NORTHWESTERN UNIVERSITY

# Modeling Bacterial Infection Risk for Data-Driven Antibiotic De-Escalation in Critically Ill Adults

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Biomedical Informatics in the Driskill Graduate Training Program in Life Sciences

By

Garrett Eickelberg

EVANSTON, ILLINOIS

September 2023

# ABSTRACT

Bacterial infections (BI) are a frequent, expensive, and life-threatening condition for critically ill patients. For patients with serious BI, minimizing the time between admission to the intensive care unit (ICU) and administration of appropriate antibiotic therapy is crucial to improve prognosis. However, the current gold-standard for identifying the appropriate antimicrobials to administer, microbiology cultures, have long resolution times and are rarely able to guide early antibiotic treatment choices in the ICU. Consequently, critical care providers are advised to broadly administer empirical antibiotics to all patients suspected of BI, then adjust the treatment regimen based on follow-up information. This approach presents challenges in cases where the infection status is uncertain and current methods of characterizing BI risk underperform. In this paradigm, populations of patients with low risk of BI are exposed to unnecessarily prolonged antibiotic regimens and experience iatrogenic harm as a result. In this thesis, we demonstrated how leveraging electronic health record data with statistical learning techniques and informatics tools can supplement existing BI detection methods to inform antibiotic de-escalation decisions in the ICU. First, we developed and optimized a modeling framework to predict patient-level BI risk using an open and de-identified ICU database. Next, we developed and validated an open-source python package (MicrobEx) to extract BI status concepts from free-text microbiology reports. Then we performed an external validation and transportability evaluation of our BI modeling architecture in two unaffiliated tertiary intensive care unit (ICU) settings and a community ICU setting. Finally, using these same data sources, we performed a retrospective impact study to estimate the treatment effect of prolonging antibiotic therapy past 96 hours in critically ill patients predicted to have low risk for BI and adjust for selection bias using propensity score matching. Our analyses showed that sensitive and

transportable performance can be achieved by using longitudinal patient features, such as temperature and white blood cell count, to predict BI status with our modeling framework. Furthermore, we present compelling evidence that critically ill patients who are predicted to be at low risk for BI can experience improved outcomes when discontinued from antibiotic therapy prior to 96 hours. To our knowledge, these analyses are the first to utilize EHR based clinical prediction modeling to help guide antibiotic de-escalation decisions in critically ill adults.

# ACKNOWLEDGMENT

Five years ago, I packed up all my earthly belongings and moved across the country to pursue a PhD in Biomedical Informatics at Northwestern University. The work I present here is the product of the most stimulating, humbling, and meaningful endeavor of my life thus far. Naturally, all of this has been made possible as a direct result of the incredible support, love and mentorship provided to me by many wonderful people.

First and foremost, I want to express sincere gratitude and appreciation to my advisor Dr. Yuan Luo and unofficial advisor Dr. Nelson Sanchez-Pinto. Thank you for your frequent support, for fostering my independence, and for guiding me to become a better scientist and thinker. Although completing my graduate work during a global pandemic was not something I was prepared for, I feel incredibly fortunate to have had both of you by my side throughout those trying times. I would also like to extend my appreciation to the additional members of my thesis committee, Drs. Justin Starren, Lee Cooper, and Ellick Chan, for their ample availability, guidance, and advocacy. At numerous times in my graduate journey, my thesis committee demonstrated that they had my best interests at heart, and I feel fortunate to have had the opportunity to work with such a caring, wise, and trustworthy people. I want to further thank Dr. Starren for, on numerous occasions, providing potent feedback and insight that helped me navigate some of the most challenging decisions during my graduate career. Similarly, I want to directly thank Drs. Steve Anderson, Pamela Carpentier, and Toni Gutierrez for the opportunity to study in the Driskill Graduate Program and for the student advocacy you provide behind-the-scenes.

Thank you to all the colleagues and collaborators in the Luo lab. This research was also made possible by the ample support provided by Anna Pawlowski, Prasanth Nannapaneni, and

Daniel Schneider in the NMEDW Information Technology team. To Dr. Adrienne Kline and Meghan Hutch, thank you for being so caring, helpful, and easygoing! I will cherish our memories of working in the trenches together. I also want to thank all my wonderful friends and climbing buddies for bringing joy and balance to my life.

Finally, I feel so blessed to have such an outgoing and loving family. I want to express my love and gratitude for my Parents (Steve and Veva), my brother and sister-in-law (Christian and Kelsey), my amazing girlfriend (Alaina), and my loving dog (Dilly). You all are truly the backbone of my support network and I have relied upon your unconditional love, support, and wisdom at every point in this journey. My heart is full, and I am incredibly grateful for you all.

# GLOSSARY OF TERMS AND ABBREVIATIONS

| | |
|---|---|
| ADE | Adverse Drug Events |
| AMR | Antimicrobial Resistance |
| ATC | Anatomical Therapeutic Chemical |
| ATM | Average Treatment effect in propensity score Matched individuals |
| AUC, AUROC | Area Under the Receiver Operator Curve |
| BI | Bacterial Infection |
| BUN | Blood Urea Nitrogen |
| CAP | Community-Acquired Pneumonia |
| CI | Confidence Interval |
| CSF | Cerebral Spinal Fluid |
| EAT | Empiric Antibiotic Therapy |
| EHR | Electronic Health Record |
| ETL | Extract-Transform-Load |
| FiO2 | Fractional Inspired Oxygen |
| FN | False Negative |
| FP | False Positive |
| GCS | Glasgow Coma Scale |
| HAP | Hospital-Acquired Pneumonia |
| HIPAA | Health Insurance Portability and Accountability Act |
| ICU | Intensive Care Unit |
| INR | hemoglobin International Normalized Ratio |
| IQR | InterQuartile Range |
| K-NN | K- Nearest Neighbors algorithm |
| LOS | Length Of Stay |
| MAP | average Arterial blood Pressure over one cardiac cycle |
| MIMICIII | Medical Information Mart for Intensive Care III dataset |
| MLP | MultiLayer Perceptron |
| MRSA | Methicillin-Resistant Staphylococcus Aureus |
| NDC | National Drug Codes |
| NLP | Natural Language Processing |
| NM, NMH | Northwestern Medicine, Northwestern Memorial Hospital |
| NM-C | Northwestern Medicine, Community ICUs |
| NM-T | Northwestern Medicine, Tertiary ICUs |
| NPV | Negative Predictive Value |
| OHDSI | Observational Health Data Sciences and Informatics |
| PaO2 | Partial pressure arterial oxygen |
| pCO2 | Partial pressure of arterial Carbon Dioxide |
| PRC | Area Under the Precision Recall Curve |
| PTT | Partial Thromboplastin Time |
| RC | matched-pairs Rank biserial correlation Coefficient |
| RCT | Randomized Control Trials |
| RXCUI | RxNorm Concept Unique Identifier |

| | |
|---|---|
| SBP | Systolic Blood Pressure |
| SMD | Standardized Mean Difference |
| SNOMED | Systemized Nomenclature of Medicine |
| SpO2 | Peripheral Oxygen Saturation |
| SQL | Structured Query Language |
| SVC | Support Vector Classifier |
| $t_0$ | time of each patient's first antibiotic dose |
| TN | True Negative |
| TP | True Positive |
| UTI | Urinary Tract Infection |
| VAP | Ventilator-Associated Pneumonia |
| WBC | White Blood cell Count |
| XGB | eXtreme Gradient Boosted decision tree |

# DEDICATION

I dedicate this thesis to my family, friends, and all the loving and supportive people I am fortunate to have in my life.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# 1. INTRODUCTION

## 1.1 Management of Bacterial Infections in Critical Care Settings

Bacterial infections (BI) pose severe threats to critically ill patients. Approximately half of all patients admitted to the intensive care unit (ICU) will have suspected or proven BI, and approximately 30-45% of them will die in the hospital (1-8). The frequency and severity of BI are elevated in the ICU for many reasons. First, patients in the ICU are critically ill and are commonly admitted with multiple comorbid conditions, severe pathologies, and/or immune suppression (9). Furthermore, ICU admission rates are frequently higher in populations with elevated intrinsic infection risk factors such as advanced age and lower socioeconomic status (9, 10). Taken together, these factors create an environment where uniquely vulnerable patients experience significant morbidity and mortality resulting from BI.

Critical care providers are particularly on the lookout for the most common and severe conditions associated with BI, namely pneumonia and sepsis. Pneumonia is a respiratory infection characterized by inflammation of the alveolar airspace and is the main cause of death due to infections around the world (11). Pneumonia is most often triggered by a BI, however, it can also be caused by other infectious microorganisms and less frequently by autoimmune processes (12). In the ICU, severe pneumonias are broadly placed into three categories with different treatment protocols: community-acquired pneumonia (CAP), hospital-acquired pneumonia (HAP), and ventilator-associated pneumonia (VAP). Sepsis is a life threatening host-response to an infection and a major global healthcare problem due to its high mortality (9). While any type of infection can lead to sepsis, respiratory and urinary tract infections (UTI) are the two most common etiologies.

Antibiotics are the bedrock of therapy for critically ill patients with serious BI. In the ICU, point prevalence studies have found that around 70% of patients receive at least one dose of antibiotic therapy and more than half of the antibiotic doses administered are given to patients without a confirmed BI (13-20). For serious BI, minimizing the time delay between ICU admission and delivery of appropriate antibiotic therapy is crucial to improve patient prognosis (2, 13, 21-24). For instance, each hour of delay in appropriate antibacterial therapy after onset has been associated with ≥8% increase in crude mortality in patients with sepsis (23, 24). Unfortunately, microbiology cultures, the current gold-standard for determining the appropriate antimicrobials to administer, are rarely able to guide initial antibiotic treatment choices on the timescales needed in the ICU (12). Therefore, critical care providers are recommended to cast a wide net and treat all BI suspected patients early, broadly, and empirically. This strategy saves lives but also incurs collateral damage.

Even in cases with confirmed BI, antibiotic administration can have harmful effects. Antibiotic exposure has been associated with numerous adverse drug events (ADE) such as allergic reactions, hepatic injury, immune cell dysfunction, microbiome dysbiosis, and increased risk for subsequent infections from *Clostridium difficile* and antibiotic-resistant organisms (25-31). The prolonged exposure to some commonly prescribed antibiotics in the ICU has also been associated with increased risk of cardiac events and mortality (32-34). Finally, unsuitable antibiotic prescriptions are known contributors to the emergence of increasingly drug resistant microorganisms. The adverse effects of antibiotics are particularly concerning in cases of antibiotic misuse. It's currently estimated that up to 60% of antibiotic prescriptions in the ICU are unnecessary, inappropriate, or suboptimal (15-20, 35, 36). Some of the most common reasons for antibiotic overuse include prescribing antibiotics for longer durations than indicated, for

nonbacterial/noninfectious illness, and for conditions resulting from contaminating organisms (25). Reducing the amount and duration of unnecessary antibiotic treatments is one of many needed changes to stem the rapid proliferation of resistant bacteria and abate antibiotic ADE (15-18).

Antibiotic stewardship programs seek to optimize the delivery of antibiotic therapy and patient outcomes by "helping providers find the right drug at the right time and the right dose for the right bug for the right duration" (37, 38) (Figure 1). Indeed, stewardship-focused antibiotic prescribing practices have been repeatedly associated with non-inferior or improved patient outcomes (39, 40). However, despite the apparent benefits, adoption of antibiotic stewardship practices has not yet become widespread. For instance, multi-site studies found that broad-spectrum therapy was narrowed in only 30-40% of patients who lacked evidence of a resistant pathogen BI (39, 40). To understand this gap, it can be helpful to think of antibiotic stewardship occurring in two stages in the ICU. First, antibiotic stewardship recommendations call for prompt delivery of empiric antibiotic therapy to maximize coverage for patients with serious BI. At this point, it's particularly important that microbiologic specimens be obtained prior to receiving antibiotic therapy (if possible), and that appropriate empirical regimens be selected based upon patient-level risk factors in addition to community and hospital stratified antibiograms (41). In the second phase, care teams work to de-escalate antibiotic therapies by either discontinuing antibiotic therapy or narrowing the spectrum of antibiotics to target an infection once characterized. Many stewardship techniques focusing on data-driven antibiotic de-escalation hinge upon sensitive and robust characterization of BI risk, and the need for large advancements in BI diagnostics remains unmet (19, 38, 42). Developing new methods for

improved BI diagnostics and pathogen characterization offer a promising route to improve the adoption and efficacy of antibiotic stewardship practices.

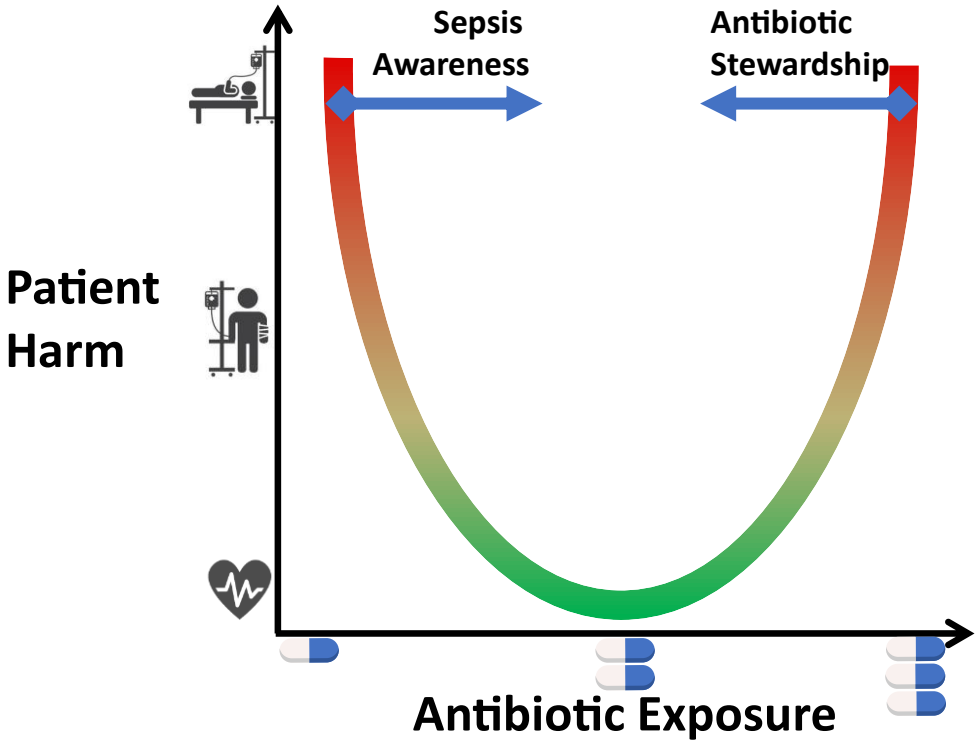


*Figure 1. Balancing antibiotic exposure with patient harm in critically ill patients.The severity, frequency, and time sensitive nature of BI in the ICU justify increased antibiotic intensity and exposure to mitigate risks of conditions like sepsis and pneumonia. On the other hand, antibiotic ADE, multidrug resistant bacteria, and antibiotic stewardship efforts justify actions to reduce the intensity, breadth, and exposure of prescribed antibiotics when appropriate. A conceptual 'goldilocks' zone exists when these forces are balanced. In this zone, patient outcomes are maximized by treating true BI with sufficient intensity and patient harm is minimized by de-escalating antibiotics when BI are counter indicated.*

## 1.2 Bacterial Infection Diagnostics

For over a century, BI diagnostics have relied primarily on microscopy, microbiologic cultures, and immunodiagnostics. Recently, the need for faster resolution and increased sensitivity and specificity has driven the development of newer technologies. However, there is

still uncertainty around how these methods will fit into the current paradigm. Microbiological cultures, often augmented with gram stains and nonamplified probes, remain the gold standard for bacterial pathogen detection and characterization (43-45). While reliable in many use cases, these techniques often suffer from slow turnaround times, varying interpretations, and suboptimal sensitivity, such as in cases of suspected bacteremia (43, 46-49). Diagnostic biomarkers, which can target either host-response or specific microbes, have become increasingly popular in critical care settings to complement traditional microbiological cultures and clinical gestalt. Most of the biomarkers utilized today target host-response and commonly encounter specificity issues in critically ill patients with altered immune function or multiple sources of inflammation (38, 50, 51). Host-response biomarkers such as C-reactive protein, procalcitonin, and white cell count have been shown to provide additional utility in the initial diagnosis of infection, but their evidence in guiding antibiotic therapy decisions is mixed. Despite recent progress in BI diagnostics, the lack of reliable microbial presence biomarkers drives research and development of new BI diagnostic methods.

Over the last 20 years, more than 250 targeted and host-response biomarkers have been proposed for sepsis diagnosis and prognosis (52). However, less than 30% of them have been evaluated in studies with over 300 participants or in multiple studies (52, 53). Despite new molecular biology technologies and the development of multiplexed, omics-based, and point-of-care assays, the promise of rapid BI diagnostics remains largely unfulfilled (38). Additional validation and impact studies are necessary, and concerns about cost continue to present significant obstacles (38, 45). Furthermore, when bacterial nucleic acids are amplified it's difficult to distinguish if they originated from a case of active infection versus a resolved infection or asymptomatic colonization (45, 51, 54). This drawback is especially relevant in

scenarios where numerous amplification targets are likely to be detected at once or in patients with elevated pretest probability of signal amplification, such as in UTI (45). Therefore, for the near future, new BI diagnostics are likely to complement traditional methods rather than supersede them (45, 55).

## 1.3 Statistical Modeling of Bacterial Infections

Leveraging electronic health record (EHR) data with statistical learning techniques offers a low-cost opportunity to supplement existing BI detection methods to help guide antibiotic therapy decisions in the ICU. The task of managing infection in the ICU is uniquely well suited to benefit from clinical informatics research because it involves highly complex, time-sensitive decision making in uncertain conditions and utilizes information collected concurrently from numerous medical specialties. Furthermore, the adoption and use of EHR systems continue to rise, providing investigators access to massive repositories of both structured and unstructured data generated through routine clinical care (56-58). Given the potential of EHR data to inform clinical decision-making, considerable work has been invested into developing predictive models for managing infections in the ICU.

The literature on predictive models for managing infections has placed a large focus on predicting the occurrence of sepsis and its associated outcomes (59-74). The performances of such models have been found to vary considerably. A review conducted by Fleuren et al. highlighted that among the 130 models evaluated, the AUROCs ranged from 0.68 to 0.99 in the ICU (61). Only 7% of these models were prospectively validated, and only 11% were successfully implemented into clinical practice, yielding mixed result. Additionally, hundreds of hospitals across the country using the Epic EHR system have adopted Epic's proprietary sepsis model (75). Despite its popularity, the model has drawn criticism for having minimal data and

algorithmic transparency, and for achieving a 67% false negative rate in a recent independent external validation study (76, 77).

Moving beyond sepsis forecasting, numerous models have been proposed to improve other aspects of BI diagnosis, management, and antibiotic stewardship efforts. Hwang et al. suggested a deep learning model that works in conjunction with clinicians to improve the interpretation of chest X-rays (78). Meanwhile, Lamping et al. introduced a machine learning approach that distinguishes between inflammation and bacterial infection, which surpassed host response biomarkers in internal validation (79). Pathogen resistance phenotyping and genotyping workflows have also been improved with machine learning models. Roux-Dalvai et al. developed a workflow that can identify bacteria responsible for 84% of UTI in just four hours without requiring culture, using advanced liquid chromatography with tandem mass spectrometry techniques and machine learning models (80). Previous studies have also proposed machine learning models to support antibiotic stewardship efforts by flagging cases for antimicrobial prescription review and prescriber feedback (81, 82). These models demonstrate both the potential for machine learning techniques to enhance antibiotic stewardship efforts and the need for more external validation and clinical impact studies. Currently no machine learning models exist to help safely guide antibiotic de-escalation decisions in critically ill patients who are at low risk for having a bacterial infection. To build a robust and useful BI risk stratification model, it's important to adequately address the barriers that exist for developing and implementing prediction models into clinical practice.

## 1.4 Barriers to EHR Predictive Modeling

Researchers face numerous different obstacles when attempting to extract and prepare EHR data for secondary use applications such as developing clinical prediction models (Figure

2). EHR data contains protected health information and must be handled carefully to safeguard patient privacy, as per the Health Insurance Portability and Accountability Act (HIPAA). Additionally, every EHR system has a unique profile of data quality challenges resulting from variations in data completeness, accuracy, consistency, accessibility, and timeliness (83-87). Inconsistencies in semantics and coding standards between hospital systems, combined with data quality issues, pose further interoperability hurdles that complicate the linkage of disparate datasets (88, 89). Moreover, since a significant portion of EHR data is stored in unstructured or semi-structured text reports, specialized informatics tools are often necessary to extract information for secondary use (88, 90-92). However, even with the growing library of published data preprocessing tools, custom solutions are currently required to tackle the unique challenges associated with each EHR-based dataset (93). Given that the quality and biases of data fed into a model are inevitably reflected in its outputs, it is recommended that researchers adhere to established data preparation workflows and meticulously document coding decisions to enhance transparency (93-95). Once a model is developed, researchers then must overcome barriers involved in establishing trust in their model.

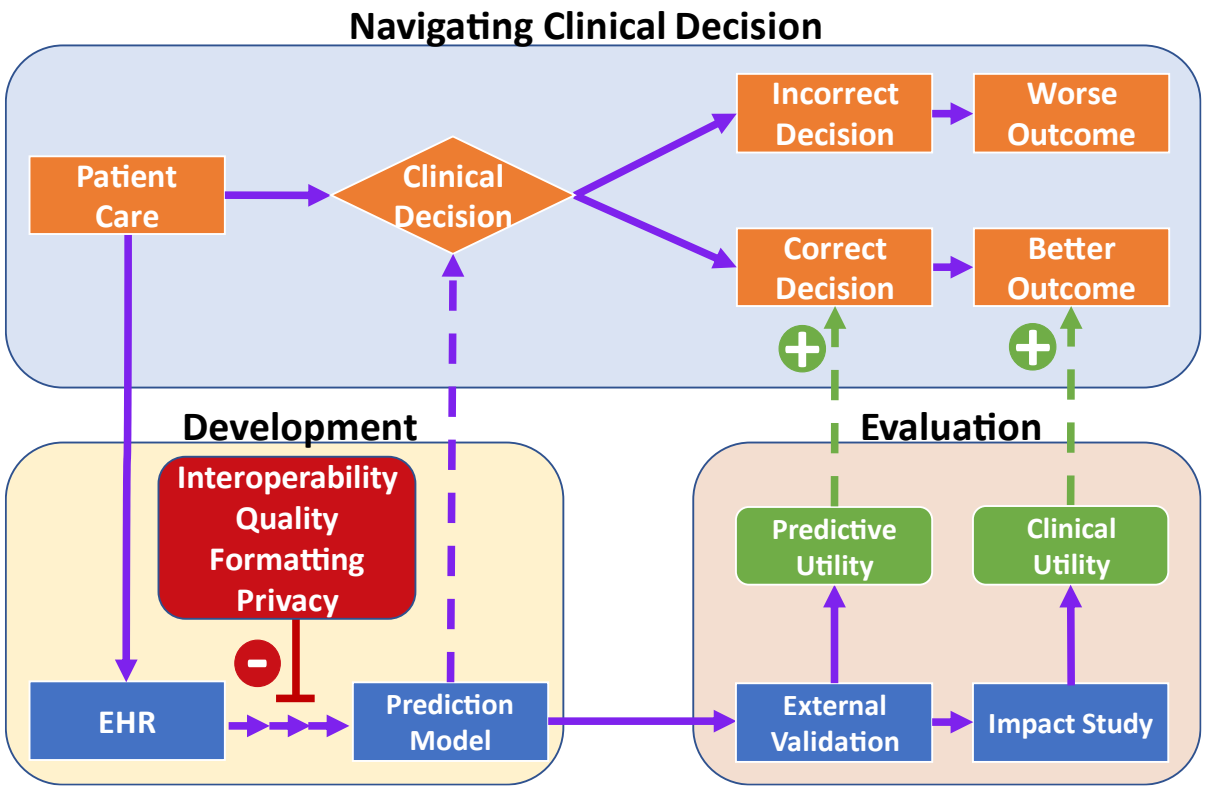


*Figure 2. Interplay between clinical prediction models and clinical decision making. Clinical prediction models are most useful when they are designed to output information that can be considered when caregivers are making a clinical decision. Most models do not get beyond the initial development stages due to issues with such as model design or challenges associated with secondary use of EHR data. Before implementing a developed mode, researchers must evaluate the model to ensure its predictions continue to provide utility in external populations. Predictive utility is necessary but not sufficient to demonstrate clinical utility. Once externally validated, a model should be tested in its intended clinical context to measure the impact on patient care.*

The British statistician George E. P. Box is attributed with the saying "All models are broken, but some models are useful". In a clinical setting, the potential for a broken and non-useful model to cause iatrogenic harm is vast. Therefore, it is crucial to gather ample evidence supporting the predictive and clinical utilities of a model before adopting it into clinical practice (Figure 2). The predictive utility of a model is a measure of the robustness and transportability of its predictions and is commonly assessed through external validation (96-99). By measuring model performance across external but plausibly similar cohorts, potential imperfections such as

overfitting or bias can be identified and subsequently addressed. Transparent reporting of data and algorithms can further these benefits by enabling others to independently debug and assess model performance across different scenarios (99). Studies such as the one by Wong et al., which performed an external independent validation of the Epic EHR system sepsis model, present a warning that even closed-source models developed with ample resources can demonstrate poor generalize poorly to new patient populations (76). Unfortunately, reviews suggest that only around 10% of published models present any form of external validation (100). These numbers highlight how barriers to secondary use of EHR data can also impact downstream validation studies.

