 (67.6)
Admit APS score, median [Q1,Q3]	91.0 [64.0,109.0]	91.0 [65.0,109.0]	91.0 [64.0,109.0]
Admit SOFA score, median [Q1,Q3]	11.0 [8.0,14.0]	11.0 [8.0,13.0]	11.0 [8.0,14.0]
Cumulative ICU days, median [Q1,Q3]	16.0 [8.0,28.0]	30.0 [17.0,45.0]	11.0 [6.0,18.2]
Number of ICU stays, median [Q1,Q3]	1.0 [1.0,1.0]	1.0 [1.0,1.0]	1.0 [1.0,1.0]
Tracheostomy, n (%)	178 (25.8)	131 (53.5)	47 (10.6)
Cumulative intubation days, median [Q1,Q3]	11.0 [5.0,24.0]	27.0 [13.0,42.0]	8.0 [4.0,14.0]
Discharge disposition, n (%)			
Died	287 (41.7)	115 (46.9)	172 (38.7)
Home	147 (21.3)	30 (12.2)	117 (26.4)
Hospice	24 (3.5)	2 (0.8)	22 (5.0)
LTACH	78 (11.3)	53 (21.6)	25 (5.6)
Rehab	112 (16.3)	34 (13.9)	78 (17.6)
SNF	41 (6.0)	11 (4.5)	30 (6.8)

Endpoints/Readouts: I will benchmark the ability of described methods detailed above to predict VAP onset within 7 days as the primary endpoint. I will evaluate and compare different scoring systems using AUROC, which combines sensitivity and specificity, precision, recall, accuracy, and F1 score (harmonic mean of precision and recall). Summary statistics at different thresholds will be reported and compared in terms of sensitivity and specificity.

Anticipated results, Pitfalls, and Alternative strategies: I anticipate that this benchmarking will be able to contribute significantly to the literature on traditional scoring systems to predict the development of VAP. This will also be informative that when validated using a clinician-curated dataset of adjudicated VAP episodes, these scores may not perform adequately. Pitfalls inherent to retrospective data analysis approaches include the inability to fully control for confounding factors; we address this issue by trying to correct for features we feel to be interfering by including them as features in modeling and would aim to validate any promising findings in a future prospective manner. The literature in VAP prediction is limited, and if our benchmarking shows that SOFA or CRP trajectory does not adequately predict VAP with an acceptable performance, we will be well poised to move on to more advanced machine learning modeling.

Aim 2. To develop a machine learning model to predict VAP onset in the ensuing seven days. Using routinely available laboratory and physiological features from the EHR in the seven days before VAP onset gives us the preliminary model presented above. I will add new clinical EHR features (such as age, immunocompromised status, therapeutics such as steroids and antibiotics) to my preliminary model and test new deep learning methods (such as recurrent neural networks) to improve our preliminary machine learning model with a goal of AUROC>0.8.

Data variables and confounders: I have worked with the [redacted] (EDW) team to capture patient demographics and outcomes. The EDW stores billions of observations from millions of patients, updates on a nightly basis, and integrates these data with research databases (see Facilities for more details).[61] I curated 44 laboratory and physiological parameters available from the EHR often on a daily basis (vitals, labs, ventilator settings, support devices such as ECMO) in our original [redacted] framework.[14] To the baseline model, I will add important features (**Table 2**) including patient age, sex, duration of ventilation and ICU length of stay, patient immunocompromised status, comorbidities, and therapeutics (prior antibiotic spectrum in the form of Narrow Antibiotic Therapy score[46,62]). By including these as features in the model, they will be incorporated into the decision trees and networks and reflected in the model's output. These additional features are known to be risk factors for the development of VAP and were chosen through discussion by a committee of critical care physicians.

Sex as a biological variable: Men develop pneumonia more frequently than women[63] and have worse outcomes.[64] Hence, I will include sex as a biological variable in the form of a model feature.

Model labels: This proposal will focus on predicting the onset of the first VAP episode for each patient, so 279 VAP cases. The seven days preceeding the VAP episode will be labeled as having VAP onset within seven days, and days prior to that will be labeled as not having VAP onset within seven days. Days following the start of the VAP episode will not be used in this model iteration. Of a dataset of 15,231 patient-ICU-days, 973 occur within 7 days of a VAP episode, and 5119 do not; 9139 days are after a VAP episode starts and not used for this model.

Table 2. Proposed new model features (original 44 features in Figure 2).