While demonstrating predictive utility is a crucial step towards model implementation, it does not guarantee clinical utility at the bedside (Figure 2). To assess clinical utility, a model's impact on health outcomes and decision making must be quantified through comparison studies. Traditionally, prospective randomized studies have been used, but observational studies leveraging causal inference methods and EHR data have become increasingly popular, primarily due to lower cost and development time barriers (101-104). Published standards such as the TRIPOD Statement have been introduced to help standardize methods and facilitate transparency of model development, validation, and impact studies (99, 103, 105). Recently, models that have been coupled with optimal real-world interventions have demonstrated improved clinical utility (106).

Watson et al. recently conducted a survey with 33 healthcare and informatics leaders on how to overcome the biggest barriers to adopting predictive modeling into clinical care where respondents called upon the informatics community to: (1) development of robust tools and evaluation methodologies and (2) development and dissemination of best practices (107).

Furthermore, designing each of the outputs of clinical prediction models to have a corresponding evidence-based intervention presents a promising route to improve the clinical utility of prediction models (104, 106, 108).

## 1.5 Objectives

The primary focus of this thesis is to develop and validate a transportable prediction model framework that will help guide antibiotic de-escalation decisions in the ICU and improve healthcare outcomes. An additional focus of this thesis is to create innovative tools for solving informatics problems, while emphasizing the importance of proper evaluation and validation at every step. In the series of studies reported herein, the following hypotheses were tested: **(H1)** there is sufficiently granular data in EHRs to accurately predict patient-level BI risk using raw clinical time series data of structured clinical variables; **(H2)** this model framework will be transportable to external validation cohorts; **(H3)** pairing the BI risk model with evidence based antibiotic de-escalation will provide clinical utility because patients with low BI risk will have worse outcomes when given prolonged vs. short empiric antibiotic therapy. **Chapters 2-3** test hypothesis **(H1)** and focus on model development topics as depicted in Figure 2. Here, we introduce novel open-source informatics tools to address data roadblocks associated with microbiology culture reports and present the development and optimization of a BI risk modeling framework. In **Chapters 4-5**, we shift the focus towards model evaluation (Figure 2) and demonstrate the predictive and clinical utility of our BI risk model with external validation **(H2)** and clinical impact **(H3)** studies.

# 2. PREDICTIVE MODELING OF BACTERIAL INFECTIONS AND ANTIBIOTIC THERAPY NEEDS IN CRITICALLY ILL ADULTS

This work was published as:

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.

## 2.1 Introduction

Antibiotics can be lifesaving for critically ill patients with bacterial infections (BIs), however, overuse or unnecessary administration can contribute to antimicrobial resistance (AMR) and antibiotic-associated morbidity (26-28, 42, 109-111). This is a critical issue, as patients with AMR infections suffer longer hospital stays, treatment complications, higher healthcare costs, and are more likely to die (112-115). Furthermore, antibiotics can cause harm through gut microbiome dysbiosis, mitochondrial toxicity, and immune cell dysfunction (26-28, 42, 109-111). Although clinicians have become more aware of the side effects of antibiotics, it is estimated that up to 50% of antibiotic prescriptions in acute care hospitals in the United States are still either inappropriate or unnecessary(15-20). Reducing both the amount and duration of unnecessary antibiotic treatments is a commonly proposed strategy to reduce the risk of antibiotic-related side effects(15-18, 116). This is particularly relevant in the intensive care unit (ICU), where concerns for bacterial infections (BI) are high and prescribing antibiotics empirically—prior to having confirmatory bacterial culture results or when an occult BI is suspected—is a common practice(1, 117).

Approximately 30-50% of all ICU patients are diagnosed with a BI and their mortality rates can reach as high as 60% in severe infections (1-4). As a result, providers in the ICU often have a low threshold to start empiric antibiotic therapy (EAT) despite the ramifications of excessive antibiotic use for patients at low risk of BI. Unfortunately, there is no uniform consensus on the appropriate duration of EAT. As a result, clinicians must continually weigh the risks of failing to treat a serious BI against the risks of prescribing inappropriate antibiotic regimens. Moreover, physicians lack objective criteria to identify low BI risk in patients receiving EAT and rely on clinical intuition and imprecise guidelines to balance EAT decisions

(42, 118-120). Strategies that shorten unnecessary antibiotic duration in ICU patients when BIs are no longer suspected offer a way to improve patient outcomes, and have been identified as a priority by the Society of Critical Care Medicine as part of their "less is more" campaign (121).

Leveraging electronic health record (EHR) data with machine learning techniques presents an opportunity to accurately identify patients with low risk of BI. The widespread adoption of EHR systems offers investigators access to massive repositories of data generated through routine clinical care and provides opportunities to develop novel prediction algorithms to aid in clinical decision making.

The primary objective of this study was to develop a novel framework to identify ICU patients with a low risk of BI as candidates for earlier EAT discontinuation. The feasibility of this approach was investigated in patients suspected of having a BI by modeling data collected for up to 24-, 48- or 72-hours following the first dose of antibiotics. We compare prediction performance across different model types, data collection windows, and prediction thresholds. The developed algorithm could be used to identify patients at low risk of BI early in their hospitalization who may benefit from early discontinuation of EAT. Furthermore, our EHR-based phenotype of patients suspected of having a BI could be generalized to other datasets and used for additional analyses on antibiotic usage and BI in the ICU.

The detailed data dictionary, code, and results have been made available at: https://github.com/geickelb/mimiciii-antibiotics-opensource.

## 2.2 Materials & Methods

### 2.2.1 **Dataset**

A summary of our data extraction and analysis workflow is presented in Figure 3 The data used in this study was retrieved from the Medical Information Mart for Intensive Care III (MIMIC-III). The MIMIC-III database is an open and de-identified database comprised of health-related data from over 40,000 ICU patients who received care at Beth Israel Deaconess Medical Center between 2001 and 2012 (122, 123). MIMIC-III includes a variety of data such as administrative, clinical and physiological types, which are organized, formatted, processed and de-identified in accordance with the Health Insurance Portability and Accountability Act (HIPAA) guidelines (122, 123).

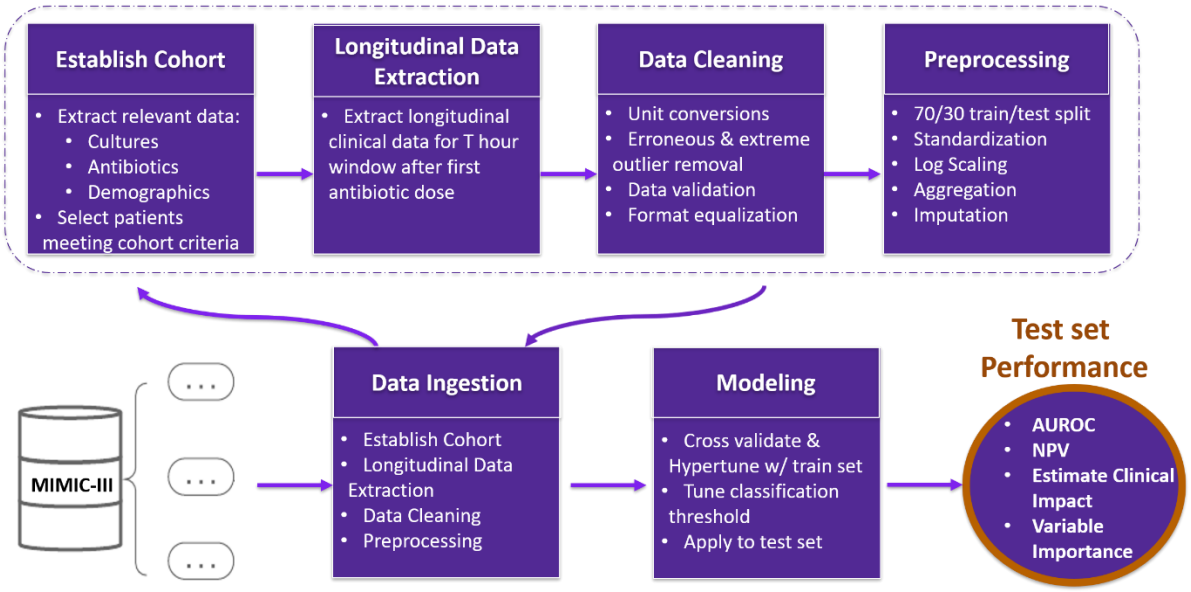


*Figure 3. Data ingestion and analysis framework overview. Raw data is ingested from the MIMIC-III database. First a cohort of adult patients suspected of having SBI is established, and both longitudinal and categorical data is extracted over the T= 24-,48-, or 72-hour window following their first antibiotic dose that corresponds with a microbiologic culture. Next, data is cleaned, formatted, and preprocessed prior to modeling. The cohort is then filtered to patients with positive microbiologic culture and prolonged antibiotics, and microbiologic culture negative with short antibiotics. A 70/30 train/test set split is then applied. Scaling and standardization are performed on each set independently. Missing values were imputed using median values from the training set. Machine learning models are hypertuned on the training set and applied to the test set. Finally, classification thresholds are tuned, and model performance metrics are output.*

### 2.2.2 **Cohort**

Adult patients who were suspected of having a BI upon admission to the ICU were eligible for our study. To match this phenotype, a patient must have: (1) received at least one dose of antibiotics within 96 hours following ICU admission and (2) had a microbiologic culture within 24 hours of their first antibiotic dose (Figure 4). Microbiologic cultures were defined as cultures obtained from any of the following: blood, joint, urine, cerebral spinal fluid (CSF), pleural cavity, peritoneum, or bronchoalveolar lavage. Patients with multiple ICU encounters

that met study inclusion criteria were analyzed independently; however, each patient's ICU encounters were assigned to the same train/test split (see *Modeling*).

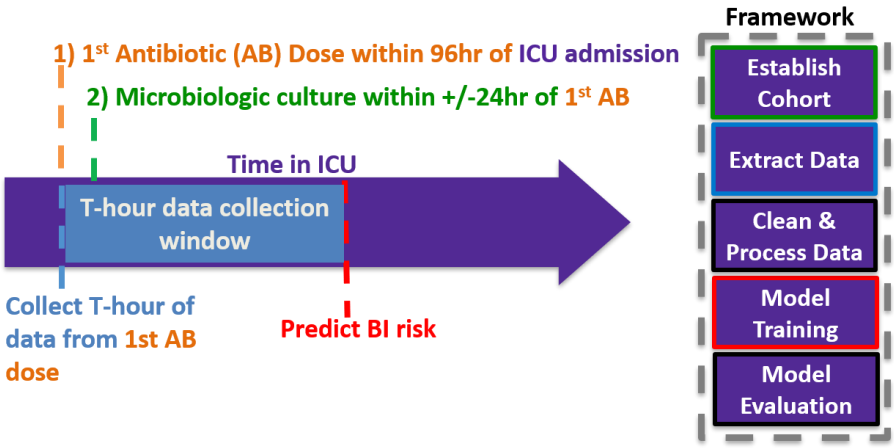


*Figure 4. Phenotype criteria for BI suspicion at ICU admission. A patient's first antibiotic (AB) dose ($t_0$) needs to: (1) be administered within 96 hours following ICU admission and (2) have an microbiologic culture within 24 hours and (1) be administered within 96 hours following ICU admission. Clinical Data is collected for up to T= 24-, 48-, or 72-hours after first antibiotic dose.*

Antibiotics prescriptions were recorded as the administration of any "antibacterial for systemic use" represented by Anatomical Therapeutic Chemical (ATC) code J01. ATC codes were obtained by first converting national drug codes (NDC) into RxNorm concept unique identifier (RXCUI) codes, and then into ATC codes. Regular expressions were used on prescription names to further filter out erroneous entries and those with missing NDC/RXCUI codes. We calculated the maximum length of cumulative antibiotic days following a microbiologic culture for each ICU encounter. Prescription information in the MIMIC-III database was stored with date level resolution. To accommodate this, the time of each patient's first antibiotic dose ($t_0$) meeting the phenotype criteria was set to 0:00:00.

Patients were allocated to one of three BI groups: serious BI, non-serious BI/no BI, and unknown BI status (Figure 5). Given the common occurrence of occult bacterial infections, a

direct inference of BI status could not be made based off microbiological culture results alone. Therefore, patient's BI statuses were assigned based both on their microbiologic culture results (positive vs. negative) and duration of their antibiotic treatment (short [≤96 hours] vs. prolonged [>96 hours]). In this paradigm, patients with positive microbiologic culture and prolonged antibiotic treatment were considered to have serious BIs (prediction events), whereas those with negative cultures and short antibiotic treatment were considered to have no BIs (prediction non-events). Additionally, patients with short antibiotic treatment and positive microbiologic culture were considered non-serious BIs. Due to the possibility of occult infections, patients who received prolonged antibiotics despite having a negative microbiologic culture had less clear infection statuses, and were thus coded as unknown BI status. Conceptually, patients in that group could be further divided into those with an occult serious BI and those with either no BI or an occult non-serious BI. These patients were separated from the dataset prior to model training and testing, and were later used assess the clinical utility of the prediction model by testing its ability to identify patients at low-risk BI in that population.

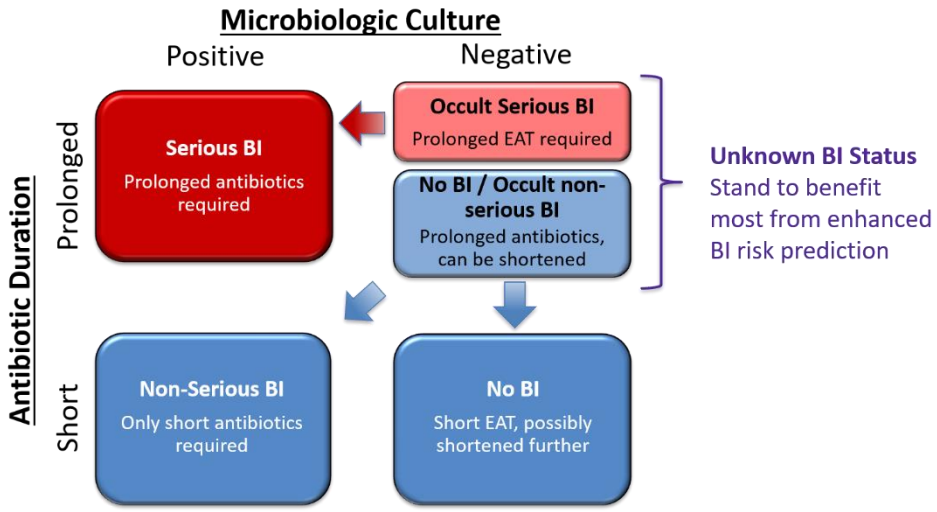


*Figure 5. Classification of BI status and framing of the clinical prediction problem. Patient BI status can be classified into three groups based on duration of antibiotics and microbiological results: "Serious BI" are those with positive microbiological cultures receiving antibiotics for >96 hours and are the cases in model training. "Non-serious BI" and "No BI" patients are those with antibiotics ≤96 hours and are the controls in the model training. "Unknown BI status" are patients who received empiric antibiotic therapy [EAT] for >96 hours despite negative microbiological cultures and are the group of patients most likely to benefit from correct BI risk prediction. The unknown BI status group may be conceptually divided into patients with "occult serious BI" and patients with "no BI or occult non-serious BI".*

To control for *Staphylococcus* culture contamination, we required two consecutive *Staphylococcus* positive cultures to be considered microbiologic culture positive. Additionally, we coded patients that died within 24 hours of their last antibiotic dose as prolonged antibiotic treatment (n=1266). To accommodate for date-level resolution on prescription timings, we utilized a conservative 96-hour threshold for short vs prolonged antibiotic duration.

### 2.2.3 **Data Extraction**

We extracted static and longitudinal patient clinical data from the MIMIC-III database using open-source code provided by the MIMIC-III team (Table 1). Longitudinal data was restricted to either the T= 24-, 48-, or 72-hour cutoff following the administration date of the first antibiotic dose ($t_0$: $t_{0+T}$) (Figure 4).

### 2.2.4 **Cleaning & Pre-processing**

The raw clinical data extracted for the purpose of this study were first cleaned and formatted to address data quality issues and then preprocessed to facilitate usability by selected machine learning models. The first cleaning step was to address desperate units of measurement by converting each variable into designated units (Table 1). Next, conservative thresholds were set to review erroneous values and data entry errors for removal based upon a combination of reference laboratory value limits, clinical knowledge, three sigma outlier criteria, and manual audit of a subset of free-text to confirm concordance. Finally, event and windowed continuous variables, such as administration of renal replacement therapy or mechanical ventilation were coded and discretized. The cleaned data were then converted into unit variances following as in (Equation 1), where $\tilde{X}_{(-/short)}$ is the median value of the patients with negative microbiologic culture and short duration EAT. Next, longitudinal and ordinal clinical variables spanning $t_0$: $t_{0+T}$ were aggregated to produce single value(s) for each parameter using the operation that conferred the highest likelihood of infection (minimum, maximum or both). Lastly, categorical variables were encoded to dummy variables using the one-hot-encoding technique. The final dataset was represented by a 52-dimension feature vector.

*Equation 1. Median based Z-score equivalent.*

$$Z = \frac{X - \widetilde{\mathrm{X}}_{(-/short)}}{IQR_{(-/short)}}$$

*Table 1. Extracted data and missingness- Raw variables and units extracted from the corresponding table in the MIMIC-III database.*

| ***MIMIC-III TABLE*** | ***Data Collected*** | ***Unit*** | ***% Missingness (T= 24-72 hour)*** |
|---|---|---|---|
| **Diagnoses** | ICD-9 codes (Elixhauser Comorbidity Index) | categorical | 0 - 0 |
| **Admissions** | Age | years | 0 - 0 |
| | Race & Ethnicity | categorical | 0 - 0 |
| | Gender | categorical | 0 - 0 |
| **ChartEvents** | Blood pressure (systolic, diastolic) | mmHg | 0.2 - 0 |
| | Glasgow Coma Scale | GCS score | 72.5 - 53.3 |
| | Glucose | mg/dL | 0.5 - 0.1 |
| | Heart rate | bpm | 0 - 0 |
| | Peripheral oxygen Saturation (SpO2) | % | 0 - 0 |
| | Temperature | deg. C | 1.6 - 0.2 |
| | Ventilation status | categorical | 1.3 - 0.9 |
| | Weight | kg | 8.4 - 8.4 |
| **InputEvents** | Dobutamine | mcg/kg/min | 98.8 - 98.3 |
| | Dopamine | mcg/kg/min | 94.9 - 94 |
| | Epinephrine | mcg/kg/min | 97.9 - 97.6 |
| | Norepinephrine | mcg/kg/min | 83.1 - 80.1 |
| | Phenylephrine | mcg/kg/min | 86.2 - 83.2 |
| | Renal replacement therapy | pos/neg | 0 - 0 |
| | Vasopressin | mcg/kg/min | 98 - 97 |
| **LabEvents** | Bands | % | 87.3 - 82.6 |
| | Serum bicarbonate | mEq/L | 2.1 - 0.3 |
| | Bilirubin | mg/dL | 60.1 - 47.8 |
| | Blood urea nitrogen (BUN) | mg/dL | 2 - 0.3 |
| | Serum chloride | mEq/L | 1.9 - 0.3 |
| | Serum creatinine | mg/dL | 2 - 0.3 |
| | Serum glucose | mg/dL | 0.5 - 0.1 |
| | Hemoglobin | g/dL | 2.6 - 0.3 |
| | International Normalized Ratio (INR) | ratio | 24.9 - 13.3 |
| | Serum lactate | mmol/L | 48.1 - 42.3 |
| | Urine leukocyte | pos/neg | 69.5 - 57.6 |
| | Urine nitrite | pos/neg | 69.5 - 57.6 |
| | Partial pressure of arterial oxygen (PaO2)/fraction of inspired oxygen (FiO2) ratio | ratio | 67.9 - 65.1 |
| | Partial thromboplastin time (PTT) | sec | 25.2 - 13.8 |
| | Partial pressure of arterial carbon dioxide (pCO2) | mmHg | 39.9 - 34 |
| | Serum pH | n/a | 41.9 - 36.7 |
| | Platelet count | K/uL | 2.6 - 0.3 |
| | Serum potassium | mEq/L | 1.6 - 0.3 |
| | White blood cell count | K/uL | 2.9 - 0.3 |
| | Serum calcium | mmol/L | 63.1 - 56.6 |

### 2.2.5 **Modeling**

The patients with positive microbiological cultures and prolonged antibiotic duration (serious BI) and those with short antibiotic duration (no BI or non-serious BI) were split into a training and test set following a 70/30 split based upon unique ICU stay identifiers. Cohort splitting was performed on unique ICU stay identifiers where individual patients were sequestered to either the training or testing set to prevent testing set contamination. We chose to impute missing values with median values from the training set to facilitate implementation into a clinical setting. Empirical studies have suggested that including imputed values with high missingness can improve model clinical utility, so we chose to include imputed values with high missingness in our model (Table 1) (124, 125).

The final dataset was modeled using a variety of machine learning algorithms, including Ridge regression (126), Random Forests (127), support vector classifier (SVC) (128), extreme Gradient Boosted decision Tree (XG Boost) (129), K-Nearest Neighbors, and Multilayer Perceptron (MLP). These models were chosen using a set of criteria that included each model's relative interpretability, approach to handling nonlinearity, and ability to model categorical and continuous features. A soft voting classifier, or ensemble of all other models, was also used to test for significant performance gains or losses.

Class imbalance was addressed by classification threshold tuning and modeling specific class balancing parameters, such as bootstrapping and class weights, during hyperparameter tuning to simplify the modeling workflow. Modeling hyperparameters were tuned using 10-fold cross validation with a binary cross entropy loss function on the training set. The binary classification threshold was tuned in 10-fold cross validation to achieve a high sensitivity (sensitivity $\geq 0.9$) and was averaged across all folds. This high sensitivity was chosen to reduce

the number of false negatives and predict low BI risk with higher certainty. Threshold tuned model performances were assessed on the test set using area under the receiver operator curve (AUC), F1 score, negative predictive value (NPV), precision, and recall.

## 2.3 Results

### 2.3.1 **Cohort**

We identified a total of 19,633 ICU encounters (15,412 unique patients) in the MIMIC-III data that met inclusion criteria for our study. Within this set, we filtered our cohort down to 12,232 ICU encounters (10,290 unique patients) that had either prolonged antibiotics and positive microbiologic culture, or short antibiotics and negative microbiologic culture (Table 2). Table 3 summarizes the breakdown of these patients across the train/test splits. Additionally, 7,401 ICU encounters (6,520 unique patients) with unknown BI status (prolonged antibiotics and negative microbiologic culture) were set aside to test the prediction model's ability to identify patients at low-risk BI in that population.

*Table 2. Distribution of cohort demographics.*

| ***Variable*** | ***Count*** |
|---|---|
| Gender- N, % | |
| Female | 5709 (47%) |
| Male | 6523 (53%) |
| Age in years (stdev.) | 64.7 +\- 17.0 |
| Race & Ethnicity- N, % | |
| Black/non-Hispanic | 1385 (11%) |
| White/non-Hispanic | 8855 (72%) |
| Hispanic | 507 (4%) |
| Other | 1485 (12%) |

a. SD denotes Standard Deviation

*Table 3. Cohort distribution of BI status classifications.*

| ***Microbiologic Culture*** | ***Antibiotic Duration[a]*** | ***BI Status Classification*** | ***Train No. (%)*** | ***Test No. (%)*** | ***Total No. (%)*** |
|---|---|---|---|---|---|
| Negative | Short | Negative | 5512 (65%) | 2355 (65%) | 7867 (65%) |
| Positive | Prolonged | Positive | 1693 (20%) | 745 (20%) | 2438 (20%) |
| Positive | Short | Negative | 1296 (15%) | 631 (15%) | 1927 (15%) |
| Negative | Prolonged | Unknown | N/A | N/A | 7401 (100%) |

a. Time on antibiotics, short (≤96 hours) vs. prolonged (>96 hours)

*Table 4* summarizes the test set results for each threshold tuned model. The performance across the models for each T-hour test set showed little variation, where XGBoost and Random Forests slightly outperformed the other models in terms of AUC, F1 score, NPV, and precision. As the data window was increased from 24 to 72 hours, there were small increases in AUC across the best performing models for each time window. Figure 6 summarizes the ROC curve for all the T=24-hour models where all, except K-nearest neighbors, performed similarly. Additionally, when tested with the 72-hour test data, the 24-hour Random Forests model obtained an AUC of 0.787 (~0.013 increase). Similarly, the 72-hour Random Forests model produced an AUC of 0.765 (~0.028 decrease) when tested on the 24-hour data. These changes in AUC suggest that both the 24-hour and 72-hour models maintain similar model performances when making predictions on data collected over 48-hour longer and shorter collection windows, respectively.

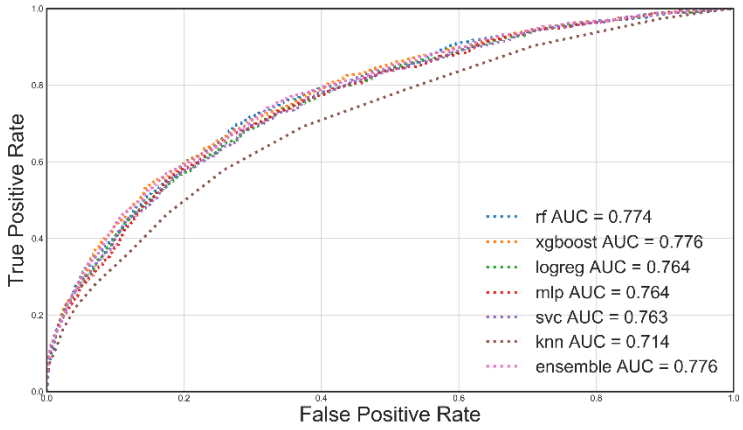


*Figure 6. Receiver operating characteristic curves (ROC) for all T=24-hour models on 24-hour Test set data. We use different colors and line styles to differentiate models. AUC: Area under the curve.*

Table 4. *Preliminary model results - Modeling parameters for each model on the test set using the high sensitivity threshold.*

| ***Model*** | ***AUC*** | ***F1*** | ***NPV*** | ***Precision*** | ***Recall*** | ***High Sensitivity Threshold*** |
|---|---|---|---|---|---|---|
| ***72-hour Test set*** | | | | | | |
| **Random Forests Classifier** | 0.793 | 0.431 | 0.941 | 0.284 | 0.891 | 0.124 |
| **XGBoost** | 0.795 | 0.439 | 0.943 | 0.291 | 0.891 | 0.096 |
| **MLP Classifier** | 0.779 | 0.395 | 0.948 | 0.25 | 0.936 | 0.09 |
| **Logistic Regression** | 0.781 | 0.423 | 0.932 | 0.278 | 0.876 | 0.298 |
| **SVC** | 0.778 | 0.425 | 0.935 | 0.28 | 0.881 | 0.101 |
| **K-NN** | 0.734 | 0.357 | 0.936 | 0.219 | 0.963 | 0.04 |
| **Voting Classifier** | 0.793 | 0.429 | 0.946 | 0.281 | 0.905 | 0.147 |
| ***48-hour Test set*** | | | | | | |
| **Random Forests Classifier** | 0.788 | 0.43 | 0.943 | 0.283 | 0.897 | 0.126 |
| **XGBoost** | 0.796 | 0.436 | 0.946 | 0.288 | 0.9 | 0.091 |
| **MLP Classifier** | 0.771 | 0.456 | 0.92 | 0.318 | 0.805 | 0.084 |
| **Logistic Regression** | 0.774 | 0.421 | 0.938 | 0.275 | 0.893 | 0.296 |
| **SVC** | 0.773 | 0.42 | 0.941 | 0.274 | 0.9 | 0.099 |
| **K-NN** | 0.733 | 0.393 | 0.922 | 0.252 | 0.887 | 0.044 |
| **Voting Classifier** | 0.788 | 0.436 | 0.939 | 0.29 | 0.881 | 0.147 |
| ***24-hour Test set*** | | | | | | |
| **Random Forests Classifier** | 0.774 | 0.424 | 0.944 | 0.277 | 0.905 | 0.258 |
| **XGBoost** | 0.776 | 0.416 | 0.94 | 0.271 | 0.901 | 0.104 |
| **MLP Classifier** | 0.764 | 0.439 | 0.925 | 0.297 | 0.84 | 0.087 |
| **Logistic Regression** | 0.764 | 0.411 | 0.94 | 0.266 | 0.907 | 0.302 |
| **SVC** | 0.763 | 0.411 | 0.937 | 0.267 | 0.9 | 0.105 |
| **K-NN** | 0.714 | 0.382 | 0.922 | 0.243 | 0.903 | 0.044 |
| **Voting Classifier** | 0.776 | 0.421 | 0.939 | 0.275 | 0.895 | 0.177 |

Figure 7 displays how variable importance changed across the models. For this plot, a list of 20 variables was selected based on the top ten most important variables for the Random Forests, logistic regression, XGBoost, and SVC models. Figure 7 suggests that although the models perform similarly, each model prioritized predictors. This interpretation is reinforced by the results of the soft voting ensemble models, which performed comparably to the best performing model within each T-hour test set. This further suggests that the models are identifying the same or similar patients regardless of the underlying algorithm.

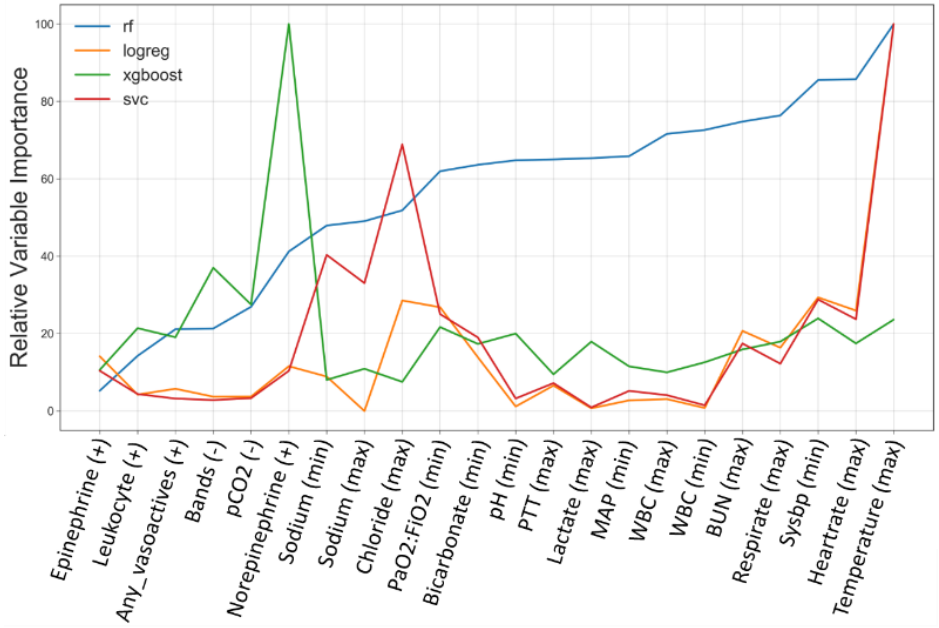


*Figure 7. Stacked relative variable importance across prediction models. Variable importance for Random Forests and XGBoost were based on standardized Gini importance, while SVC and logistic regression used standardized coefficients. Variable importance values from all models were scaled relative to the value of the most important variable for all 20 values in the variable list. $pCO_2$: carbon dioxide partial pressure; PaO2:FiO2: ratio of arterial oxygen partial pressure to fractional inspired oxygen; PTT: platelets; MAP: average arterial blood pressure over one cardiac cycle; WBC: white blood cell count; BUN: Blood urea nitrogen; sysBP: Systolic blood pressure*

The T=24-hour Random Forests model was chosen for the subsequent analyses given that the T=24-hour timepoint provides more clinical utility, and thus the Random Forests model was

the best performing model within this timepoint. Table 5 summarizes the confusion matrix for this model with a high sensitivity classification threshold (0.26) in the test set. The model achieved an NPV of 0.944 in the test set; however, this figure is based on the 0.20 BI prevalence from the training and testing set. Figure 8 displays how the model NPV changes as a function of population BI prevalence and classification threshold. We found that as the BI prevalence changed from 0.5 to 0.1, the NPV of the T=24-hour random forests model changed from 0.82 to 0.98 when using a high sensitivity threshold, and 0.59 to 0.93 when using a 0.5 threshold. These results suggest that our model performance will be more robust to changes in prevalence when using the high sensitivity prediction threshold. The remaining patients falsely predicted as negatives by all of the T=24-hour models were investigated for observable patterns. These investigations suggested that the false negatives are a heterogeneous group with no reproducible patterns.

*Table 5. Confusion matrix statistics- Test set classification summary for the T=24-hour Random Forests model with a high sensitivity threshold.*

| | **True Negatives (%)** | **False Positives (%)** | **False Negatives (%)** | **True Positives (%)** |
|---|---|---|---|---|
| **High Sensitivity Threshold** | 1208 (32.4%) | 1773 (47.5%) | 71 (1.9%) | 679 (18.2%) |
| **0.5 Probability Threshold** | 2826 (75.7%) | 155 (4.2%) | 521 (14.0%) | 229 (6.1%) |

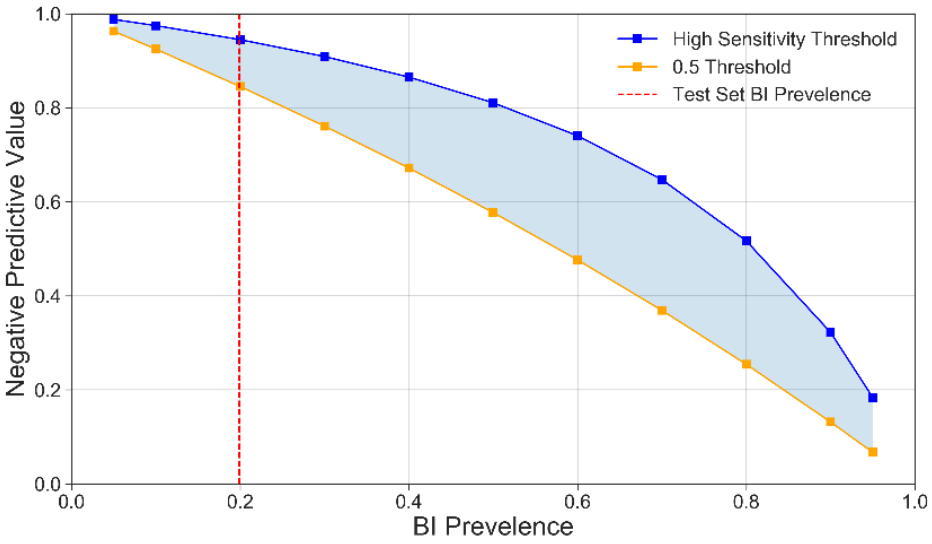


*Figure 8. NPV across BI prevalence for T=24-hour Random Forests tuned and 0.5 prediction thresholds. NPV was simulated for a variety of BI prevalence values using the sensitivity and 1-specficity for the high sensitivity and 0.5 prediction thresholds from the test set.*

Of the 1208 true negatives, 458 (37.9%) cases received antibiotics for 24 hours or less, while 750 (62.1%) received antibiotics greater than 24 hours. We estimated that 1,289 out of the 2,375(53.2%) total antibiotic days administered to patients in the true negative group could have been avoided if our model with a high sensitivity threshold were hypothetically used to stop EAT early.

### 2.3.2 **Performance in patient set with unknown BI**

Finally, the best performing T=24-hour Random Forests model was applied to the patient group with unknown BI status, which are those who stand to benefit the most from correct BI risk prediction. Using the high sensitivity and 0.5 probability thresholds, the model predicted 861 out of 7,401 (11.6%) and 5,525 out of 7,401 (74.7%) patients to be at low risk of BI, respectively (Table 6). Using the NPV from the test set with high sensitivity and 0.5 thresholds (NPV=94.5%, 84.3%) we estimated that approximately 48 (0.6%) and 860 (11.6%) of all unknown BI status

patients would have been predicted to have a low BI risk but actually have had a BI (false negatives). By subtracting these estimated false negative patients from the total negative predictions, we estimated that the high sensitivity and 0.5 thresholds would have theoretically benefited 813 (11.0%) and 4,664 (63.0%) patients, respectively. We estimated that as a lower bound, our T=24-hour Random Forests model with a high sensitivity threshold could have reduced approximately 5,684 (9.5%) antibiotic days administered to this group, and as an upper bound with the 0.5 probability threshold could have reduced approximately 35,831 (60.0%) antibiotic days. A manual chart review and clinical assessment of 10 patient records with unknown BI status (5 predicted high BI risk, 5 predicted low BI risk) found that 8 out 10 model BI risk classifications matched the clinical reviewer's assessment of BI risk, 2 out of 10 were probably correct but remained indeterminate, and 0 out of 10 were misclassified (Supplementary material A).

*Table 6. Prolonged antibiotic negative microbiologic culture predictions- Prediction distribution for the T=24-hour Random Forests model.*

| | **High Sensitivity Threshold** | **0.5 Threshold** |
|---|---|---|
| **Predicted low BI risk** | 861 (11.6%) | 5525 (74.7%) |
| **Predicted high BI risk** | 6540 (88.4%) | 1876 (25.3%) |

## 2.4 Discussion

In this study, we developed a novel framework to extract patient features from raw clinical data and identify patients at low risk of BI who, in theory, could benefit from earlier EAT discontinuation within 24 hours of initiation. Our main finding is that our models can predict patients with low risk of BI with good performance when applied to structured clinical data collected for T= 24-hours after the EAT initiation. We also found that increasing the data collection time and model complexity yielded only slight performance increases. Finally, our

results suggest by that applying our T=24-hour Random Forests model with a high sensitivity threshold to the patient set with unknown BI status (prolonged antibiotics and negative microbiologic culture), we would be able to identify around 11.6% of patients as candidates for EAT removal with high confidence and could reduce total antibiotic days by approximately 9.5%.

Designing data-driven approaches to accurately stratify patients based on their BI risk has the potential to greatly improve antibiotic stewardship efforts. Antibiotic stewardship in the ICU can be viewed as a two-stage process. The first stage requires administering broad-spectrum antibiotics to maximize treatment of serious BI. In the second stage, physicians either stop EAT for patients at low risk of BI or narrow the spectrum of antibiotics once the infection is characterized (42). Many stewardship techniques focusing on the later stage hinge upon sensitive and specific identification and monitoring of BI risk. Bacterial cultures and inflammatory biomarkers are currently the most common methods of monitoring BI risk in the ICU but are not necessarily optimal. Bacterial cultures, the current gold standard for diagnosing BI, may take days to result and are often unreliable in detecting all Bis (43). To address this, bacterial cultures are frequently supplemented with Gram staining, which provide additional information more immediately about a patient's BI risk. However, Gram staining suffers from high variability and low reliability that results from individual differences in slide preparation and interpretation (130-132). Assays based on inflammatory biomarkers, such as C-reactive protein and procalcitonin, have improved sensitivity and specificity for detecting community-acquired infections, but have high rates of false-positives and -negatives for hospital-acquired infections(42, 133-135). Newer rapid multiplex diagnostics for infectious organisms have also been introduced; however, these are still being tested for efficacy, costly, and not yet widely

available(136). Designing better methods to identify patients with low risk for BI is critical to shorten the duration of unnecessary EAT and facilitate antibiotic stewardship.

Numerous prior studies have presented EHR-based machine learning models and clinical decision support systems to predict infection related conditions, such as bacteremia, sepsis, and ICU mortality(62-69, 137-140). The goal of such models has been to ensure all septic and/or bacteremic patients are identified and treated early with appropriate antibiotic regimens(62-69). For instance, Nemati et al. achieved AUROCs ranging from 0.83-0.85 in predicting the early onset of sepsis using data collected during the 12, 8, 6, and 4 hours prior to diagnosis for patients across two Emory University hospitals and the MIMIC-III dataset (68). In contrast to these prior studies, the models we present differ by clinical timeframe (it is intended to be used after a patient is already suspected of having BI and has started EAT) and by the goal of the model (identify patients on EAT who are candidates for EAT discontinuation). Currently, no other prominent EHR-based prediction models exist with the goal of identifying patients on EAT with low risk of having BI who are candidates for EAT discontinuation. Existing methods for forecasting patient-level BI risk have focused on the use of protein and genetic biomarkers (109, 133, 134). The models we present rely on data commonly recorded in the ICU and do not require any specialized laboratory diagnostics or data from current BI risk prediction methods. Our study adds to the body of research surrounding EHR-based prediction models and provides a complementary approach to biomarker-based forecasting of patient-level BI risk. When used in combination with current BI risk metrics and clinical intuition, our model promises to help assist care providers in the de-escalation process of antibiotic stewardship.

For our clinical use case, false negative patients, i.e., those with a serious BI who were predicted as unlikely to have an infection, encompass the largest source of potential patient harm

given the risk of untreated BIs in the ICU and therefore need to be minimized. Similarly, the largest source of potential patient benefit of our model from the current standard of care comes from reducing the number of antibiotic days given to patients who don't have known BI. Our T=24-hour Random Forests model uses a high sensitivity decision threshold to reduce false negative predictions and therefore improve the potential clinical utility in an ICU setting.

We recognize several limitations of this study. First, the retrospective data used was collected for clinical care purposes at a single academic medical center. The retrospective design of our study required us to infer information regarding BI suspicion, consecutive antibiotic days, and culture results based upon sensible criteria that may not completely reflect real world conditions. To address this, chart review and a variety of other quality checks were performed throughout the workflow to ensure appropriate coding of outcomes. Results from our 10-patient chart review of unknown BI status patients found two indeterminate cases and zero misclassifications by our proposed model. Details in the chart notes of one of these indeterminate cases suggested that this patient experienced a prolonged stay in the emergency department prior to transferring to the ICU and that the data from the emergency department was not available in the MIMIC-III dataset. This case suggests that the performance of our phenotype and model can be improved with more complete data on patients prior to ICU transfer. Future work will include retrospective data from additional ICU centers for external model validation and assessment of clinical utility, including data prior to ICU admission. Next, our estimates of antibiotic reduction provide an upper and lower bound on the potential clinical impacts of our model and makes numerous assumptions. To better understand the clinical utility of our model, further study is necessary to test the hypothesis that discontinuing antibiotic therapy on the patients predicted as low risk of BI would clinically benefit them. In future work, we will perform a propensity-

matched analysis to estimate the effects of receiving short vs. prolonged antibiotics on outcome in patients with a predicted low risk of BI. Finally, the longitudinal patient data collected over T=24-,48-, or 72-hours was aggregated prior to modeling using the aggregation function(s) most associated with increased BI risk for each variable. With this design, the time for patients to exhibit symptoms most indicative of BI risk increases as the data collection window increases; however, time-window aggregation methods do not fully capture temporal patterns. To better leverage the longitudinal nature of our data, future work will focus on testing more complex algorithms to explore temporal trends and improve model performances.

## 2.5 Conclusion

The goal of this paper was to detail the design and initial application of a novel collection of algorithms which extract patient features from clinical data and identify patients at low risk of BI who can be safely removed from EAT at 24-hours after initiation. Our models achieved up to 0.8 AUC and demonstrate the feasibility of forecasting BI risk in a critical care setting using patient features found in the EHR. Future work will focus on validating models with external datasets, measuring clinical utility more accurately, and improving model performance by accounting for temporal information in patient data. Overall, these results call for more extensive research in this promising, yet relatively understudied, area.

# 3. DEVELOPMENT AND VALIDATION OF MICROBEX: AN OPEN-SOURCE PACKAGE FOR MICROBIOLOGY CULTURE CONCEPT EXTRACTION

This work was published as:

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).

## 3.1 Introduction

Microbiology culture reports are relied upon for myriad healthcare applications ranging from guiding clinical treatment decisions to global disease surveillance. In a clinical setting, microbiology culture reports are helpful in answering if an infection is present and what organisms are driving that infection (141). Outside of the clinical setting, microbiology data are used to monitor disease outbreaks, improve healthcare operations (e.g. monitor nosocomial infection rates), and are leveraged in a variety of observational studies (142-145). Thus, the data within microbiology reports impacts clinical treatment and public policy decisions, and are therefore critical for secondary use (142, 146).

Unlike many other structured laboratory test results, microbiology culture reports are often complex, semi-structured reports that pose unique challenges for large-scale secondary use applications. Samples sent to a microbiology laboratory routinely undergo numerous tests, such as gram stains and antibiotic susceptibility tests, each of which have different turnaround times, can produce more than a single result, and need to be linked to the original accession number (141, 142). Additionally, results from each test can include both quantitative and qualitative data and need to be reported as they become available to facilitate treatment decisions (141, 142). Unfortunately, although there are efforts to standardize reporting and analysis of clinical microbiology data, the suitability of existing microbiology reports for secondary use are hindered by reporting variability and analysis practices (49, 147, 148). Finally, microbiology reports contain varying amounts of protected health information as defined by the HIPAA, thus limiting the flexibility of this data for data sharing projects.

There is critical need for informatic tools that can navigate microbiology report data challenges and extract information to facilitate their secondary use. The goal of this study was to

develop, validate, and release an open-source microbiology concept extraction (MicrobEx) system to facilitate secondary use of microbiology reports.

## 3.2 Materials & Methods

### 3.2.1 **Datasets**

The two derivation datasets for this study were extracted from two source systems (Epic Systems Corporation and Cerner Corporation) within the Northwestern Medicine (NM) Enterprise Data Warehouse (EDW). Data from source systems one and two were extracted into separate respective derivation sets to reflect different world conditions and preserve their unique microbiology report structures and language characteristics. The regular expressions and logic flow of our extraction system were developed using 216,372 raw free-text microbiology reports extracted from critical care patients treated at one of 10 Northwestern Medicine intensive care units between 1/1/2010-1/1/2020. To define microbiology reports, we queried the NMEDW and manually curated 235 unique procedures associated with microbiology culture orders. The collection of microbiology reports had highly heterogeneous formatting and lacked consistent template features such as concept-value pairs and table structures. Additionally, our corpora contained full microbiology reports, as well as individual microbiology components such as gram stains and antibiotic susceptibility reports. To address these challenges, rules were crafted to separate reports into sections wherever possible. For cultures with multiple report entries tied to the same accession number, only the notes with the latest report update time were selected for downstream processing and analysis. Testing and validation of our extraction system was performed on two external datasets with 65,448 expertly annotated free-text microbiology reports from University of Chicago (validation 1) and Ann & Robert H. Lurie Children's Hospital (validation 2). The validation sets of microbiological culture results were part of prior

study and details have been previously published (146). The reports from both hospitals were annotated by the same senior clinical research coordinator. All four datasets included microbiologic cultures reports from blood, urine, respiratory, and cerebral spine fluid samples.