Patient-level features
Age
Sex
Immunocompromised status
Comorbidities
Day-level features
Duration of ventilation thus far
ICU length of stay thus far
Antibiotic exposure thus far (mean Narrow Antibiotic Therapy score)
Steroid exposure thus far (cumulative hydrocortisone anti-inflammatory equivalents)

Model development: I will partition the current data into a 70% training set and 30% validation set, with the future [redacted] enrollees being the true test set. Preliminary data have been built using an *XGBoost* model (results above in Figure 3).[51] I will retest our baseline model using the additional features highlighted above, and also examine the ability of other models to predict VAP episode onset, including *LightGBM*,[65] *CatBoost*,[66] Random Forest,[67] logistic regression, and deep learning using recurrent neural networks (**Figure 4**).[68] I will re-approach feature selection to pare down the model to the most critical features, using *scikit-learn*'s feature selection package.[69] I will explore the use of novel encoding methods such as *FIDDLE*,[70] which was developed specifically for EHR data and has been used to build models that are currently in clinical practice to predict high risk *C. difficile* infections.[71] This method encodes data missingness as its own feature, which is important for EHR data as clinically-appropriate data collection is guided by patient condition and clinician decision making. For example, a patient who is improving will likely be sampled less frequently than a patient who is failing to improve unless their care is focused on comfort measures. Co-mentor [redacted] has conducted extensive work and analysis on the handling of missing data, including an ongoing R01 modeling the incompleteness of health data. These methods can be directly applied to my dataset.

Our preliminary machine learning model using XGBoost treats each day as an independent day. We will explore deep learning models that take into account sequential data, that is, uses the trends from multiple days rather than only one day at a time. Deep learning methods perform their own feature selection and can capture time-relative data, so that data from a patient's earlier ICU days are incorporated into the model of their current ICU day.