### 3.2.2 **Algorithm Overview**

A summary of our algorithm workflow is presented in Figure 9. Our concept extraction algorithm uses a comprehensive set of rules, as well as context, keyword, and morphologic features that capture overall bacterial infection status and identify bacterial species present in a microbiology report. Rulesets and regular expressions were developed through an iterative process based on document structural and context features in addition to clinical criteria and domain knowledge. For bacterial species captures, we wrote regular expressions to capture the genus and species for bacteria present in a dictionary of clinically relevant organisms collated from knowledgebases (141, 149). Organisms captured were mapped to Observational Health Data Sciences and Informatics (OHDSI) and Systemized Nomenclature of Medicine (SNOMED) IDs via a dictionary included in the source code. The mapping dictionary for microorganism to OHDSI and SNOMED IDs was constructed by passing the collated microorganism list into Usagi software indexed on SNOMED vocabulary and restricted to class ‘ORGANISM’ and domain ‘OBSERVATION’ (150). During each iteration, concept extraction performance was reviewed manually using a variety of different pattern occurrence-based audits on our training data sets. Customized regular expressions were created to capture remaining complex patterns. Each regular expression was developed with generalizability in mind to maximize dissemination and reusability. For all false positive and negative cases, we reviewed the associated case context, assigned a reason for misclassification. We addressed the cases by either refining existing rules or implementing new ones. This iteration process was repeated until all remaining

uncaptured cases were caused by report noise, uncommon misspellings, or lack of report clarity (92).

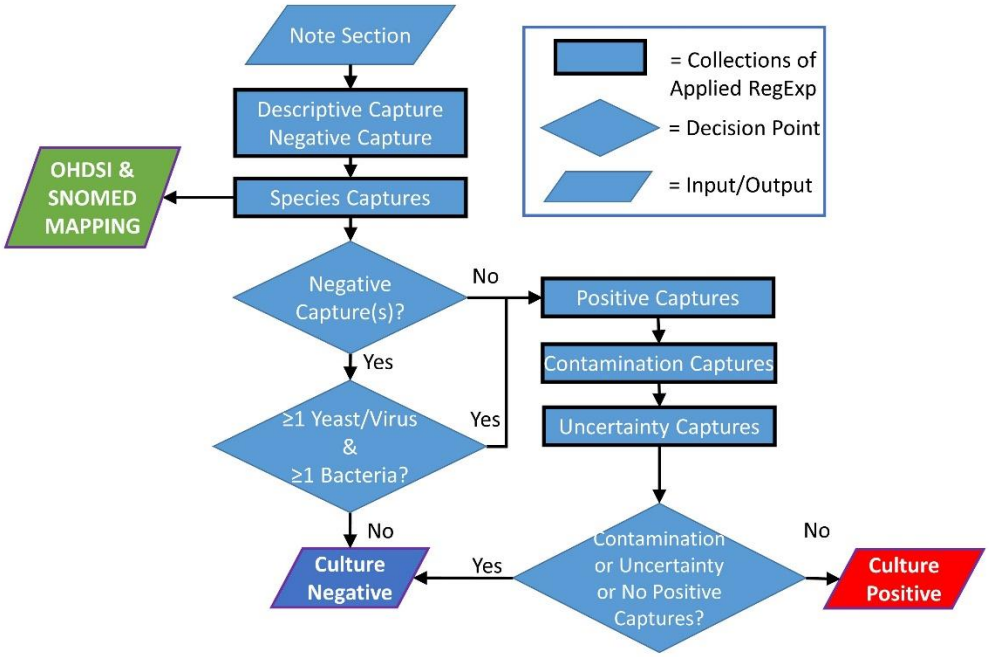


*Figure 9. The MicrobEx algorithm structure. The input of our algorithm is a whole or parsed section of a free-text microbiology culture reports. Within the algorithm, a series of regular expression collections are applied to the text input and the captures are associated with bacterial absence (negative), bacterial presence (positive), microbiological species, potential bacterial contamination, and uncertainty. Bacterial species captured are subsequently mapped to both OHDSI and SNOMED concept IDs. Hierarchical decisions are applied to the regular expression collection captures to categorize the culture as positive or negative for bacteria.*

### 3.2.3 **Validation**

Figure 10 includes example reports annotated with extracted concepts, species, and estimated bacterial culture positive status. Both species extraction and binary bacteria positive culture status (yes/no) were evaluated as outcomes for validation of our algorithm and compared to the manually annotated results in the validation sets. For species extraction, we compared species captured across all report sections by our algorithm and the expert annotation. We

encoded our species extraction binary outcome as positive only if MicrobEx captured every species identified by the expert. Cases where MicrobEx captured bacterial species not identified by the expert were manually reviewed and coded on a case-by-case basis. For positive culture status, MicrobEx assigned a binary classification to all report sections. Report section classifications were aggregated using a maximum function and compared to expert annotation at the report-level.

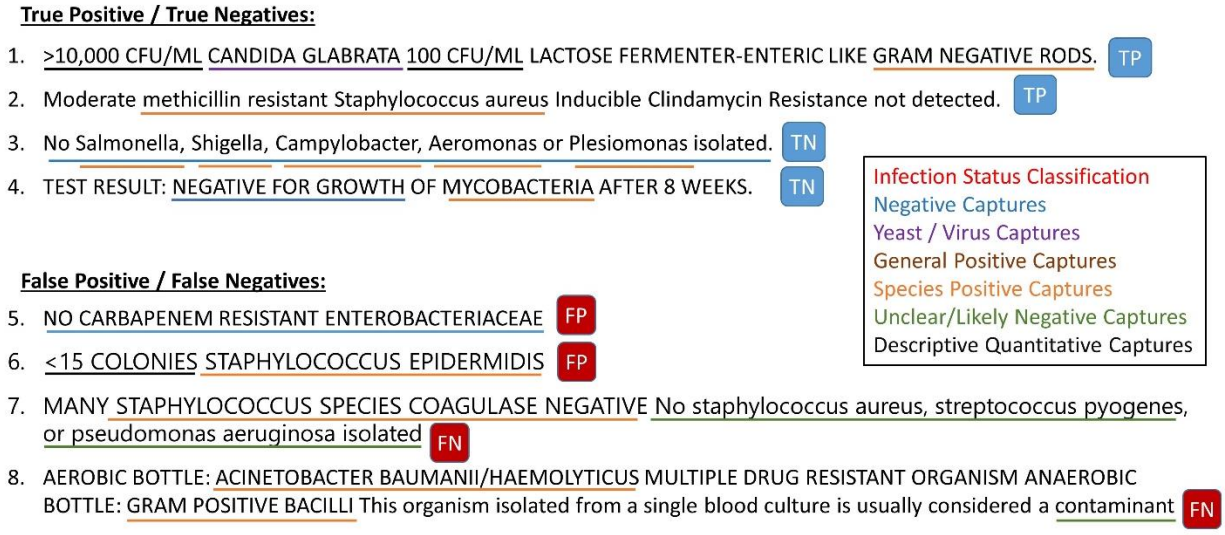


*Figure 10. Examples of annotated validation set reports for error analysis. Colored underlines correspond to parts of the report captured by the associated regular expression collection. For bacterial culture positive status classification, the concepts captured in each block are considered in a hierarchical decision structure according to Figure 9. Examples 1 to 4 demonstrate algorithm annotation on cases found to be correctly classified as positive and negative. Examples 5 to 8 depict four examples representative of common misclassifications. Abbreviations: TP, true positive; TN, true negative; FP, false positive; FN, false negative.*

### 3.2.4 **Performance Benchmark**

In order to benchmark our algorithm's performance against a well-established clinical natural language processing (NLP) tool, we applied MetaMap(90) to both validation sets and

built a rule-based decision workflow to predict positive bacterial culture status and capture bacterial species.

### 3.2.5 **Dataset Customization**

To identify and address dataset-specific patterns capable of causing misclassifications, we audited our workflow as described in the *supplementary use guide* prior to final validation. The generalizable regular expressions we added during the audits were both appended into the codebase prior to our validation studies. The detailed code and Python package installation instructions have been made available at: https://github.com/geickelb/microbex. See the *supplementary use guide* section for Regular expression examples and a description on how to deploy and customize our package to a new dataset.

## 3.3 Results

### 3.3.1 **Validation**

Table 7 summarizes the distribution of positive bacterial culture status in the four datasets. The ratio of positive to negative cases across our training set predictions is consistent with that seen in the two curated validation sets.

*Table 7. Bacterial culture positive status distribution.*

| | **Positive bacterial culture** | **Negative bacterial culture** |
|---|---|---|
| **Derivation set 1** | 14,376 (20.7%) | 55,065 (79.3%) |
| **Derivation set 2** | 23,549 (16%) | 123,382 (84%) |
| **Validation set 1** | 2,184 (14.5%) | 12,916 (85.5%) |
| **Validation set 2** | 7,391 (14.7%) | 42,957 (85.3%) |

Table 8 summarizes the validation results across both species and positive bacterial culture status classification tasks. The algorithm had excellent and consistent performance, with validation sets 1 and 2 having F1-scores of 0.99 and 0.96 for positive culture classification and

species capture, respectively. To estimate the improvements made by introducing customized regular expressions from the data audits, each validation set was reanalyzed using a codebase with the associated regular expressions deactivated. From this, we estimate that culture positivity classification increased from 0.93 to 0.96 and 0.69 to 0.96 for validation sets 1 and 2, respectively. The addition of customized regular expressions was found to cause little-to-no effect on species capturing across both validation sets.

*Table 8. Infection classification and species capture performance across Validation sets.*

| | True Negative | False Positive | False Negative | True Positive | Precision | Recall | NPV | F-1 |
|---|---|---|---|---|---|---|---|---|
| ***Validation set 1.*** | | | | | | | | |
| **Species capture** | 12,463 (82.54%) | 2 (0.01%) | 209 (1.38%) | 2,426 (16.07%) | 0.998 | 0.921 | 0.984 | 0.958 |
| **Positive culture status** | 12,909 (85.48%) | 7 (0.05%) | 22 (0.15%) | 2,162 (14.32%) | 0.995 | 0.990 | 0.998 | 0.992 |
| ***Validation set 2.*** | | | | | | | | |
| **Species capture** | 42,391 (84.20%) | 4 (0.01%) | 68 (0.14%) | 7,885 (15.66%) | 0.999 | 0.991 | 0.999 | 0.995 |
| **Positive culture status** | 42,950 (85.31%) | 7 (0.01%) | 606 (1.20%) | 6,785 (13.48%) | 0.998 | 0.918 | 0.986 | 0.956 |

Table 9 presents the results from our customized MetaMap based benchmarking algorithm against both validation sets. Across both positive culture classification and species capture, MicrobEx matched or surpassed the benchmark algorithm performance. These results suggest that our task-specific classifier can outperform more general-use clinical NLP tools like MetaMap. Supplemental Run Report 1. presents our MicrobEx “Run Report” for validation sets 1&2, detailing report- and report section-level data regarding regular expression captures, binary classification decision data, and descriptive statistics.

*Table 9. Baseline MetaMap classifier infection classification and species capture performance across Validation sets.*

| | True Negative | False Positive | False Negative | True Positive | Precision | Recall | NPV | F-1 |
|---|---|---|---|---|---|---|---|---|
| ***Validation set 1.*** | | | | | | | | |
| **Species capture** | 12,418 (82.24%) | 47 (0.31%) | 287 (1.90%) | 2,348 (15.55%) | 0.980 | 0.891 | 0.977 | 0.934 |
| **Positive culture status** | 12,558 (85.49%) | 358 (2.37%) | 299 (1.98%) | 1,885 (12.48%) | 0.840 | 0.863 | 0.977 | 0.852 |
| ***Validation set 2.*** | | | | | | | | |
| **Species capture** | 42,395 (84.20%) | 0 (0.00%) | 1578 (1.35%) | 6,375 (14.44%) | 1.00 | 0.914 | 0.984 | 0.955 |
| **Positive culture status** | 36,612 (72.72%) | 6345 (12.60%) | 1130 (2.24%) | 6,261 (12.43%) | 0.497 | 0.847 | 0.97 | 0.626 |

### 3.3.2 **Error Analysis**

In the error analysis we identified a collection of five patterns that account for most of the errors observed in our concept extraction workflow. Figure 10 presents annotated visual examples of the classification hierarchical logic for the different patterns observed, with examples for both correct classifications as well as misclassifications. Examples 5 and 6 depict the two most common types of false positive patterns and examples 7 and 8 present the most common patterns found in false negatives in the validation sets (Figure 10). We can summarize these patterns as a combination of multiple positive and negative organisms where the negative regex capture supersedes the positive captures, and the use of the term “contaminant” leading to a false negative classification.

## 3.4 Discussion

In this study, we developed and validated an open-source, rule-based framework to extract and map clinical concepts from microbiology reports to standardized terminologies to facilitate secondary use of microbiology reports. Our main finding is that our algorithm can reliably estimate binary bacterial culture status, extract bacterial species, and map these to SNOMED organism observations when applied to semi-structured, free-text microbiology reports from different institutions with relatively low customization.

Top performing rule-based concept extraction applications commonly employ a well-established clinical NLP tool that can map mentions to a corresponding medical concept(s) for broad medical corpora, such as cTAKES(91) and MetaMap(90). Like the well-established tools, MicrobEx performs concept matching by leveraging existing microbiology knowledgebases as described in Materials & Methods. In contrast to these tools however, MicrobEx uses custom rules and regular expressions tailored to microbiology reports for negation detection. MicrobEx’s

higher performance on bacterial positive culture status prediction suggests that for this classification task, MicrobEx's more tailored approach provides advantages over an out-of-the-box approach using a well-established NLP tool. To illustrate, validation set 1 had a language pattern (n=88 report-level occurrences) where the results from the antibiotic susceptibility report were mentioned alongside the microorganism summary (E.G "'Many methicillin resistant Staphylococcus aureus Inducible Clindamycin Resistance not detected"). Our MetaMap benchmarking algorithm, which used a Negex negation detection engine, classified culture status negative while MicrobEx correctly classified culture status positive for such cases. By including specific regular expressions to distinguish between susceptibility or resistance detection from microorganism detection (E.G "*(?<!resistance)(?<!susceptibility)\s+not\sdetected|indicated*"), MicrobEx was able to correctly classify binary bacteria culture status. To further improve MicrobEx's prediction performance, additional institution-specific customized rules could be added. Figure 10 depicts four representative examples of cases misclassified for positive culture status that could be addressed with institution-specific custom rules.

To our best knowledge, three previously published studies have applied clinical concept extraction methods to microbiology notes (151-153). Jones et al. (151) applied a set of crafted rules to blood culture reports from the Salt Lake City Healthcare system to extract organism information, antibiotic susceptibilities, and infer if methicillin-resistant staphylococcus aureus (MRSA) was present. An evaluation was performed against approximately 10,000 expertly annotated reports to measure successful identification of MRSA. Matheny et al. and Yim et al. (152, 153) used hybrid and rule-based systems to capture combinations of microorganism species and antibiotic susceptibilities from blood and multiple sample types, respectively. Our algorithm is notably different from the previously published systems in the following ways: 1)

we estimate positive bacterial culture status, 2) our algorithm was designed to work with a variety of disparate microbiology report formats from different institutions, 3) we performed external validation on two expertly annotated microbiology datasets, and 4) our software is entirely open-source and available as a python package that can be further adapted to the reports of other institutions as described in the *supplementary use guide* and supported by our results.

We recognize several limitations of our study. First, for users of this software, classifying positive culture status is the prediction task with the largest potential error. Compared to species extraction, which is largely string matching, estimating infection status requires significantly more complex logic. The hierarchical logic involved with positive bacterial culture status estimation is potentially susceptible to syntactic heterogeneity and report complexity, as depicted in Figure 10. Additionally, we focused on bacterial cultures for the development and validation of the algorithm given the importance of antibiotic stewardship, antibiotic resistance, and bacterial sepsis in hospitalized patients. While our algorithm captures other microorganism species (including fungal and viral species), we did not validate the performance on those. Finally, we included logic to extract relevant quantitative and semi-quantitative concepts, however the performance of this was variable due to syntactic heterogeneity. As a result, we continue to provide quantitative captures as a feature of the MicrobEx algorithm, however these were not included in our validation.

## 3.5 Conclusion

In this article we detail the development, validation, and use of our open-source microbiology concept extractor (MicrobEx) algorithm and package. Our workflow achieved excellent performance in two independent validation sets with minimal customization, improved performance versus a well-established alternative, and comparable performance to manual chart

review by an expert. Our concept extraction Python package is designed to be reused and adapted to individual institutions as an upstream process for other clinical applications such as machine learning, clinical decision support, and disease surveillance systems.

# 4. TRANSPORTABILITY OF BACTERIAL INFECTION PREDICTION MODELS FOR CRITICALLY ILL PATIENTS

## 4.1 Introduction

For patients in the intensive care unit (ICU), bacterial infections (BI) 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 25-30%, more than twice that of patients without an infection (1, 5). 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 (117, 154). 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 (19, 38, 42, 131). 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 (25-28). Developing data-driven strategies to help providers stratify patient-level BI risk shortly after an ICU admission offers a promising avenue in antibiotic stewardship (38, 121).

We previously proposed a model to predict BI risk in patients at 24-hours 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 (145) (**Chapter 2**). It's widely acknowledged that the performance of any clinical prediction model should be evaluated in different populations using equivalent

information prior to clinical implementation (103, 146, 155-158). For a classification model trained and tested in two independent populations, differences between the populations in terms of predictor and/or outcome distributions can lead to variation in model class discrimination and calibration performance (96, 159-162). 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. (76) 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 (145) (**Chapter 2**) on two 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 two tertiary ICUs would improve the performance in the community ICU setting.

## 4.2 **Methods**

### 4.2.1 **Datasets**

Data were obtained from two 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 de-identified dataset collected from

over 40,000 patients who received care at Beth Israel Deaconess Medical Center ICU between 2001 and 2012 (122, 123). 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 Northwestern Medicine health system. NMEDW intensive care unit (ICU) encounter information was sourced from a manually curated subset of 55,989 ICU encounters across 6 different NM affiliated hospitals in Northeastern Illinois (including several community hospitals) for patients admitted between 2011-10-01 and 2020-01-01.

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 is predominantly comprised of 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 two 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 $Ensemble_M$ (model details described below).

### 4.2.2 **Cohort**

Cohort selection and computational phenotype labeling were performed on all patients as detailed in our prior work (145) (**Chapter 2**). Briefly, patients 16 years or older suspected of having a bacterial infection (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 hours of ICU admission and a microbiology culture sampled from a sterile site within 24 hours of the first antibiotic dose. Patients who matched these cohort criteria were allocated to one of three groups based upon their BI status: serious BI, non-serious BI/no BI, and unknown BI status (Figure 11), detailed below. Due to the common occurrence of occult bacterial infections, 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.

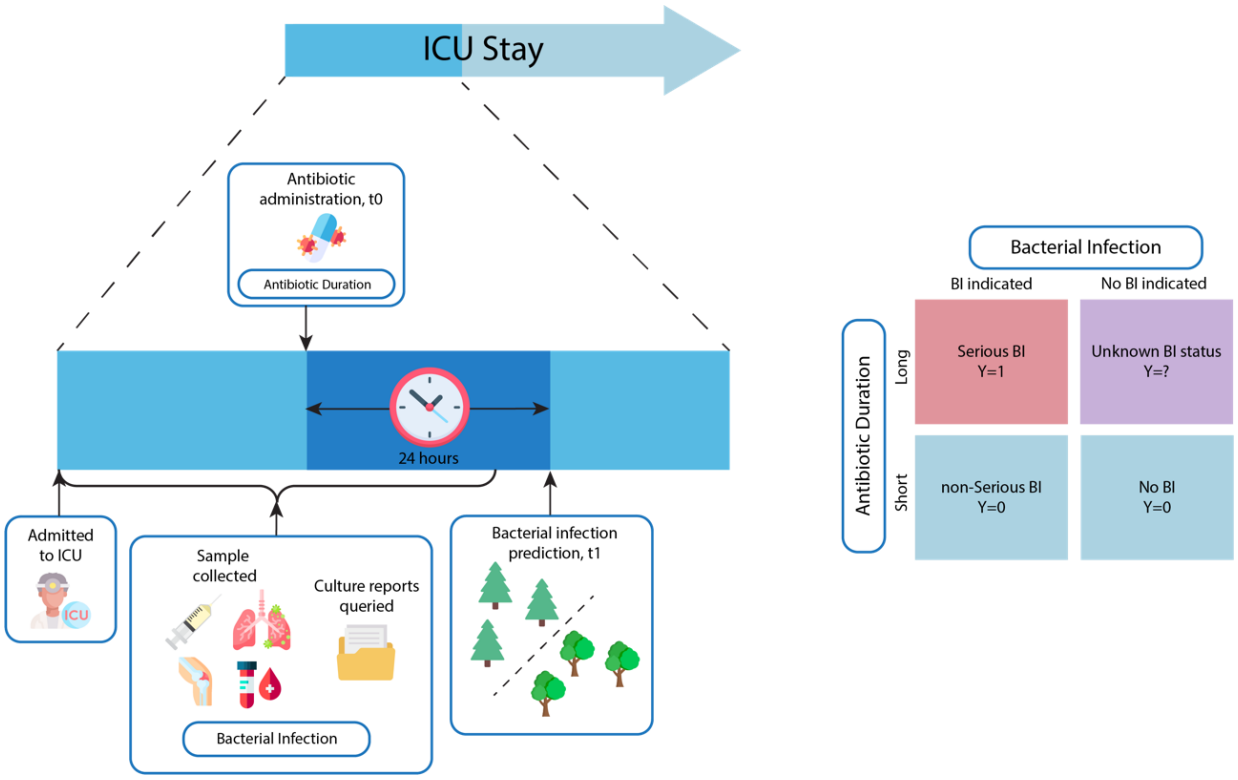


*Figure 11. BI status labeling and classification. Our phenotype for BI suspicion upon ICU admission requires that (1) an antibiotic be administered within 96h following ICU admission and (2) a microbiology culture be drawn within 24 hours of (1). Clinical data were collected for 24-hours after the first 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 ("non-serious BI") or negative culture ("no BI") were labeled prediction non-events.*

### 4.2.3 **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-hours 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* (163). 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 (163). Two consecutive positive cultures were required for positive BI for coagulase-negative *Staphylococcus* and other common contaminate species to warrant inclusion.

### 4.2.4 **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 (164) were selected. In NMEDW, the same was identified using regular expressions, manual curation, and medical expert review (145) (**Chapter 2**). 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 hours, 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 hours prior to ICU admission and continue to accumulate up until hospital discharge if the medication was also administered during the patient's ICU stay. As performed in **Chapter 2**, patients who died within 24 hours of their final antibiotic dose were coded as having received prolonged antibiotic therapy (n= $MIMIC_D$:1266, NM-$T_D$: 1238, NM-$C_{val}$: n=484).

### 4.2.5 **Outcome**

Patients with both a positive bacterial culture and prolonged antibiotic therapy were classified as serious BI status (prediction event) (Figure 11). Patients with negative bacterial culture and a short antibiotic timeline were considered to have no BI (prediction non-event). Patients with a positive BI culture but short antibiotic treatment duration were considered to have non-serious BIs (prediction non-event). 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 (145) (**Chapter 2**) to categorize these patients as unknown BI status and exclude them from modeling for this study.

### 4.2.6 **Data Extraction, Cleaning, & Pre-processing**

We follow our previous studies' in-depth descriptions and open-source code for the extraction, cleaning, and pre-processing of static and longitudinal data from the MIMIC-III database (145) (**Chapter 2**). Data extraction, cleaning, and pre-processing for both NM-$T_D$ and NM-$C_{val}$ followed the same framework and is 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-hour window after the first antibiotic dose ($t_0$: $t_{0+24}$) (Table 10). 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 three datasets and manually reviewed (Table 10, Appendix Figure S 1-Figure S 10). 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 non-events (Equation 1). Finally, all continuous values within the 24-hour collection window were aggregated using functions (minimum, maximum, or both), based on our previous model (145) (**Chapter 2**). After one-hot encoding categorical variables, our final feature list included 55 variables.

$$Z = \frac{X - \tilde{X}_{(neg\ /\ short)}}{IQR_{(neg\ /\ short)}} \quad (1)$$

Table 10. Extracted predictor data and percent missing across datasets from NM tertiary referral hospitals (NM-T) NM-affiliated community hospitals (NM-C) and MIMIC-III.

| DATA COLLECTED | UNIT | % MISSING MIMIC | % MISSING NM-T | % MISSING NM-C |
|---|---|---|---|---|
| **ELIXHAUSER COMORBIDITY INDEX** | categorical | 0 (0, 0) | 0 (0, 0) | 0 (0, 0) |
| **AGE** | years | 0 (0, 0) | 0 (0, 0) | 0 (0, 0) |
| **ETHNICITY** | categorical | 0 (0, 0) | 0 (0, 0) | 0 (0, 0) |
| **GENDER** | categorical | 0 (0, 0) | 0 (0, 0) | 0 (0, 0) |
| **SYSTOLIC, DIASTOLIC BLOOD PRESSURE (SBP, DBP)** | mmHg | 0.2 (0.2, 0.4) | 0 (0, 0) | 0 (0, 0) |
| **GLASGOW COMA SCALE (GCS)** | GCS score | 43.6 (41.6, 49.6) | 11.4 (9.2, 13.5) | 8 (7.6, 8.2) |
| **GLUCOSE** | mg/dL | 0.4 (0.2, 0.8) | 1.1 (0.5, 1.7) | 1.1 (0.5, 1.5) |
| **HEART RATE** | beat/minute | 0.1 (0.1, 0.1) | 0 (0, 0) | 0 (0, 0) |
| **RESPIRATION RATE** | breath/minute | 0.1 (0.1, 0.1) | 0 (0, 0) | 0 (0, 0) |
| **PERIPHERAL OXYGEN SATURATION (SPO2)** | % | 0.1 (0.1, 0.1) | 0 (0, 0) | 0 (0, 0) |
| **TEMPERATURE** | deg. C | 1 (0.6, 2) | 0 (0, 0) | 0 (0, 0) |
| **VENTILATION STATUS** | categorical | 0 (0, 0) | 0 (0, 0) | 0 (0, 0) |
| **WEIGHT** | kg | 8.4 (8.2, 8.8) | 18.5 (17, 19.9) | 7.2 (6.9, 7.7) |
| **DOBUTAMINE** | mcg/kg/min | 0 (0, 0) | 0 (0, 0) | 0 (0, 0) |
| **DOPAMINE** | mcg/kg/min | 0 (0, 0) | 0 (0, 0) | 0 (0, 0) |
| **EPINEPHRINE** | mcg/kg/min | 0 (0, 0) | 0 (0, 0) | 0 (0, 0) |
| **NOREPINEPHRINE** | mcg/kg/min | 0 (0, 0) | 0 (0, 0) | 0 (0, 0) |
| **PHENYLEPHRINE** | mcg/kg/min | 0 (0, 0) | 0 (0, 0) | 0 (0, 0) |
| **VASOPRESSIN** | mcg/kg/min | 0 (0, 0) | 0 (0, 0) | 0 (0, 0) |
| **RENAL REPLACEMENT THERAPY** | pos/neg | 0 (0, 0) | 0 (0, 0) | 0 (0, 0) |
| **BANDS** | % | 87.3 (80, 92) | 92 (88.3, 95.2) | 59 (48.6, 68.9) |
| **SERUM BICARBONATE** | mEq/L | 1.4 (1.2, 2.6) | 1.4 (0.5, 2.2) | 1.2 (0.5, 1.6) |
| **BILIRUBIN** | mg/dL | 37.6 (26.1, 41.4) | 24.7 (18.3, 30.3) | 18.8 (13.4, 23.5) |
| **BLOOD UREA NITROGEN (BUN)** | mg/dL | 1.4 (1.4, 2.4) | 1.8 (1, 2.6) | 1.9 (1.3, 2.4) |
| **SERUM CHLORIDE** | mEq/L | 1.3 (1, 2.4) | 1.8 (1, 2.6) | 1.9 (1.3, 2.3) |
| **SERUM CREATININE** | mg/dL | 1.4 (1.4, 2.4) | 1.6 (1, 2.4) | 1.7 (1.2, 2) |
| **HEMOGLOBIN** | g/dL | 1.8 (2.3, 3.1) | 2 (1, 2.9) | 2.3 (0.9, 3.2) |
| **INTERNATIONAL NORMALIZED RATIO (INR)** | ratio | 23.5 (20.8, 27) | 37.9 (33.5, 40.2) | 39.3 (32.1, 45.8) |
| **SERUM LACTATE** | mmol/L | 30.7 (35.2, 54.1) | 26.7 (18.4, 33.6) | 21.8 (13, 28.4) |
| **URINE LEUKOCYTE** | pos/neg | 40.1 (20, 48.1) | 39.4 (30, 46.9) | 32.1 (22.7, 39.6) |
| **URINE NITRITE** | pos/neg | 40.1 (20, 48.1) | 39.5 (30.1, 47) | 32.1 (22.7, 39.6) |
| **PARTIAL PRESSURE OF ARTERIAL OXYGEN: FRACTION OF INSPIRED OXYGEN (PAO2:FIO2)** | ratio | 67.9 (52.1, 69.1) | 44 (35, 51.2) | 51.5 (41.7, 59.4) |
| **PARTIAL THROMBOPLASTIN TIME (PTT)** | sec | 26 (21, 27.6) | 55.7 (55, 57) | 54.9 (48.2, 61) |
| **PARTIAL PRESSURE OF ARTERIAL CARBON DIOXIDE (PCO2)** | mmHg | 39.9 (33, 43.8) | 41.6 (33.7, 48.6) | 41.5 (31.2, 49.6) |
| **SERUM PH** | n/a | 44.8 (29.9, 50) | 33.6 (25.9, 40.7) | 42.9 (33, 50.9) |
| **PLATELET COUNT** | K/uL | 2.9 (2.2, 3.1) | 2.6 (1.4, 3.5) | 3 (1.4, 4.1) |
| **SERUM POTASSIUM** | mEq/L | 1.7 (1, 1.9) | 1.4 (0.7, 2.1) | 1.4 (1, 1.7) |
| **WHITE BLOOD CELL COUNT (WBC)** | K/uL | 3.2 (2.3, 3.4) | 2.5 (1.4, 3.5) | 2.9 (1.2, 4.1) |
| **SERUM CALCIUM** | mmol/L | 62.1 (56, 68.6) | 47.3 (43.1, 51.5) | 63.8 (58.2, 68.8) |

a. % missingness is presented for the entire dataset (first) and for the y=0 (second) and y=1 (third) subsets.

### 4.2.7 **Modeling and Statistical Analyses**

ICU encounters from $MIMIC_D$ and $NM\text{-}T_D$ were split 70/30 into independent train ($_{train}$) and test ($_{test}$) sets, while encounters in NM-C were set aside for model validation ($NM\text{-}C_{val}$). Individual patients with more than one 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=4,637) and $Pooled_{test}$ (n=1,989) 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 (127, 165). 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 (165, 166). Fit models applied to test sets from differing institutions (e.g., $MIMIC_M$ on $NM\text{-}T_{test}$) were first recalibrated on the associated training set (e.g., $NM\text{-}T_{train}$). High sensitivity ($\geq 0.9$) 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 (167), the high-sensitivity threshold was

determined using a fit 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 (168). 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 (169, 170). 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 (161, 171, 172). 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 (173). We set $\alpha = 0.005$ by default, as previously recommended for large datasets (174). In cases with more than 10 comparisons, a Bonferroni correction was applied.

### 4.2.8 **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.

### 4.2.9 **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).

## 4.3 **Results**

### 4.3.1 **Cohort Characteristics**

We identified ICU encounters in $MIMIC_D$ (n=19,633; 37.7% of all ICU encounters), NM-$T_D$ (n=11,076; 40.2% of all ICU encounters), NM-$C_{val}$ (n=4,059; 38.8% of all ICU encounters) that met our study inclusion criteria. The demographics of patients in the three datasets used are presented in Table 11. Table 12 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-$T_D$ and NM-$C_{val}$ had BI prevalence of 44.2% and 44.6%, respectively.

*Table 11. Demographics of BI positive and negative labeled patients across hospital datasets.*

| **Variable** | **$MIMIC_D$** | **NM-$T_D$** | **NM-$C_{Val}$** |
|---|---|---|---|
| *Gender- N, %* | | | |
| Female | 5340 (47%) | 3112 (47%) | 1241 (50%) |
| Male | 6013 (53%) | 3514 (53%) | 1244 (50%) |
| Age in years (stdev.) | 65.3 +\− 17.0 | 64.1+\- 17.1 | 66.7 +\- 17.8 |
| *Ethnicity- N, %* | | | |
| Black/non-Hispanic | 1294 (11%) | 782 (12%) | 151 (6%) |
| White/non-Hispanic | 8218 (72%) | 4586 (69%) | 2040 (82%) |
| Hispanic | 468 (4%) | 606 (9%) | 166 (7%) |
| Other | 1373 (12%) | 652 (10%) | 128 (5%) |

a. Abbreviations: NM tertiary referral hospitals (NM-T); NM community hospitals (NM-C); MIMIC-III (MIMIC)

Table 12. Cohort stratified by BI status and hospital datasets.

| Microbiology Culture | Antibiotic Duration | BI Status Classification | MIMIC N (% cohort; % ICU) | NM-T N (% cohort; % ICU) | NM-C N (% cohort; % ICU) |
|---|---|---|---|---|---|
| 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%; 5.9%) | 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.
b. Abbreviations: NM tertiary referral hospitals (NM-T); NM community hospitals (NM-C); MIMIC-III (MIMIC)

### 4.3.2 **Model Evaluation**

In Table 13 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 (e.g., $MIMIC_M$ on $MIMIC_{test}$) had significantly ($P<0.002$) higher AUROC than models trained on external development cohorts (e.g., $MIMIC_M$ 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 13. $MIMIC_M$ and $NM\text{-}T_M$ classification discrimination & performance.*

| Model | 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 |
| $NM\text{-}T_M$ | $NM\text{-}T_{test}$ | 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 |

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

Table 14 summarizes the results of two 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 significantly different on $NM\text{-}T_{test}$ ($P$=2x10$^{-4}$) but not significantly different on $MIMIC_{test}$ ($P$=0.037) with a Bonferroni-adjusted $P$<0.002. 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 14. $Ensemble_M$ vs. $Pooled_M$ classification discrimination & performance.*

| Model | Evaluation Set | BI (%) | AUROC | F1 | NPV | Precision | Recall | High Sensitivity Threshold | DeLong *P*-Value |
|---|---|---|---|---|---|---|---|---|---|
| $Pooled_M$ | $MIMIC_{test}$ | 24.8 | 0.774 | 0.538 | 0.878 | 0.436 | 0.703 | 0.131 | 0.037 |
| $Ensemble_M$ | $MIMIC_{test}$ | 24.8 | 0.767 | 0.458 | 0.937 | 0.303 | 0.942 | 0.131 | 0.037 |
| $Pooled_M$ | $NM\text{-}T_{test}$ | 44.3 | 0.788 | 0.696 | 0.773 | 0.662 | 0.734 | 0.267 | 2x10$^{-4}$ |
| $Ensemble_M$ | $NM\text{-}T_{test}$ | 44.3 | 0.798 | 0.695 | 0.879 | 0.556 | 0.926 | 0.267 | 2x10$^{-4}$ |

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

Table 15 and Figure 12 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 ($P$<0.002) 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 appear to be highly distinct based on the C-statistics presented in Table 15 and cohort BI status distributions presented in Table 12. Higher C-statistic values also appear to correspond with

lower AUROC values in NM-$C_{val}$, 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 13. When comparing each model's performance across all evaluation sets, the range of AUC values for multisite learning models was lower (more stable) compared to single institution models.

*Table 15. Modeling classification discrimination & performance on NM-$C_{val}$.*

| 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-$T_M$ | 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 |

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

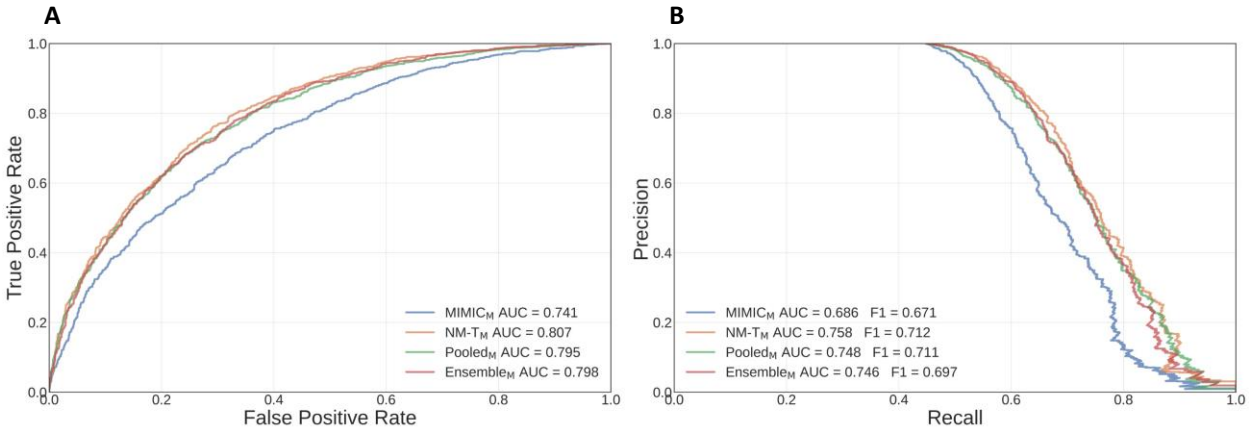


*Figure 12. NM-$C_{val}$ model evaluation. A. Receiver operating characteristic curves (ROC). B Precision recall curves (PRC). The ROC generated by NM-$T_M$ was the highest of any model on NM-$C_{val}$ and was significantly different than $Pooled_M$ and $MIMIC_M$ but not significantly different than $Ensemble_M$ at adjusted P≤ 0.002 via DeLong's test. However, there was no difference observed between the ROC of $Ensemble_M$ and $Pooled_M$. Finally, $MIMIC_M$'s ROC was significantly different from all other models at P≤ 0.002. All significant differences observed in ROC additionally observed in PRC.*

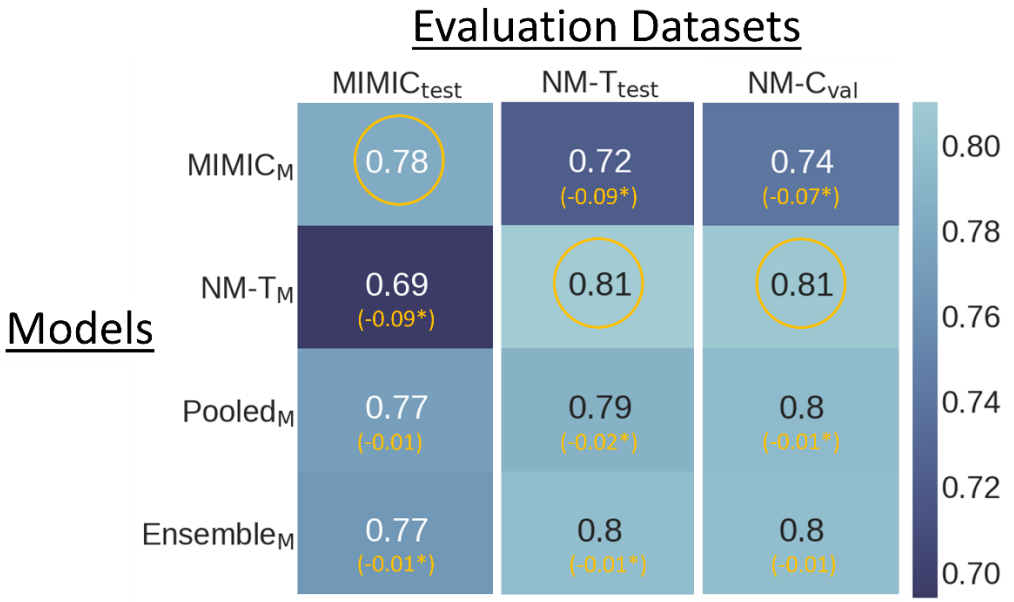


*Figure 13. AUROC heatmap between models and evaluation sites. Gold rings indicate the best performing model on each evaluation cohort while gold numbers present the AUROC delta relative to the gold ring model. * Denotes significant difference from gold ring model using DeLong test at P<0.002. Although NM-T$_M$ and MIMIC$_M$ 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).*

### 4.3.3 **Predictor Effects**

Figure 14 displays the relative variable importance for each model in NM-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 (Appendix Figure S 1-Figure S 10). 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 (Appendix Figure S 1-Figure S 10).

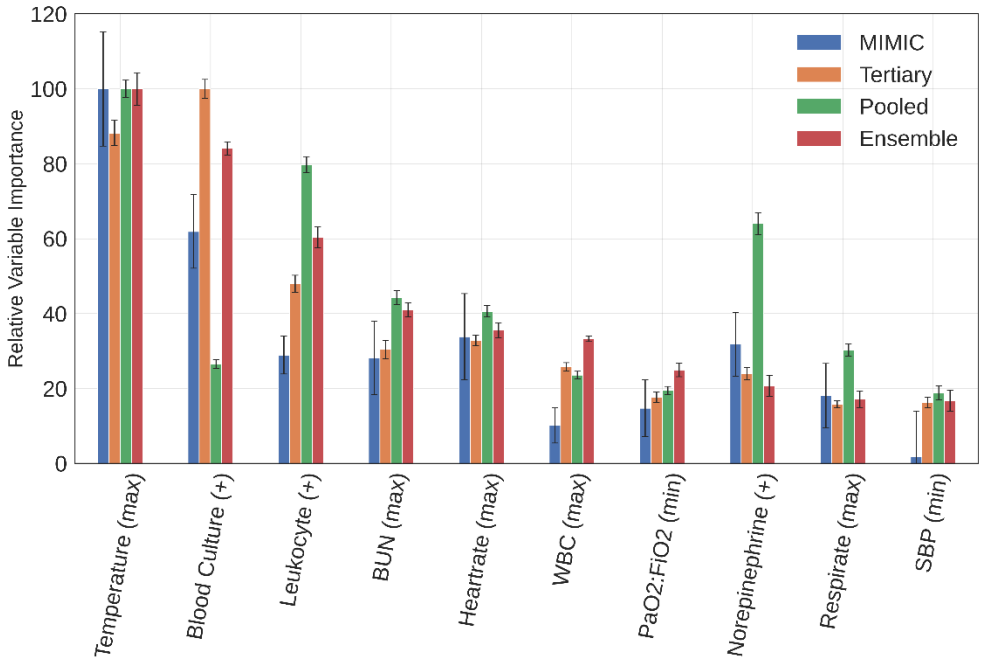


*Figure 14. Relative variable importance across models for top 10 important variables. All models are calibrated and permuted on NM-$T_{train}$. Variable importance values for each model were scaled relative to each model's most important variable. pCO2: carbon dioxide partial pressure; Blood Culture: indication for microbiology culture performed on blood sample; Leukocyte: indication for leukocytes in urine; BUN: Blood urea nitrogen; WBC: white blood cell count; PaO2:FiO2: ratio of arterial oxygen partial pressure to fractional inspired oxygen; Norepinephrine: indication for having received norepinephrine; Respirate: max respiration rate; SBP: Systolic blood pressure.*

### 4.3.4 **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 16). The reliability diagram for patients in NM-$C_{val}$ (Figure 15) and calibration statistics presented in Table 17 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 17) and all BI

prevalence patient bins (Figure 15), likely contributing to the mismatch between observed recall and desired recall in Table 15.

*Table 16. Model predicted BI probability average vs BI prevalence across evaluation datasets (Mean calibration).*

| **Model** | **BI Prevalence Decimal** | **$MIMIC_M$** | **NM-$T_M$** | **$Pooled_M$** | **$Ensemble_M$** |
|---|---|---|---|---|---|
| $MIMIC_{test}$ | 0.25 | 0.25 ± 0.20 | 0.25 ± 0.15 | 0.23 ± 0.28 | 0.25 ± 0.16 |
| NM-$T_{test}$ | 0.44 | 0.43 ± 0.18 | 0.44 ± 0.25 | 0.43 ± 0.43 | 0.43 ± 0.20 |
| NM-$C_{val}$ | 0.45 | 0.45 ± 0.19 | 0.47 ± 0.25 | 0.49 ± 0.44 | 0.46 ± 0.21 |

a. Mean calibration is assessed based on the agreement between BI prevalence decimals and the average predicted probability of a model.
b. Abbreviations: NM tertiary referral hospitals (NM-T); NM community hospitals (NM-C); MIMIC-III (MIMIC)

*Table 17. Calibration evaluation statistics for all models on NM-$C_{val}$.*

| **Model** | **Brier score** | **Spiegelalter-Z** | **Spiegelalter-P** | **Cox slope** | **Cox intercept** |
|---|---|---|---|---|---|
| $MIMIC_M$ | 0.206 | -1.61 | 0.054 | 1.069 | -0.036 |
| NM-$T_M$ | 0.179 | -1.55 | 0.06 | 1.036 | -0.037 |
| $Pooled_M$ | 0.232 | 58.88 | 0 | 0.545 | 0.180 |
| $Ensemble_M$ | 0.185 | -5.99 | $1x10^{-9}$ | 1.23 | -0.120 |
| $MIMIC_M$ | 0.206 | -1.61 | 0.054 | 1.069 | -0.036 |

a. Abbreviations: NM tertiary referral hospitals (NM-T); NM community hospitals (NM-C); MIMIC-III (MIMIC); equal sized samples from MIMIC and NM-T concatenated together ($Pooled_M$); soft-voting ensemble of NM-$T_M$ and $MIMIC_M$ ($Ensemble_M$)

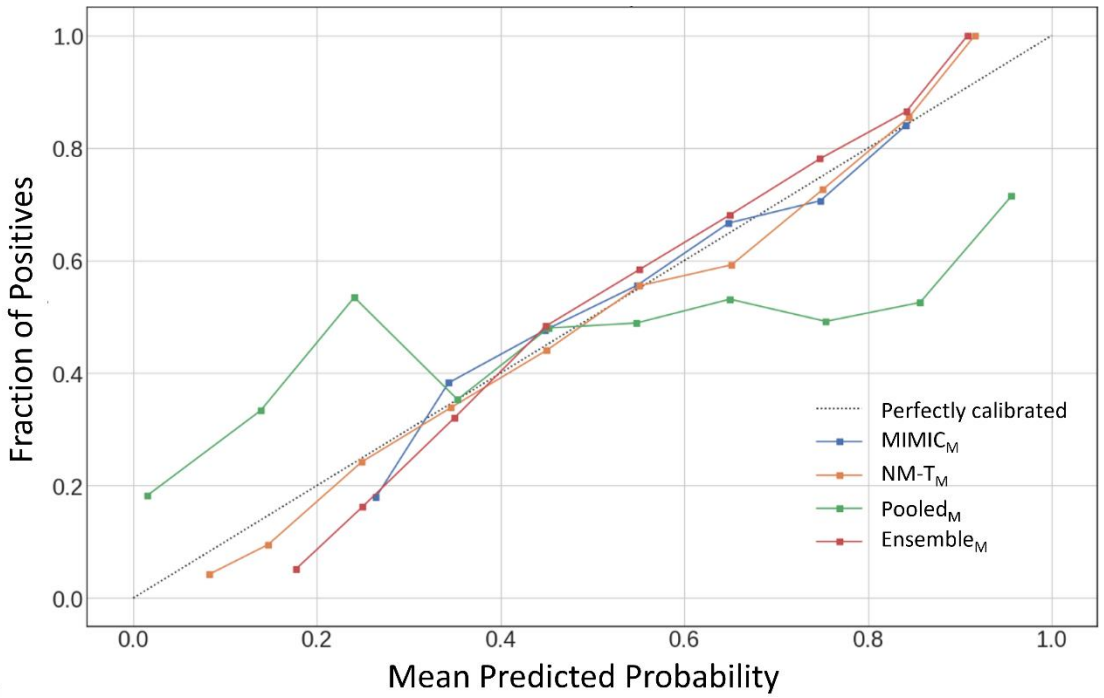


*Figure 15. Model calibration plot for NM-$C_{val}$. All models were calibrated or recalibrated to NM-$T_{D\text{-}Train}$ using Platt scaling. MIMIC$_M$ (Platt scaled), NM-$T_M$ and Ensemble$_M$ all achieved adequate calibration despite some deviations from perfect calibration on the lowest and highest BI fractions. Our Pooled RandomForests model demonstrated poor model calibration across a range of BI fractions and was resistant to Platt scaling.*

## 4.4 **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 hours of initiation (145, 163) (**Chapters 2,3**). In the current study, we carried out an in-depth transportability assessment of the previously developed BI prediction model architecture on two 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 (103, 157, 158, 173). 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 (i.e., electronic phenotype, predictor input, etc.) may be transportable between different sites, the models themselves may not translate with the same level of 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 (e.g., distribution of predictor variables and setting characteristics) (76, 157-159, 162, 175). 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 (26, 42, 119, 176, 177). 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. 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 (84, 178).