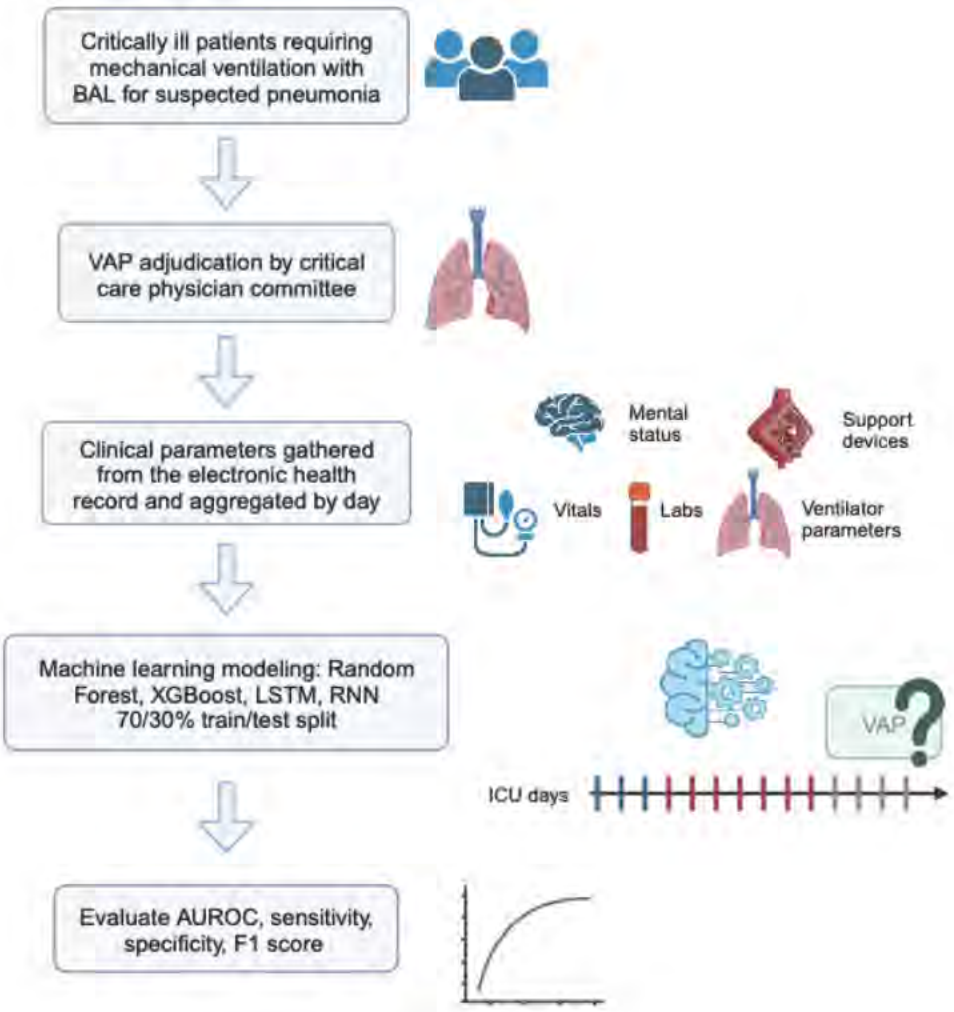


Figure 4. Modeling schematic of day-by-day breakdown of training and test sets and model evaluation metrics.

Endpoints/Readouts: I will use the model's AUROC of predicting VAP onset within 7 days as the primary endpoint. AUROC combines sensitivity and specificity, and I will also evaluate additional modeling measures including precision, recall, accuracy, and F1 score (harmonic mean of precision and recall). Summary statistics at different thresholds will be reported in terms of sensitivity and specificity. I will plot calibration curves and calculate Brier scores (the difference between estimated prediction and actual probability of event outcome)[72] for the different models. Visualization of feature importance will be done with *SHAP* (SHapley Additive exPlanations), which shows directionality and influence on model predictions.[73]

Statistical analysis: Numerical values that are not normally distributed will be compared using Mann-Whitney U tests with false-discovery rate (FDR) correction as appropriate. Categorical values will be compared using Fisher's Exact tests. Different models AUROCs will be compared using Delong's method[74] and the Python package *pyROC*.[75] A $p<0.05$ will be [redacted] *Respiratory and Critical Care Medicine* (see Letter of Support).

Anticipated results, Pitfalls, and Alternative strategies: My preliminary model already predicts VAP onset within 7 days with moderate accuracy (AUROC 0.7). I hypothesize that the addition of pertinent features will only improve model performance. As an example, my preliminary model is agnostic to patient characteristics and timing in ICU stay. I plan to include a new feature such as days of ventilation, which

are known risk factors for VAP. Pitfalls inherent to an observational method include the inability to fully exclude unmeasured confounders and difficulties in capturing all the nuance involved in the bedside care of critically ill patients. I address this by including a host of important features identified by critical care physicians that were not present in our original dataset. I will take measures to avoid overfitting by tuning model hyperparameters and using cross-validation.[76] Trialing new strategies for imputation may lead to a more robust model that is better able to predict pneumonia cure; a specific algorithm to try would be multivariate imputation by chained equation (*MICE*).[77] It may also be difficult to predict 7 days ahead of VAP onset with high accuracy; we can further examine whether forecasting a shorter period ahead of time, such as 5 or 3 days, may provide a more accurate model.

Aim 2b. To build the scaffolding for prospective deployment of this VAP-prediction ML model in our EHR. I will deploy my model and validate it prospectively in newly enrolled patients.

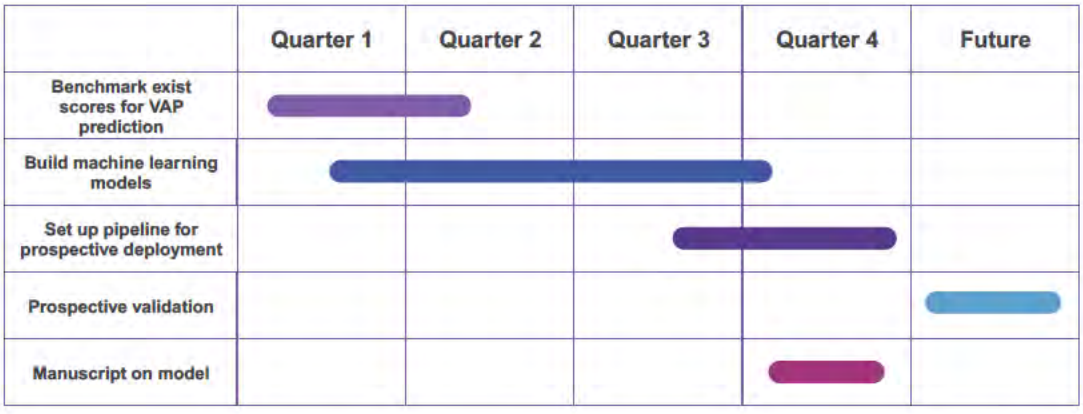
Rationale: It is critically important to validate models on entirely new datasets and prospectively,[78] and this key step is currently only performed in a small fraction of clinical machine learning studies (~2%).[79] External validation consistently has lower performance than the original development cohort.[80] Dataset shift, a difference between the original development cohort and the data from new patients, may cause the model to underperform when deployed.[81] Also, small variations in patient selection and even technical changes in timing of model calculation (temporal shift and infrastructure shift) can cause unexpected changes in model performance.[82]