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 hours prior to diagnosis (68, 70) and AUROCs in the 0.65 to 0.80 range for predicting mortality in septic patients (71-73). The most important variables in these studies (systolic blood pressure, blood urea nitrogen, respiratory rate, and temperature) were also among the top 10 most important variables in all four of our models. Next, several related publications have reported improvements in model performance after employing multisite learning techniques

such as federated and transfer learning (70, 168, 179). We found that models trained using two 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., who found that across five datasets and 21 outcomes, weighted fusion ensembles produced more stable class discrimination when transported to new databases compared to single database models (168). 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 (163). 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-$T_M$ (Appendix Annotation S 1-Annotation S 10). In this chart review, we identified that in seven out of the ten 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 (49, 180). To better understand the clinical utility of our model, further study is necessary to test the hypothesis that discontinuing antibiotic therapy on the patients predicted as low risk of BI would clinically benefit them.

## 4.5 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 two 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.

# 5. EMPIRIC ANTIMICROBIAL TREATMENT DURATION VS OUTCOMES IN CRITICALLY ILL PATIENTS WITH LOW PREDICTED RISK OF BACTERIAL INFECTION

## 5.1 Introduction

Critical care providers must navigate the challenging decision space between failing to treat a patient with a potentially serious bacterial infection (BI) and treating a low BI risk patient with an overzealous or inappropriately protracted antibiotic regimen. Patients with a BI in the intensive care unit (ICU) have more than two-fold higher risk of ICU mortality than those without an infection, and that risk has been shown to decrease following prompt and broad empirical antibiotic therapy (EAT) (1, 5, 154, 181, 182). Upon characterizing the pathogen(s) and scrutinizing information collected during follow-up, antibiotic regimens are typically de-escalated to provide more targeted coverage and mitigate undesired antibiotic effects as more information becomes available (117, 154). When done appropriately, antibiotic de-escalation has been associated with lower morbidity and mortality rates in numerous prospective and observational studies (183-187). However, within this paradigm, there are areas of uncertainty and variability in patients whose infection status is unclear, this issue is further compromised by the presence of relatively non-uniform guidelines for the duration of EAT. Recent efforts have aimed to address this issue through enhancing diagnostic methods for BI and antibiotic stewardship efforts (19, 38, 42, 131). Despite this, current practices often result in patients with low risk of BI being exposed to prolonged and unnecessary antibiotic regimens.

Prolonged antibiotic exposure is not benign and has been associated with a spectrum of adverse drug effects such as gut microbiome dysbiosis, hematologic abnormalities, hepatic injury, and increased rates of antimicrobial resistance, all of which are drug dependent (25-30). The impact of inappropriate empiric antibiotic regimen and antibiotic de-escalation on in-hospital mortality and other common outcomes has been studied across numerous independent variables and cohorts such as effect of time delay (188), infection sites (189, 190) specific

organisms (191, 192), or conditions (e.g. sepsis, pneumonia) (193, 194). Indeed, the impact of antibiotic regimen have been well characterized in cohorts with confirmed BI or high likelihood of having a BI. However, there exists a population of critically ill patients who receive prolonged antibiotic regimens despite having a low BI risk. The impact of prolonged antibiotic regimens in these patients is not currently well characterized and addressing this blind spot will help critical care providers balance risks when making antibiotic de-escalation decisions.

Prospective randomized control trials (RCTs) are the most rigorous way of assessing cause-effect relationship between a treatment and outcome (195). In RCT, randomization helps reduces the risk of intergroup biases by balancing potential confounders across treatment arms. Although the gold standard, conducting RCT can require considerable investments of time and resources. Similarly, an RCT may not be feasible or ethical for some interventions, such as administering prolonged antibiotics to patients with low risk of infection (196, 197). For situations such as these, observational studies offer the possibility to potentially detect causal treatment effects using approaches to minimize confounders and treatment selection bias in nonrandomized cohorts, such as propensity score matching (198-203).

The aim of this study was to evaluate the marginal treatment effect of prolonging empiric antibiotic therapy in critical ill patients who were at low risk of having a community acquired BI. To this end, we leveraged previously validated computational phenotypes and BI risk models to identify critical care patients with low community acquired BI risk who received empiric antibiotic therapy in three unique ICU datasets. Using a propensity score-matching approach and analyzing each dataset independently, we evaluated the association between prolonging empiric antibiotic therapy past 96 hours on both in-hospital mortality and length of hospital stay (LOS).

## 5.2 **Methods**

### 5.2.1 **Study design and setting**

This was a retrospective, multi-center, observational study to compare in-hospital mortality and LOS between low BI risk patients who received short vs prolonged duration antibiotic therapy after admission to the ICU. Data for this study were obtained from two clinical data repositories and represented three distinct patient populations spanning different healthcare systems, timeframes, and locations (**Chapter 4**). A 1:1 propensity score matching case-control analysis was used, where outcomes were compared between patients who received 'short' (<96 hours) or 'prolonged' (≥96 hours) of consecutive antibiotic days and met criteria for 1) suspected of having a BI upon ICU admission, 2) had ≥1 microbiologic culture taken, 3) had no positive microbiologic culture before or during antibiotic therapy and 4) assigned Low BI risk from our previously published/validated BI risk model (145) (**Chapters 2,4**) (Figure 16). Patients who met these criteria and were treated with prolonged or short antibiotic therapy were considered cases and controls respectively. The primary endpoint was all-cause in-hospital mortality evaluated up to 28-days after the first qualifying antibiotic dose administered in the ICU ($t_0$). The secondary endpoint was length of hospital stay (LOS) associated with the qualifying ICU admission. The data in this study were analyzed and presented in accordance with Reporting and Guidelines in Propensity Score Analysis (204).

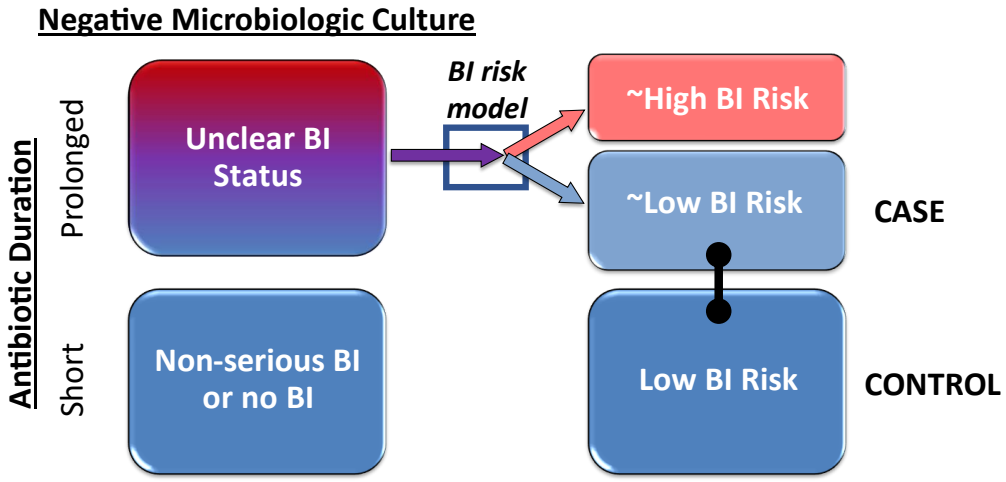


*Figure 16. Bacterial Infection risk labeling and Case-Control assignment. Patients who were suspected of having a BI upon being admitted to the ICU and had negative microbiologic cultures taken between $t_0$-24h: $t_0$+96h were selected for this study. Study patients who received short (<96 hours) therapy were considered to have low BI risk and were labeled as Control patients. Study patients who received prolonged antibiotic therapy were considered to have unclear BI risk and were predicted as low or high BI risk using previously published BI risk prediction models. Patients who received prolonged antibiotics with low predicted BI risk were labeled as Case patients, while those predicted high risk were excluded from this present study.*

### 5.2.2 **Datasets and cohort selection**

Data used for this study were obtained from the Medical Information Mart for Intensive Care III (MIMIC-III) repository and from the Northwestern Medicine (NM) Enterprise Data Warehouse (EDW) (145) (**Chapters 2,4**). MIMIC-III is a freely available and de-identified dataset collected from over 40,000 patients who received care at Beth Israel Deaconess Medical Center ICU between 2001 and 2012 (122, 123). The NMEDW is a comprehensive clinical and research data repository that encompasses encounters from across the Northwestern Medicine health system. ICU encounter information used in this study was sourced from a curated subset of 55,989 ICU encounters spanning six NM affiliated ICU in Illinois that occurred between 2011-10-01 to 2020-01-01. A detailed discussion of similarities and differences between NMEDW and MIMIC-III primary sources is presented in **Chapter 4**. NMEDW ICU encounters

were further subdivided into encounters from NM tertiary referral hospitals (NM-T) and encounters at NM-affiliated community hospitals (NM-C).

The patients used in this study were selected from the broader cohort of patients previously used to develop and externally validate our BI risk prediction model (145) (**Chapters 2,4**). In short, we selected adult patients from MIMIC, NM-T and NM-C who were suspected of having a potential BI upon admission to an ICU. Patients met these criteria if they had both antibiotics administered and had a microbiologic culture performed within 24 hours of ICU admission. Patients in the identified cohorts were allocated to one of three phenotype groups based upon their BI status: serious BI, non-serious BI/no BI, and unknown BI status. Because bacterial infections can evade microbiological culture detection in some cases, patients were assigned to BI status groups based upon their aggregated microbiology culture result and the duration of antibiotic therapy received. A detailed discussion of antibiotic selection criteria and data extraction methodology is presented in (145) (**Chapters 2,4**). Microbiology cultures were considered for this study if they were sampled during the $t_0$-24h: $t_0$+96h time window and if they were from the following specimen: blood, joint, cerebral spinal fluid, pleural cavity, peritoneum, or bronchoalveolar lavage. Qualifying microbiology cultures were classified (BI-indicated vs not-indicated) as described in detail in **Chapter 3** (163). Patient-level microbiology culture status was coded as positive if any microbiology qualifying microbiology cultures were positive for a BI. A thorough discussion of microbiologic culture selection, data extraction, and classification is presented in **Chapters 2&3** (145, 163). Within the patients who had no BI based on the microbiology results, patients who received short and prolonged antibiotics were assigned to the “non-serious BI/no BI” and “unknown BI status” groups respectively. Patients in the “non-

serious BI/no BI" group were used as prediction non-events, and patients in the "unknown BI status" group were excluded from development and validation studies.

Patients assigned to the "unknown BI" status and the "non-serious BI/no BI" status groups form the basis of our cohort for the present study (Figure 16). Here, we sought to derive a cohort of patients normalized on having a low BI risk and stratified by duration of antibiotic therapy. To achieve this, we stratified patients within the "unknown BI status" group on BI risk using our previously developed and externally validated BI risk models (**Chapters 2, 4**). Patients who received prolonged antibiotics ("unknown BI status") and had a predicted BI risk lower than the published high sensitivity decision thresholds were labeled as Cases and patients in the "non-serious BI/no BI" group were labeled as Controls. Patients in the "non-serious BI/no BI" control group were selected exclusively from the test set split (30%) used in **Chapters 2-4** (145, 163) to avoid any potential for bias introduced during model development. All data used in BI risk prediction and in subsequent propensity score modeling followed identical extraction, cleaning, and pre-processing procedures detailed at length in (145) (**Chapters 2,4**).

### 5.2.3 **Propensity score matching and analyses**

Propensity score derivation and matching were performed with the PsmPy python package using a logistic regression-based propensity model (205). Matching was performed with a K-nearest neighbor algorithm using 1:1 matching without replacement, a Euclidean distance measure, and a caliper of 0.2 standard deviations of the logit, which has been broadly supported (199, 202, 206). Our primary outcome was all-cause in-hospital mortality measured up to 28 days following $t_0$. The secondary endpoint was LOS following $t_0$. Variables for our propensity score matching were selected a priori following published recommendations (199, 207, 208) and accounted for the most common conditions associated with in-hospital mortality and hospital

stay duration in critically ill patients: multiorgan dysfunction, central nervous system failure, and cardiovascular failure (3, 70, 209-211) (Table 18). To ensure that established confounders were accounted for, we additionally included covariates into our propensity score matching from the most important predictive features in previously published sepsis, mortality, and infection status prediction models (70, 145, 212-214) (**Chapter 4**). Clinical data were modeled as minimum or maximum aggregations of all measurements taken between $t_0$ and $t_0$+24-hours (Table 18). Missing values were imputed using the median values from the associated training set (Table 18). Partial pressure of arterial oxygen ($PaO_2$)/fraction of inspired oxygen ($FiO_2$) ratios were calculated for patients who received mechanical ventilation and were imputed as 476 for everyone else. When missing (17-25% of ventilated patients across datasets), $PaO_2$ values were estimated from $SpO_2$ using the Ellis Severinghaus inversion equation when direct measurements were unavailable (215).

*Table 18. Description and missingness of PSM baseline covariates.*

| DATA COLLECTED | UNIT | AGGREGATION FUNCTION | % MISSING MIMIC | % MISSING NM-T | % MISSING NM-C |
|---|---|---|---|---|---|
| **ICD-9 CODES (ELIXHAUSER COMORBIDITY INDEX)** | ordinal | N/A | 0 | 0 | 0 |
| **AGE** | years | N/A | 0 | 0 | 0 |
| **SYSTOLIC, DIASTOLIC BLOOD PRESSURE (SBP, DBP)** | mmHg | Minimum | 0.2 | 0 | 0 |
| **GLASGOW COMA SCALE (GCS)** | GCS score | Minimum | 43.6 | 12.2 | 8.5 |
| **RESPIRATION RATE** | breath/minute | Maximum | 0.1 | 0 | 0 |
| **TEMPERATURE** | deg. C | Maximum | 1.4 | 0 | 0 |
| **VENTILATION STATUS** | categorical | N/A | 0 | 0 | 0 |
| **WEIGHT** | kg | Maximum | 11.4 | 22.9 | 8.8 |
| **VASOACTIVE ADMINISTRATION** | categorical | N/A | 0 | 0 | 0 |
| **BLOOD UREA NITROGEN (BUN)** | mg/dL | Maximum | 3.9 | 2.5 | 2.3 |
| **PARTIAL PRESSURE OF ARTERIAL OXYGEN (PAO2)/FRACTION OF INSPIRED OXYGEN (FIO2) RATIO** | ratio | Minimum | 67.9* | 41.6* | 41.5* |
| **WHITE BLOOD CELL COUNT (WBC)** | K/uL | Maximum | 5.5 | 3.2 | 3.9 |

a. Missing values were imputed using the median values from the associated training set.
b. *$PaO_2$:$FiO_2$ ratio were calculated for patients who received mechanical ventilation and imputed as 476 for everyone else.
c. Abbreviations: NM tertiary referral hospitals (NM-T); NM community hospitals (NM-C); MIMIC-III (MIMIC)

Although common, using statistical hypothesis testing to assess differences in baseline characteristics has been criticized due to their results being impacted by sample size (199, 216, 217). Here, we assessed balance of baseline characteristics by calculating standardized mean differences (SMD) between case and control groups both before and after matching on propensity scores. SMD <0.1 were considered balanced based upon published recommendations (199, 218-220).

Average treatment effect in propensity score matched individuals (ATM) was estimated for 28-day in-hospital mortality with odds ratios calculated from exact McNemar test (199, 201, 217, 221). Standard error and confidence intervals were calculated using the binomial method for exact McNemar test (222). ATM for LOS was estimated between matched case and control

patients using a paired Wilcoxon signed rank test (223-225). LOS effect estimates were presented as matched-pairs rank biserial correlation coefficient (RC) for the effect estimates as recommended for Wilcoxon paired analyses (226). Standard error of RC was calculated with bootstrapping using the wilcoxonPairedRC function in the rcompanion R (R Foundation for Statistical Computing, Vienna, Austria) package (227).

Sensitivity analyses were performed on the treatment and control groups of propensity matched patients to assess the susceptibility of our estimated ATMs to unmeasured or uncontrolled confounding. E-values were calculated on the odds ratio and RC scale to approximate the minimum strength of association that an unmeasured confounder would need to have with both the treatment selection and the outcome, conditional on the measured confounders, to fully explain away the estimated effect of treatment on outcome (228, 229). Similar to Rosenbaum's methods (225), the E-value makes no assumptions on the distributions of the estimated unmeasured confounders, and have been found to perform well for a variety of propensity score estimators and measurement error structures (230-233).

For all analyses, 2-tailed tests where $\alpha<.05$ were considered statistically significant.

## 5.3 **Results**

### 5.3.1 **Cohort**

A total of 3483, 2495, and 1487 patients fulfilled our low BI risk inclusion criteria from MIMIC, NM-T and NM-C datasets respectively and were included in the study (Table 19). In the unmatched MIMIC, NM-T and NM-C cohorts, prolonged antibiotics were administered in 40.2%, 66.4%, and 33.7% of cases. This variation in case-to-control ratio across datasets is result of differences in baseline BI prevalence observed among the three datasets (14.4%, 26.4%, and

27.3% respectively). Furthermore, the control cases for MIMIC and NM-T were sourced from patients in the corresponding test set used in model development, which represented 30% of the total "non-serious BI/no BI" cohort, while NM-C's control cases served as a validation set and were therefore never split into training and testing sets.

*Table 19. Distribution of cohort treatment assignment and outcomes before and after matching.*

| | | CONTROL SHORT ANTIBIOTICS | CASE PROLONGED ANTIBIOTICS | UNADJUSTED MORTALITY % | UNADJUSTED LOS MEDIAN (IQR) |
|---|---|---|---|---|---|
| **MIMIC** | **ALL** | 2083 | 1400 | 11.1 | 6.7 (7.0) |
| | **MATCHED** | 1347 | 1347 | 10.8 | 7.5 (7.3) |
| **NM-T** | **ALL** | 839 | 1656 | 14.0 | 7.2 (7.5) |
| | **MATCHED** | 832 | 832 | 12.6 | 6.6 (7.8) |
| **NM-C** | **ALL** | 987 | 500 | 11.7 | 4.4 (4.9) |
| | **MATCHED** | 500 | 500 | 11.1 | 4.9 (5.0) |

a. Abbreviations: NM tertiary referral hospitals (NM-T); NM community hospitals (NM-C); MIMIC-III (MIMIC); Low BI risk patients matched on propensity score (Matched); All low BI risk patients prior to matching (All)

### 5.3.2 **Propensity Score Evaluation**

The distribution of propensity scores across treatment groups and datasets before and after matching are presented in Figure 17. In all three datasets before matching, the distributions of the calculated propensity score in the treatment groups were both centered around higher means than the control groups and had a high degree of overlap.

Table 20 presents the SMD for the groups before and after matching patients on their propensity scores. All variables in MIMIC and NM-T datasets had post-match SMD of <0.1. Temperature and Glasgow coma scale (GCS) were found to be 0.16 and 0.11 respectively in the matched NM-C cohort. Although above the 0.1 arbitrary threshold, we considered these variables as having acceptable balance because they still fall below the low <0.2 effect size threshold set by Cohen (218) and have variance ratios (1.47 and 1.19 for temperature and GCS respectively) that fall into the suggested range for acceptable balance (234-236).

*Table 20. A comparison of standardized mean differences (SMD) between short and prolonged antibiotic treatment patients in unmatched and matched cohorts.*

| PROPENSITY COVARIATES | MIMIC | | NM-T | | NM-C | |
|---|---|---|---|---|---|---|
| | Pre-match SMD | Post-match SMD | Pre-match SMD | Post-match SMD | Pre-match SMD | Post-match SMD |
| **WEIGHT** | 0.002 | 0.015 | 0.126 | 0.022 | 0.074 | 0.007 |
| **YEARSOLD** | 0.054 | 0.005 | 0.023 | 0.000 | 0.158 | 0.044 |
| **ELIXHAUSER COMORBIDITY COUNT** | 0.013 | 0.009 | 0.072 | 0.019 | 0.093 | 0.054 |
| **DIASTOLIC BLOOD PRESSURE (DBP)** | 0.222 | 0.018 | 0.157 | 0.015 | 0.252 | 0.019 |
| **SYSTOLIC BLOOD PRESSURE (SBP)** | 0.320 | 0.004 | 0.121 | 0.017 | 0.285 | 0.040 |
| **VENTILATION TYPE RECEIVED** | | | | | | |
| **NONE** | 0.016 | 0.035 | 0.055 | 0.034 | 0.076 | 0.035 |
| **OXYGEN** | 0.108 | 0.010 | 0.036 | 0.002 | 0.038 | 0.100 |
| **VASOACTIVE MEDICATION RECEIVED** | 0.326 | 0.072 | 0.227 | 0.035 | 0.164 | 0.049 |
| **WHITE BLOOD CELL COUNT (WBC)** | 0.202 | 0.000 | 0.163 | 0.067 | 0.114 | 0.026 |
| **TEMPERATURE** | 0.182 | 0.034 | 0.079 | 0.014 | 0.184 | 0.162 |
| **GLASGOW COMA SCALE (GCS)** | 0.025 | 0.004 | 0.155 | 0.002 | 0.223 | 0.114 |
| **RESPIRATION RATE** | 0.109 | 0.002 | 0.229 | 0.035 | 0.117 | 0.001 |
| **BLOOD UREA NITROGEN (BUN)** | 0.241 | 0.007 | 0.105 | 0.004 | 0.126 | 0.053 |
| **PARTIAL PRESSURE OF ARTERIAL OXYGEN: FRACTION OF INSPIRED OXYGEN (PAO2:FIO2)** | 0.126 | 0.021 | 0.042 | 0.044 | 0.112 | 0.071 |

a. Abbreviations: NM tertiary referral hospitals (NM-T); NM community hospitals (NM-C); MIMIC-III (MIMIC); Low BI risk patients matched on propensity score (Matched)

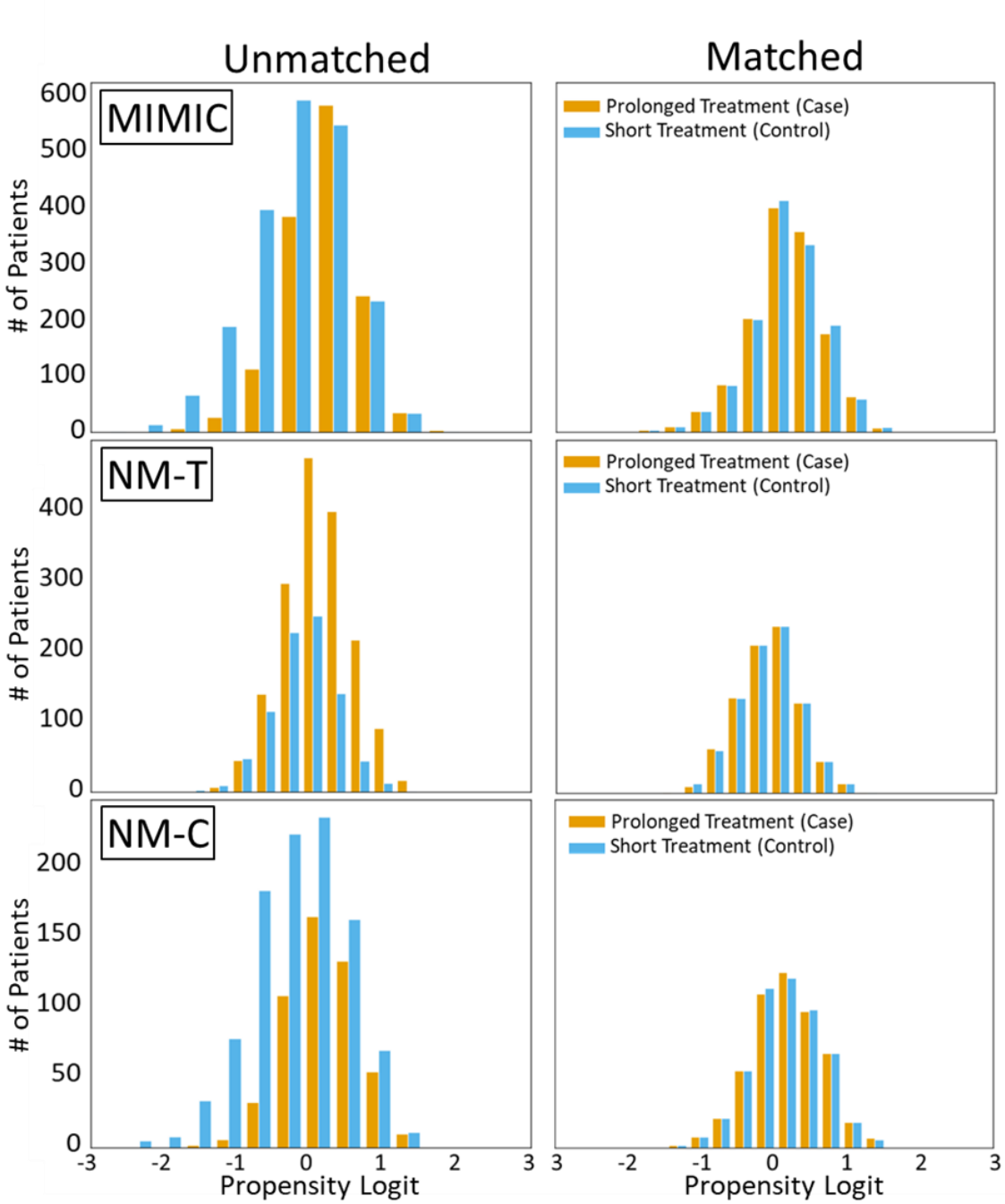


*Figure 17. Propensity score distribution before and after matching across all datasets. Prior to matching, the propensity logit distributions of patients in the prolonged treatment group are centered around larger means than patients in the short treatment groups. All patients have non-zero probabilities of being assigned to either treatment group. Similarly, most propensity scores fall within the area of common support between treatment groups. After matching, the distributions of propensity scores are largely indistinguishable.*

### 5.3.3 **Treatment Effect Assessment**

In the MIMIC, NM-T, and NM-C low BI risk cohorts matched on propensity scores, we observed significantly higher odds of 28-day in-hospital mortality in matched patients who received prolonged antibiotic therapy (Table 21). Similarly, matched patients were also found to have significant negative RC values for LOS, indicating that the LOS in the prolonged antibiotic treatment group tended to be larger than that of the short treatment group. It has been suggested that RC values above an arbitrary cutoff of >0.5 can be interpreted to indicate large effect magnitudes (226, 227).

*Table 21. Average treatment effects in matched cohorts for primary and secondary outcomes.*

| | **MIMIC** | | | **NM-T** | | | **NM-C** | | |
|---|---|---|---|---|---|---|---|---|---|
| | Short Treatment *(%/IQR)* | Prolonged Treatment *(%/IQR)* | Effect Estimate (95% CI) | Short Treatment *(%/IQR)* | Prolonged Treatment *(%/IQR)* | Effect Estimate (95% CI) | Short Treatment *(%/IQR)* | Prolonged Treatment *(%/IQR)* | Effect Estimate (95% CI) |
| Patients | 1347 (50%) | 1347 (50%) | | 832 (50%) | 832 (50%) | | 500 (50%) | 500 (50%) | |
| *In-hospital mortality* | | | | | | | | | |
| Survived | 1248 (92.6%) | 1155 (85.7%) | | 761 (91.5%) | 702 (84.4%) | | 457 (91.4%) | 425 (85.0%) | |
| Died | 99 (7.4%) | 192 (14.3%) | **2.12**** (1.82, 2.79) | 71 (8.5%) | 130 (15.6%) | **1.97**** (1.43, 2.72) | 43 (8.6%) | 75 (15.0%) | **1.94**** (1.27, 3.03) |
| *Length of stay* | | | | | | | | | |
| Median | 4.73 (5.25) | 9.58 (7.92) | **-0.57**** (-0.57, -0.63) | 4.79 (5.56) | 8.79 (8.4) | **-0.48**** (-0.55, -0.41) | 3.61 (2.99) | 6.7 (4.88) | **-0.61**** (-0.69, -0.54) |

a. For 28-day in-hospital mortality (primary outcome, dichotomous), effect estimate is presented as odds ratios and 95% confidence intervals are calculated using the binomial method for exact McNemar test.
b. For length of stay and 28-day hospital free days (secondary outcomes, continuous), effect estimates are presented as matched-pairs rank biserial correlation coefficient for Wilcoxon paired analysis and 95% confidence intervals are calculated by bootstrapping.
c. ** Denotes significance at P<0.005
d. Abbreviations: NM tertiary referral hospitals (NM-T); NM community hospitals (NM-C); MIMIC-III (MIMIC); Antibiotic treatment <96hours (Short); treatment ≥96hours (Prolonged)

We calculated E-values to assess how strong an unmeasured confounder would have to be to explain away each of the observed treatment-outcome associations (Table 22). In the matched MIMIC dataset, the 28-day in-hospital mortality E-value for the limit of the CI closest to the null was calculated as 3.04. This can be interpreted as follows: an unmeasured confounder

would have to be associated with a 3-fold increase in 28-day in-hospital mortality and 3-fold increase in prolonged antibiotic duration to explain away the lower confidence limit (229). This interpretation can be applied to all the calculated E-values presented in Table 22. Overall, the lowest E-value for the confidence interval closest to the null observed was 1.86 for 28-day in-hospital mortality in NM-C.

*Table 22. Sensitivity analysis for estimated average treatment effect in patients treated with prolonged therapy and matched individuals.*

| DATASET | OUTCOME | E-VALUE (POINT, 95% CI) |
|---|---|---|
| **MIMIC** | Mortality | 3.66 (3.04-4.28) |
| **MIMIC** | LOS | 2.75 (2.60-2.90) |
| **TERT** | Mortality | 3.35 (2.21-4.49) |
| **TERT** | LOS | 2.47 (2.26-2.68) |
| **COM** | Mortality | 3.29 (1.86-4.72) |
| **COM** | LOS | 2.88 (2.64-3.12) |

a. Abbreviations: NM tertiary referral hospitals (NM-T); NM community hospitals (NM-C); MIMIC-III (MIMIC); Length of hospital stay (LOS)

## 5.4 Discussion

This retrospective, multi-center, propensity score matched analyses investigated the effects of treating low BI risk ICU patients with prolonged vs short duration antibiotic therapy on 28-day in-hospital mortality and LOS. We found that across all three matched cohorts, patients who received prolonged antibiotics had between 1.94 - 2.12 (1.27 - 3.03 at 95% CI) higher odds of 28-day in-hospital mortality and longer LOS. Furthermore, our sensitivity analysis suggests that the estimated treatment-outcome associations were resilient against potential unmeasured confounders for antibiotic treatment assignment and outcomes. The results from our study provide additional support for the importance of de-escalating antibiotic therapy in the ICU, especially in cases where a BI risk is low.

Our results suggest that the PSM-based analyses presented here meet the two critical assumptions required for valid conclusions to be drawn from propensity score methods: 1) "strong ignorability" and 2) "common support" or "positivity" (202). Strong ignorability assumes that measured outcomes and treatment assignment are independent after accounting for confounding variables (199, 202, 224). Although not directly measurable, we indirectly evaluated strong ignorability by first evaluating the covariate balance between treatment groups after matching on propensity scores, and by performing a sensitivity analysis to estimate the strength an unmeasured confounder would need to have to explain away our observed treatment effects. Our analysis of SMD presented in Table 20 found that all variables achieved balance in matched case-control groups consistent with widely used SMD thresholds and variance-ratio criteria (199, 218-220, 237). Furthermore, our sensitivity analysis found that an unmeasured confounder would need to more than triple the odds of 28-day in-hospital mortality and would need to be more than three times as prevalent among the prolonged treatment individuals to explain away the observed association between antibiotic treatment assignment and 28-day in-hospital mortality. These results provide strong support for our propensity model to have satisfied strong ignorability. Common support assumes that, given the measured covariates, a patient needs to have a positive probability of receiving both treatment options (199, 202). The distributions of propensity scores presented in Figure 17 show that: 1) all patients have a non-zero propensity-based probability of receiving either treatment, and 2) the propensity scores of each treatment group have a high degree of overlap. Overall, our propensity model appears to satisfy the major assumptions in propensity score analysis.

Our results are aligned with other studies exploring treatment effects associated with antibiotic prescribing practices in a variety of different settings and subpopulations. The impact

of de-escalating empiric antibiotics has been broadly studied and has largely reported as having associations with reduced mortality. In two independent reviews of observational studies, empiric antibiotic de-escalation, compared to continuation of empirical treatment, was associated with a relative risk reduction in mortality of 56% (95% CI 34%-70%) and an risk ratio of 0.74 (95% CI 0.54-1.03) respectively (186, 187). Similarly, Garnacho-Montero et al. performed a prospective RCT on the impact of de-escalating empiric antibiotics in patients with severe sepsis and found de-escalation to have a significant mortality odds ratio of 0.58 (0.36-0.93) after adjusting by a propensity score (184). Leone et al. conducted a similar RCT and observed no significant reduction in risk for mortality, however this study has been criticized for being underpowered (238). In a related study, Takahashi et al. found lower rates of 28-day mortality in bacterial sepsis patients treated with short (≤7day) antibiotic courses (239). In addition to de-escalation studies, numerous other studies have found positive mortality associations related to antibiotic prescribing practices, such as rates of inappropriate empiric antibiotic therapies (188-191, 193, 240, 241) and contextual use of specific antibiotic and antibiotic classes (32-34). Our results mirror the findings that prolonged durations of antibiotic therapy can confer higher odds of in-hospital mortality and extended LOS. Our study adds to the robust body of literature on the outcome effects of prolonging antibiotic therapy by performing this observational study in a critical care cohort that has been identified as having lower risks for bacterial infection.

This study has several limitations. First, our BI prediction model was validated to predict patient-level BI risk in patients suspected of having a BI upon admission to the ICU using data collected between $t_0$: $t_0$+24h. In this study, we made the design choice to control for confounding using data collected during the same timepoints to observe the downstream cascade indicated by clinical characteristics in the first 24 hours. This creates a limitation since a patient's state can

change between the $t_0$+24h: $t_0$+96h timepoint, and those changes can have confounding effects on both their assignment to prolonged vs short antibiotic duration and on their outcomes. To account for this limitation, patients who had any documented positive microbiologic culture within the $t_0$-24h: $t_0$+96h timeframe were not eligible for this study. A second limitation of our study is that we relied upon two of our previously published classification algorithms to assemble a cohort of patients who had no microbiologic culture indication for BI and were predicted to have low BI risk based on clinical characteristics (145, 163) (**Chapters 2,4**). Although both algorithms performed well in validation studies, their use has the potential to introduce additional unmeasured bias. Similarly, our propensity model accounted for number of prior comorbidities and did not account for specific comorbidities such as immunocompromised status or oncology treatments, which could additionally impact their treatment assignment. These discussed limitations all have potential to contribute unmeasured confounding effects. To assess the severity of these limitations, we performed a sensitivity analysis to measure the strength of unmeasured confounding needed to explain away the significant association between antibiotic duration and outcomes observed (Table 22). Although residual confounding is almost certainly present after our adjustments because of design choices in this study, we believe the observed associations are both plausible and significant based upon the E-scores observed in our sensitivity analysis, and the consistency of our results with prior observational and prospective RCT studies. A final limitation of this study is that our estimand runs the risk of not corresponding to any broader target population. Because we chose to discard unmatched cases using a 0.2 caliper during our matching process to reduce bias, the estimand presented in this study no longer correspond to either the population average treatment effect nor the average treatment in the treated population, but rather to the average treatment effect in the remaining

matched sample (206). We believe that the consistency of the observed treatment effects across three distinct ICU populations supports the notion that our measured ATM approximates marginal average treatment effect in treated individuals for our primary and secondary outcomes.

## 5.5 Conclusion

In this retrospective propensity score match study, we identified a mortality and hospital duration benefit associated with discontinuing antibiotic therapy prior to 96 hours in critical ill patients predicted to have low risk for BI. Observed treatment effects were found to be resistant to unmeasured confounders and were replicated across three ICU environments from varying locations, timeframes, and healthcare organizations. The results from our study provide additional support for the importance of de-escalating antibiotic therapy in critically ill patients, especially in patients at low risk of BI.

# 6. CONCLUSION

## 6.1 Summary

In the collection of studies presented herein, we have detailed the conception, development, and evaluation of a prediction model framework that can help safely guide antibiotic de-escalation decisions in the ICU. In each of the chapters, special focus was placed on following consensus guidelines and promoting transparency. Our analyses showed that using longitudinal patient features to predict BI status with our model framework can be accomplished with high discrimination performance and can be effectively transported to external populations. Additionally, our analyses presented compelling evidence that, for critically ill patients who are predicted to be at low risk for BI, discontinuing antibiotic therapy prior to four days was associated with improvements in mortality and hospital duration. To our knowledge, these analyses are the first to utilize EHR-based clinical prediction modeling to help guide antibiotic de-escalation decisions in critically ill adults.

In **Chapter 2**, we developed and optimized a modeling framework to predict patient-level BI risk. To this end, we first established a computational phenotype for adult patients who were suspected of having a bacterial infection upon admission to the ICU. Within this cohort, we stratified patients based on the cumulative duration of empiric antibiotic therapy received (<96 or ≥96hours) and whether a bacterial infection was detected on a microbiologic culture obtained from a sterile site prior to starting EAT. Using structured longitudinal data collected up to 24, 48, and 72 hours after starting EAT, our best models identified patients at low risk of BI with AUROCs up to 0.8 and negative predictive values >93%. The work presented in **Chapter 2** provides sufficient evidence to support our first hypothesis **(H1)**: there are sufficiently granular

data in EHRs to accurately predict patient-level BI risk using raw clinical time series data of structured clinical variables. Guided by the results of our modeling studies, we determined that our Random Forests model using structured data collected for 24 hours after $t_0$ with a high sensitivity classification threshold offered the best balance between performance, parsimony, and usability. The results demonstrate the feasibility of forecasting BI risk and call for model evaluation studies to assess predictive and clinical utility.

In **Chapter 3** described the development and validation of an open-source microbiology concept extractor (MicrobEx) algorithm and package. Pursuant to the goal of externally validating our BI prediction model, MicrobEx was developed to overcome the numerous data challenges associated with extracting BI status from free-text microbiology reports. MicrobEx achieved excellent performance in two independent validation sets with minimal customization, improved performance versus a well-established alternative, and comparable performance to manual chart review by an expert. Our results suggest that MicrobEx can be used to reliably interpret binary bacterial culture status, extract bacterial species, and map these to SNOMED organism observations when applied to semi-structured, free-text microbiology reports from different institutions with relatively low customization. MicrobEx was designed to be reused and adapted to individual institutions as an upstream process for other clinical applications such as machine learning, clinical decision support, and disease surveillance systems. For our use case, MicrobEx was used as an upstream component in our data processing for external model validation and served as an invaluable tool for extracting BI status information used in our computational phenotype.

In **Chapter 4**, we performed an external validation of BI modeling architecture (**Chapter 2**) in two tertiary intensive care unit (ICU) settings and a community ICU setting. We

additionally explored how simple multisite learning (data pooling and model ensembling) techniques impacted model transportability. 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. These results provide evidence that our novel BI modeling framework has predictive utility when transported to external validation cohorts and therefore support our second hypothesis **(H2)**. However, our results also demonstrate that differences in case-mix and predictor effects can impact model transportability, even when the same framework for data extraction, processing, and model recalibration are followed in each dataset. While this result is not surprising, it is an important reminder that developing a prediction model that is valid and consistent across diverse populations may be unrealistic. We recommend that institutions seeking to implement an externally developed prediction model: 1) select model(s) developed from data that share similar case-mix and predictor effects and 2) evaluate and update the chosen model(s) in the intended cohort(s) prior to evaluating the model's clinical utility. Although models developed using multisite learning have the potential to enhance class differentiation and performance stability, the effectiveness of such models cannot be guaranteed and should be assessed individually. Finally, predictive utility is necessary but not sufficient to demonstrate clinical utility. Once externally validated, a model should be tested in its intended clinical context to measure the impact on patient care.

In the current paradigm of ICU antibiotic prescribing practices, there exists populations of critically ill patients who despite having low risk for BI, receive unnecessarily prolonged empiric antibiotic regimens. In fact, it's estimated that up to 33% of all inpatient antibiotic therapy days are unnecessary (15-20, 35, 36). In **Chapter 5**, we performed a retrospective impact study to estimate the treatment effect of prolonging antibiotic therapy past 96 hours in critical ill patients predicted to have low risk for BI using data from three distinct ICU environments and adjusting for selection bias using propensity score matching. After adjusting for selection bias, we found a strong association between prolonged empiric antibiotic therapy and increased risk of in-hospital mortality and longer length of hospital stay. Additionally, sensitivity analysis suggests the observed associations are strongly resistant to unmeasured confounding on either the treatment or outcome side. Identifying critically ill patients with low risk of infection and testing the treatment impact of short vs prolonged antibiotic therapies in a prospective setting would be logically challenging and ethically complicated. However, leveraging our validated BI risk model and utilizing observational study confounding adjustments allowed us to perform this experiment retrospectively. The analyses performed in **Chapter 5** produced evidentiary support for our third hypothesis **(H3)**: patients identified as low BI risk by our model have worse outcomes when empiric antibiotic therapy is prolonged past 96 hours. Improper antibiotic de-escalation practices in cohorts with confirmed BI or high likelihood of having a BI have been associated with higher morbidity and mortality rates in previous observational and RCT studies (184, 186, 187). Our findings represent a significant contribution to the already robust body of literature surrounding antibiotic prescribing practices and patient outcomes by performing this observational study in a critical care cohort that has been identified as having lower risks for bacterial infection. Furthermore, our findings support the notion that when paired with current

evidence-based practices, our BI risk model will have a reasonable chance to positively impact antibiotic de-escalation decision making and subsequent outcomes.

## 6.2 Future Direction

While this dissertation adds novel findings, tools, and a validated BI risk prediction model to the literature, there are remaining research areas left to explore. Before a model is used in real-time, it's critical to ensure that the model is likely to benefit clinical decision making and is unlikely to cause iatrogenic harm (97, 105, 242). To this end, the results presented in **Chapter 4** provide a verification that our BI model maintains a sufficient level of predictive utility in multiple groups of individuals other than from which it was developed. Similarly, the work presented in **Chapter 5** demonstrates that the decision to prolong antibiotic therapy past 4 days is associated with worse outcomes in patients predicted to have low BI risk. Together, these results provide evidence that the use of our BI prediction model to identify patients with low BI risk will indeed have a positive effect on both antibiotic de-escalation decisions and patient outcomes. The next step in our research is to test this explicitly with a cluster-randomized clinical impact study (for example, with a cluster-randomized controlled trial [RCT]).

Designing an appropriate large-scale prediction model comparison study has numerous important considerations. Following published guidelines, our RCT would recruit two groups of critical care providers where one group would be exposed to the output of our prediction model (index group) and one will proceed with their current standard-of-care (control) (97, 105, 242). The clinical impact of our model will be evaluated based upon the observed group differences for antibiotic de-escalation decision making (i.e., impact of the model on decisions) and patient outcomes (i.e., impact of the decisions on outcome). While a prospective and randomized study design is the most effective way of achieving balance between index and control groups,

observational impact studies require fewer resources and emerging methods to perform them are becoming increasingly popular (101-104). In addition to study design, researchers in Kappen et al. also stress the importance of tailoring the modeling approach to each specific setting where our model will be studied (242). While no clear set of guidelines exists for this tailoring process, our findings and recommendations in **Chapter 4** can help serve as a starting point. Here, it will be important to find setting(s) where case-mix differences and predictor effects are expected to be like that of our development cohorts. Additionally, model predictive utility in the study setting should be confirmed and tailored via calibration and or updating prior to studying model impact.

## 6.3 Concluding Remarks

The work presented in this thesis represents numerous significant and original contributions to informatics knowledge. Collectively, our results suggest that 1) it is feasible to predict patient-level BI risk using structured EHR features collected in the first day after antibiotic treatment is initiated, 2) trained BI risk models can be transported effectively to external populations, and perform best when choosing populations with similar case-mix and predictor effects, 3) critically ill patients predicted to be at low risk for BI are associated with improved outcomes after discontinuing antibiotic therapy prior to 96 hours. From a methodological perspective, developing novel data preparation procedures and performing robust model evaluations proved to be a powerful approach for developing a transparent and transportable modeling framework. The findings outlined above have laid the foundation for a future large-scale prediction model implementation study.

# 7. REFERENCES

1. Vincent J-L, Rello J, Marshall J, Silva E, Anzueto A, Martin CD, et al. International Study of the Prevalence and Outcomes of Infection in Intensive Care Units. JAMA. 2009;302(21):2323-9.

2. Vincent JL, Abraham E, Annane D, Bernard G, Rivers E, Van den Berghe G. Reducing mortality in sepsis: new directions. Crit Care. 2002;6 Suppl 3:S1-18.

3. Mayr FB, Yende S, Angus DC. Epidemiology of severe sepsis. Virulence. 2014;5(1):4-11.

4. Angus DC, Linde-Zwirble WT, Lidicker J, Clermont G, Carcillo J, Pinsky MR. Epidemiology of severe sepsis in the United States: analysis of incidence, outcome, and associated costs of care. Critical care medicine. 2001;29(7):1303-10.

5. Vincent J-L, Sakr Y, Singer M, Martin-Loeches I, Machado FR, Marshall JC, et al. Prevalence and Outcomes of Infection Among Patients in Intensive Care Units in 2017. JAMA. 2020;323(15):1478-87.

6. Lakbar I, Munoz M, Pauly V, Orleans V, Fabre C, Fond G, et al. Septic shock: incidence, mortality and hospital readmission rates in French intensive care units from 2014 to 2018. Anaesthesia Critical Care & Pain Medicine. 2022;41(3):101082.

7. Vincent J-L, Jones G, David S, Olariu E, Cadwell KK. Frequency and mortality of septic shock in Europe and North America: a systematic review and meta-analysis. Critical care. 2019;23(1):1-11.

8. Wolfertz N, Böhm L, Keitel V, Hannappel O, Kümpers P, Bernhard M, et al. Epidemiology, management, and outcome of infection, sepsis, and septic shock in a German emergency department (EpiSEP study). Front Med (Lausanne). 2022;9:997992.

9. Delia M, Monica L. Infections and Multidrug-Resistant Pathogens in ICU Patients. In: Erbay RzH, editor. Current Topics in Intensive Care Medicine. Rijeka: IntechOpen; 2018. p. Ch. 2.

10. Garland A, Olafson K, Ramsey CD, Yogendran M, Fransoo R. Epidemiology of critically ill patients in intensive care units: a population-based observational study. Crit Care. 2013;17(5):R212.

11. Kaplan V, Angus DC, Griffin MF, Clermont G, Scott Watson R, Linde-Zwirble WT. Hospitalized community-acquired pneumonia in the elderly: age- and sex-related patterns of care and outcome in the United States. American journal of respiratory and critical care medicine. 2002;165(6):766-72.

12. Morris AC. Management of pneumonia in intensive care. Journal of Emergency and Critical Care Medicine. 2018;2.

13. Gulani P, Chen J, Keene A. Antimicrobials in the ICU. In: Oropello JM, Pastores SM, Kvetan V, editors. Critical Care. New York, NY: McGraw-Hill Education.

14. Timsit J-F, Bassetti M, Cremer O, Daikos G, de Waele J, Kallil A, et al. Rationalizing antimicrobial therapy in the ICU: a narrative review. Intensive care medicine. 2019;45(2):172-89.

15. More evidence on link between antibiotic use and antibiotic resistance. ScienceDaily: European Centre for Disease Prevention and Control (ECDC); 2017 07/27/2017.