Model deployment: I will work with [redacted], leader of the [redacted] team (see Letter of Support), and Analyst [redacted], to build an automatic pipeline to extract the features from Aim 1a to run nightly to predict future pneumonia episodes in patients as they are admitted to the MICU and require mechanical ventilation. I have worked closely with this team over the last three years to build clinical databases and pipelines for critically ill patients.

Endpoints, Readouts, Statistics: as described above. Model predictions will be run automatically, in real time, and the predictors recorded in a REDCap database. The critical care adjudication committee will be blinded to the results of the model when they are performing their pneumonia episode adjudication. I will use the physician adjudication as gold standard and compare the model against this with AUROC.

Anticipated results, Pitfalls, and Alternative strategies: Based on our preliminary data, I expect the model I developed will report an AUROC>0.7 for predicting VAP episode onset. If the model fails to perform at this level, I will investigate potential reasons for discrepancy including temporal shift and infrastructure shift, as well as dataset shift as outlined above. This will provide me critical training in the practical skills and processes of model deployment and validation.

Study timeline:



Future directions: If the model performs well in prospective validation, our next step would be to design a randomized controlled trial to study usual care versus model-driven care in preventing VAP in high-risk patients (**Figure 5**). My multidisciplinary mentorship team is experienced in the design of ICU clinical trials and supports a full-time research team who have successfully conducted more than one hundred trials. This work will serve as preliminary data foundation for extramural funding via an R01 application. This proposal will also enhance my skills in the development and validation of machine learning models, supporting my goal of becoming an independent physician-scientist focused on this area. In the long

term, I hope to use predictive machine learning models to identify high risk populations that might benefit from interventions shown in preclinical studies to prevent VAP such as aerosolized antibiotics, GMCSF, or interferons, which have been proposed but suffer from lack of a high-risk group. This may help improve the care of critically ill patients with respiratory failure.

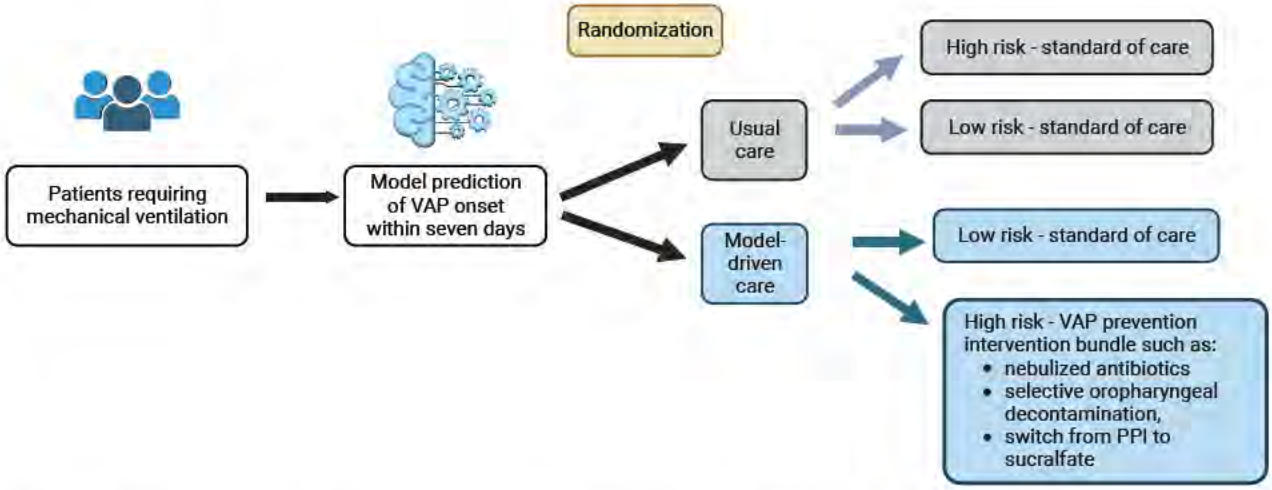


Figure 5. Design schematic for a potential future randomized controlled trial of usual care versus model-directed care to prevent ventilator-associated pneumonia.