16. Antimicrobial resistance: global report on surveillance. World Health Organization; 2014.

17. Shallcross LJ, Davies DSC. Antibiotic overuse: a key driver of antimicrobial resistance. Br J Gen Pract. 2014;64(629):604-5.

18. Michael CA, Dominey-Howes D, Labbate M. The Antimicrobial Resistance Crisis: Causes, Consequences, and Management. Frontiers in Public Health. 2014;2:145.

19. Core Elements of Hospital Antibiotic Stewardship Programs | Antibiotic Use | CDC. 2019.

20. Camins BC, King MD, Wells JB, Googe HL, Patel M, Kourbatova EV, et al. Impact of an antimicrobial utilization program on antimicrobial use at a large teaching hospital: a randomized controlled trial. Infect Control Hosp Epidemiol. 2009;30(10):931-8.

21. Kollef MH, Shorr AF, Bassetti M, Timsit J-F, Micek ST, Michelson AP, et al. Timing of antibiotic therapy in the ICU. Critical Care. 2021;25(1):360.

22. Liu VX, Fielding-Singh V, Greene JD, Baker JM, Iwashyna TJ, Bhattacharya J, et al. The Timing of Early Antibiotics and Hospital Mortality in Sepsis. American journal of respiratory and critical care medicine. 2017;196(7):856-63.

23. Kumar A, Roberts D, Wood KE, Light B, Parrillo JE, Sharma S, et al. Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. Critical care medicine. 2006;34(6):1589-96.

24. Seymour CW, Gesten F, Prescott HC, Friedrich ME, Iwashyna TJ, Phillips GS, et al. Time to treatment and mortality during mandated emergency care for sepsis. New England Journal of Medicine. 2017;376(23):2235-44.

25. Tamma PD, Avdic E, Li DX, Dzintars K, Cosgrove SE. Association of Adverse Events With Antibiotic Use in Hospitalized Patients. JAMA Internal Medicine. 2017;177(9):1308-15.

26. Claridge JA, Pang P, Leukhardt WH, Golob JF, Carter JW, Fadlalla AM. Critical analysis of empiric antibiotic utilization: establishing benchmarks. Surgical infections. 2010;11(2):125-31.

27. Francino MP. Antibiotics and the Human Gut Microbiome: Dysbioses and Accumulation of Resistances. Frontiers in Microbiology. 2015;6:1543.

28. Thomas Z, Bandali F, Sankaranarayanan J, Reardon T, Olsen KM. A Multicenter Evaluation of Prolonged Empiric Antibiotic Therapy in Adult ICUs in the United States. Critical care medicine. 2015;43(12):2527-34.

29. Teshome BF, Vouri SM, Hampton NB, Kollef MH, Micek ST. Evaluation of a ceiling effect on the association of new resistance development to antipseudomonal beta-lactam exposure in the critically ill. Infection Control & Hospital Epidemiology. 2020;41(4):484-5.

30. Teshome BF, Vouri SM, Hampton N, Kollef MH, Micek ST. Duration of Exposure to Antipseudomonal β-Lactam Antibiotics in the Critically Ill and Development of New Resistance. Pharmacotherapy. 2019;39(3):261-70.

31. Björnsson ES. Drug-induced liver injury due to antibiotics. Scand J Gastroenterol. 2017;52(6-7):617-23.

32. Gorelik E, Masarwa R, Perlman A, Rotshild V, Abbasi M, Muszkat M, et al. Fluoroquinolones and Cardiovascular Risk: A Systematic Review, Meta-analysis and Network Meta-analysis. Drug Safety. 2019;42(4):529-38.

33. Mosholder AD, Lee J-Y, Zhou EH, Kang EM, Ghosh M, Izem R, et al. Long-term risk of acute myocardial infarction, stroke, and death with outpatient use of clarithromycin: a retrospective cohort study. American journal of epidemiology. 2018;187(4):786-92.

34. Heianza Y, Ma W, Li X, Cao Y, Chan AT, Rimm EB, et al. Duration and life-stage of antibiotic use and risks of all-cause and cause-specific mortality: prospective cohort study. Circulation research. 2020;126(3):364-73.

35. Versporten A, Zarb P, Caniaux I, Gros M-F, Drapier N, Miller M, et al. Antimicrobial consumption and resistance in adult hospital inpatients in 53 countries: results of an internet-based global point prevalence survey. The Lancet Global Health. 2018;6(6):e619-e29.

36. Magill SS, Edwards JR, Beldavs ZG, Dumyati G, Janelle SJ, Kainer MA, et al. Prevalence of antimicrobial use in US acute care hospitals, May-September 2011. Jama. 2014;312(14):1438-46.

37. Dryden M, Johnson AP, Ashiru-Oredope D, Sharland M. Using antibiotics responsibly: right drug, right time, right dose, right duration. J Antimicrob Chemother. 2011;66(11):2441-3.

38. Wunderink RG, Srinivasan A, Barie PS, Chastre J, Cruz CSD, Douglas IS, 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. Annals of the American Thoracic Society. 2020;17(5):531-40.

39. Braykov NP, Morgan DJ, Schweizer ML, Uslan DZ, Kelesidis T, Weisenberg SA, et al. Assessment of empirical antibiotic therapy optimisation in six hospitals: an observational cohort study. The Lancet Infectious Diseases. 2014;14(12):1220-7.

40. Bergmans D, Bonten M, Gaillard C, Van Tiel F, Van Der Geest S, De Leeuw P, et al. Indications for antibiotic use in ICU patients: a one-year prospective surveillance. The Journal of antimicrobial chemotherapy. 1997;39(4):527-35.

41. Edition AGF. CLSI document M29-A4. May; 2014.

42. Luyt C-E, Bréchot N, Trouillet J-L, Chastre J. Antibiotic stewardship in the intensive care unit. Critical Care. 2014;18(5):480.

43. Luna C, Blanzaco D, Niederman M, Matarucco W, Baredes N, Desmery P, et al. Resolution of ventilator-associated pneumonia: Prospective evaluation of the clinical pulmonary infection score as an early clinical predictor of outcome*. Critical care medicine. 2019;31(3):676-82.

44. Wu J, Tang B, Qiu Y, Tan R, Liu J, Xia J, et al. Clinical validation of a multiplex droplet digital PCR for diagnosing suspected bloodstream infections in ICU practice: a promising diagnostic tool. Crit Care. 2022;26(1):243.

45. Schmitz JE, Stratton CW, Persing DH, Tang Y-W. Forty Years of Molecular Diagnostics for Infectious Diseases. Journal of Clinical Microbiology. 2022;60(10):e02446-21.

46. Coburn B, Morris AM, Tomlinson G, Detsky AS. Does this adult patient with suspected bacteremia require blood cultures? Jama. 2012;308(5):502-11.

47. Yu Y, Li X-X, Jiang L-X, Du M, Liu Z-G, Cen Z-R, et al. Procalcitonin levels in patients with positive blood culture, positive body fluid culture, sepsis, and severe sepsis: a cross-sectional study. Infectious diseases. 2016;48(1):63-9.

48. Laukemann S, Kasper N, Kulkarni P, Steiner D, Rast AC, Kutz A, et al. Can we reduce negative blood cultures with clinical scores and blood markers? Results from an observational cohort study. Medicine. 2015;94(49).

49. Turner P, Fox-Lewis A, Shrestha P, Dance DAB, Wangrangsimakul T, Cusack TP, 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.

50. Heilmann E, Gregoriano C, Schuetz P. Biomarkers of Infection: Are They Useful in the ICU? Semin Respir Crit Care Med. 2019;40(4):465-75.

51. Heffernan AJ, Denny KJ. Host Diagnostic Biomarkers of Infection in the ICU: Where Are We and Where Are We Going? Curr Infect Dis Rep. 2021;23(4):4.

52. Pierrakos C, Velissaris D, Bisdorff M, Marshall JC, Vincent J-L. Biomarkers of sepsis: time for a reappraisal. Critical Care. 2020;24(1):1-15.

53. Póvoa P, Coelho L, Dal-Pizzol F, Ferrer R, Huttner A, Conway Morris A, et al. How to use biomarkers of infection or sepsis at the bedside: guide to clinicians. Intensive care medicine. 2023;49(2):142-53.

54. Denny KJ, De Wale J, Laupland KB, Harris PNA, Lipman J. When not to start antibiotics: avoiding antibiotic overuse in the intensive care unit. Clinical Microbiology and Infection. 2019(1469-0691 (Electronic)).

55. Mandell LA, Wunderink RG, Anzueto A, Bartlett JG, Campbell GD, Dean NC, et al. Infectious Diseases Society of America/American Thoracic Society consensus guidelines on the management of community-acquired pneumonia in adults. Clinical infectious diseases. 2007;44(Supplement_2):S27-S72.

56. Friedman CP, Wong AK, Blumenthal D. Achieving a nationwide learning health system. Sci Transl Med. 2010;2(57):57cm29.

57. Non-federal Acute Care Hospital Electronic Health Record Adoption. In: Technology OotNCfHI, editor. 2019.

58. Kruse CS, Kristof C, Jones B, Mitchell E, Martinez A. Barriers to Electronic Health Record Adoption: a Systematic Literature Review. Journal of Medical Systems. 2016;40(12):252.

59. Goh KH, Wang L, Yeow AYK, Poh H, Li K, Yeow JJL, et al. Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. Nature communications. 2021;12(1):711.

60. van Doorn WP, Stassen PM, Borggreve HF, Schalkwijk MJ, Stoffers J, Bekers O, et al. A comparison of machine learning models versus clinical evaluation for mortality prediction in patients with sepsis. PLoS One. 2021;16(1):e0245157.

61. Fleuren LM, Klausch TLT, Zwager CL, Schoonmade LJ, Guo T, Roggeveen LF, et al. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. Intensive care medicine. 2020.

62. Ward L, Paul M, Andreassen S. Automatic learning of mortality in a CPN model of the systemic inflammatory response syndrome. Math Biosci. 2017;284:12-20.

63. Paul M, Andreassen S, Nielsen AD, Tacconelli E, Almanasreh N, Fraser A, et al. Prediction of bacteremia using TREAT, a computerized decision-support system. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America. 2006;42(9):1274-82.

64. Sheetrit E, Nissim N, Klimov D, Shahar Y. Temporal Probabilistic Profiles for Sepsis Prediction in the ICU. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; Anchorage, AK, USA. 3330747: ACM; 2019. p. 2961-9.

65. Gultepe E, Green JP, Nguyen H, Adams J, Albertson T, Tagkopoulos I. From vital signs to clinical outcomes for patients with sepsis: a machine learning basis for a clinical decision support system. J Am Med Inform Assoc. 212014. p. 315-25.

66. Brause R, Hamker F, Paetz J. Septic Shock Diagnosis by Neural Networks and Rule Based Systems. Computational intelligence techniques in medical diagnosis and prognosis: SpringerLink; 2002.

67. Peelen L, de Keizer NF, Jonge E, Bosman RJ, Abu-Hanna A, Peek N. Using hierarchical dynamic Bayesian networks to investigate dynamics of organ failure in patients in the Intensive Care Unit. J Biomed Inform. 2010;43(2):273-86.

68. 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. Critical care medicine. 2018;46(4):547-53.

69. Parente JD, Möller K, Shaw GM, Chase JG. Hidden Markov Models for Sepsis Classification. IFAC-PapersOnLine. 2018;51(27):110-5.

70. 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. Annals of Emergency Medicine. 2021;77(4):395-406.

71. 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):95.

72. 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. 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2021: IEEE.

73. Wang H, Li Y, Naidech A, Luo Y. Comparison between machine learning methods for mortality prediction for sepsis patients with different social determinants. BMC medical informatics and decision making. 2022;22(2):1-13.

74. Brown SM, Jones J, Kuttler KG, Keddington RK, Allen TL, Haug P. Prospective evaluation of an automated method to identify patients with severe sepsis or septic shock in the emergency department. BMC Emergency Medicine. 2016;16(1):31.

75. Tarabichi Y, Cheng A, Bar-Shain D, McCrate BM, Reese LH, Emerman C, et al. Improving timeliness of antibiotic administration using a provider and pharmacist facing sepsis early warning system in the emergency department setting: a randomized controlled quality improvement initiative. Critical care medicine. 2021;50(3):418-27.

76. Wong A, Otles E, Donnelly JP, Krumm A, McCullough J, DeTroyer-Cooley O, et al. External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients. JAMA Internal Medicine. 2021;181(8):1065-70.

77. Zhang J, Mattie H, Shuaib H, Hensman T, Teo JT, Celi LA. Addressing the "elephant in the room" of AI clinical decision support through organisation-level regulation. PLOS Digit Health. 2022;1(9):e0000111.

78. Hwang EJ, Park S, Jin K-N, Kim JI, Choi SY, Lee JH, et al. Development and Validation of a Deep Learning–Based Automated Detection Algorithm for Major Thoracic Diseases on Chest Radiographs. JAMA Network Open. 2019;2(3):e191095-e.

79. Lamping F, Jack T, Rübsamen N, Sasse M, Beerbaum P, Mikolajczyk RT, et al. Development and validation of a diagnostic model for early differentiation of sepsis and non-infectious SIRS in critically ill children-a data-driven approach using machine-learning algorithms. BMC pediatrics. 2018;18(1):1-11.

80. Roux-Dalvai F, Gotti C, Leclercq M, Hélie M-C, Boissinot M, Arrey TN, et al. Fast and Accurate Bacterial Species Identification in Urine Specimens Using LC-MS/MS Mass Spectrometry and Machine Learning*[S]. Molecular & cellular proteomics. 2019;18(12):2492-505.

81. Bystritsky RJ, Beltran A, Young AT, Wong A, Hu X, Doernberg SB. Machine learning for the prediction of antimicrobial stewardship intervention in hospitalized patients receiving broad-spectrum agents. Infection Control & Hospital Epidemiology. 2020;41(9):1022-7.

82. Beaudoin M, Kabanza F, Nault V, Valiquette L. Evaluation of a machine learning capability for a clinical decision support system to enhance antimicrobial stewardship programs. Artificial intelligence in medicine. 2016;68:29-36.

83. Mashoufi M, Ayatollahi H, Khorasani-Zavareh D. A review of data quality assessment in emergency medical services. The open medical informatics journal. 2018;12(1).

84. 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 Annual Symposium proceedings AMIA Symposium. 2022;2022:196-205.

85. Tu K, Widdifield J, Young J, Oud W, Ivers NM, Butt DA, et al. Are family physicians comprehensively using electronic medical records such that the data can be used for secondary purposes? A Canadian perspective. BMC medical informatics and decision making. 2015;15(1):1-12.

86. Greiver M, Barnsley J, Glazier RH, Harvey BJ, Moineddin R. Measuring data reliability for preventive services in electronic medical records. BMC health services research. 2012;12:1-9.

87. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. Journal of the American Medical Informatics Association. 2013;20(1):144-51.

88. Hersh WR, Weiner MG, Embi PJ, Logan JR, Payne PR, Bernstam EV, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. Med Care. 2013;51(8 Suppl 3):S30-7.

89. Reisman M. EHRs: The Challenge of Making Electronic Data Usable and Interoperable. P t. 2017;42(9):572-5.

90. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc. 2010;17(3):229-36.

91. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc. 2010;17(5):507-13.

92. Fu S, Chen D, He H, Liu S, Moon S, Peterson KJ, et al. Clinical concept extraction: A methodology review. Journal of Biomedical Informatics. 2020;109:103526.

93. van Os HJA, Kanning JP, Wermer MJH, Chavannes NH, Numans ME, Ruigrok YM, et al. Developing Clinical Prediction Models Using Primary Care Electronic Health Record Data: The Impact of Data Preparation Choices on Model Performance. Frontiers in Epidemiology. 2022;2.

94. Maletzky A, Böck C, Tschoellitsch T, Roland T, Ludwig H, Thumfart S, et al. Lifting Hospital Electronic Health Record Data Treasures: Challenges and Opportunities. JMIR Med Inform. 2022;10(10):e38557.

95. Terry AL, Stewart M, Cejic S, Marshall JN, de Lusignan S, Chesworth BM, et al. A basic model for assessing primary health care electronic medical record data quality. BMC Medical Informatics and Decision Making. 2019;19:1-11.

96. Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Med Res Methodol. 2014;14:40.

97. Steyerberg EW. Clinical prediction models : a practical approach to development, validation, and updating. Cham, Switzerland: Springer; 2019. Available from: https://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=2204948

https://public.ebookcentral.proquest.com/choice/publicfullrecord.aspx?p=5836937

https://doi.org/10.1007/978-3-030-16399-0

https://link.springer.com/book/10.1007%2F978-3-030-16399-0

http://www.vlebooks.com/vleweb/product/openreader?id=none&isbn=9783030163990

https://link.springer.com/book/10.1007/978-3-030-16399-0.

98. Pencina MJ, Goldstein BA, D'Agostino RB. Prediction models-development, evaluation, and clinical application. The New England journal of medicine. 2020;382(17):1583-6.

99. Van Calster B, Wynants L, Timmerman D, Steyerberg EW, Collins GS. Predictive analytics in health care: how can we know it works? Journal of the American Medical Informatics Association. 2019;26(12):1651-4.

100. Yang C, Kors JA, Ioannou S, John LH, Markus AF, Rekkas A, et al. Trends in the conduct and reporting of clinical prediction model development and validation: a systematic review. J Am Med Inform Assoc. 2022;29(5):983-9.

101. Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. PLoS medicine. 2013;10(2):e1001381.

102. Moons KGM, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart. 2012;98(9):691-8.

103. Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. Annals of internal medicine. 2015;162(1):W1-W73.

104. Marwaha JS, Kvedar JC. Crossing the chasm from model performance to clinical impact: the need to improve implementation and evaluation of AI. npj Digital Medicine. 2022;5(1):25.

105. Binuya MAE, Engelhardt EG, Schats W, Schmidt MK, Steyerberg EW. Methodological guidance for the evaluation and updating of clinical prediction models: a systematic review. BMC Medical Research Methodology. 2022;22(1):316.

106. Golas SB, Nikolova-Simons M, Palacholla R, op den Buijs J, Garberg G, Orenstein A, et al. Predictive analytics and tailored interventions improve clinical outcomes in older adults: a randomized controlled trial. npj Digital Medicine. 2021;4(1):97.

107. Watson J, Hutyra CA, Clancy SM, Chandiramani A, Bedoya A, Ilangovan K, et al. Overcoming barriers to the adoption and implementation of predictive modeling and machine learning in clinical care: what can we learn from US academic medical centers? JAMIA Open. 2020;3(2):167-72.

108. Jung K, Kashyap S, Avati A, Harman S, Shaw H, Li R, et al. A framework for making predictive models useful in practice. Journal of the American Medical Informatics Association. 2021;28(6):1149-58.

109. Zilahi G, McMahon MA, Povoa P, Martin-Loeches I. Duration of antibiotic therapy in the intensive care unit. Journal of Thoracic Disease. 2016;8(12):3774-80.

110. Weiss CH, Persell SD, Wunderink RG, Baker DW. Empiric antibiotic, mechanical ventilation, and central venous catheter duration as potential factors mediating the effect of a checklist prompting intervention on mortality: an exploratory analysis. BMC health services research. 2012;12:198.

111. Arulkumaran N, Routledge M, Schlebusch S, Lipman J, Conway Morris A. Antimicrobial-associated harm in critical care: a narrative review. Intensive care medicine. 2020.

112. Surveillance of Antimicrobial Resistance in Europe. In: Control ECfDPa, editor. 2017.

113. Prestinaci F, Pezzotti P, Pantosti A. Antimicrobial resistance: a global multifaceted phenomenon. Pathog Glob Health. 2015;109(7):309-18.

114. Dadgostar P. Antimicrobial Resistance: Implications and Costs. Infection and Drug Resistance. 2019;Volume 12:3903-10.

115. Antibiotic resistance threats in the United States, 2013. In: Prevention CfDCa, Services UDoHaH, editors. 2013.

116. Bell BG, Schellevis F, Stobberingh E, Goossens H, Pringle M. A systematic review and meta-analysis of the effects of antibiotic consumption on antibiotic resistance. BMC Infectious Diseases. 2014;14(1):13.

117. Goff DA, File TM. The risk of prescribing antibiotics “just-in-case” there is infection. Seminars in Colon and Rectal Surgery. 2018;29(1):44-8.

118. Andre K, Mark M, Michael K, John M, Daniel S, Lucy P, et al. Guidelines for the management of adults with hospital-acquired, ventilator-associated, and healthcare-associated pneumonia. American journal of respiratory and critical care medicine. 2005;171(4):388-416.

119. Dellinger RP, Levy MM, Rhodes A, Annane D, Gerlach H, Opal SM, et al. Surviving sepsis campaign: international guidelines for management of severe sepsis and septic shock: 2012. Critical care medicine. 2013;41(2):580-637.

120. Solomkin JS, Mazuski JE, Bradley JS, Rodvold KA, Goldstein EJ, Baron EJ, et al. Diagnosis and management of complicated intra-abdominal infection in adults and children: guidelines by the Surgical Infection Society and the Infectious Diseases Society of America. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America. 2010;50(2):133-64.

121. Zimmerman JJ. Society of Critical Care Medicine Presidential Address−47th Annual Congress, February 2018, San Antonio, Texas. Critical care medicine. 2018;46(6):839-42.

122. Johnson AEW, Pollard TJ. The MIMIC-III Clinical Database. 2016.

123. Johnson AEW, Pollard TJ, Shen L, Lehman L-wH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Scientific Data. 2016;3(1):160035.

124. Luo Y, Szolovits P, Dighe AS, Baron JM. 3D-MICE: integration of cross-sectional and longitudinal imputation for multi-analyte longitudinal clinical data. (1527-974X (Electronic)).

125. Luo Y, Szolovits P, Dighe AS, Baron JM. Using Machine Learning to Predict Laboratory Test Results. (1943-7722 (Electronic)).

126. Le Cessie S, Van Houwelingen JC. Ridge Estimators in Logistic Regression. Journal of the Royal Statistical Society Series C (Applied Statistics). 1992;41(1):191-201.

127. Breiman L. Random Forests. Machine Learning. 2001;45(1):5-32.

128. Cortes C, Vapnik V. Support-Vector Networks. Machine Learning. 1995;20(3):273-97.

129. Friedman JH. Stochastic gradient boosting. Computational Statistics & Data Analysis. 2002;38(4):367-78.

130. Blot F, Raynard B, Chachaty E, Tancrède C, Antoun S, Nitenberg G. Value of Gram Stain Examination of Lower Respiratory Tract Secretions for Early Diagnosis of Nosocomial Pneumonia. http://dxdoiorg/101164/ajrccm16259908088. 2000.

131. Campion M, Scully G. Antibiotic Use in the Intensive Care Unit: Optimization and De-Escalation. J Intensive Care Med. 2018;33(12):647-55.

132. Samuel LP, Balada-Llasat J-M, Harrington A, Cavagnolo R, Bourbeau P. Multicenter Assessment of Gram Stain Error Rates. 2016.

133. de Jong E, van Oers JA, Beishuizen A, Vos P, Vermeijden WJ, Haas LE, et al. Efficacy and safety of procalcitonin guidance in reducing the duration of antibiotic treatment in critically ill patients: a randomised, controlled, open-label trial. The Lancet Infectious Diseases. 2016;16(7):819-27.

134. Schuetz P, Wirz Y, Sager R, Christ-Crain M, Stolz D, Tamm M, et al. Procalcitonin to initiate or discontinue antibiotics in acute respiratory tract infections. The Cochrane database of systematic reviews. 2017;10:Cd007498.

135. Cals JW, Ebell MH. C-reactive protein: guiding antibiotic prescribing decisions at the point of care. Br J Gen Pract. 2018;68(668):112-3.

136. Paonessa JR, Shah RD, Pickens CI, Lizza BD, Donnelly HK, Malczynski M, et al. Rapid Detection of Methicillin-Resistant Staphylococcus aureus in BAL: A Pilot Randomized Controlled Trial. Chest. 2019;155(5):999-1007.

137. Ward L, Møller JK, Eliakim-Raz N, Andreassen S. Prediction of Bacteraemia and of 30-day Mortality Among Patients with Suspected Infection using a CPN Model of Systemic Inflammation. IFAC-PapersOnLine. 2018;51(27):116-21.

138. Vieira SM, Carvalho JP, Fialho AS, Reti SR, Finkelstein SN, Sousa JMC. A Decision Support System for ICU Readmissions Prevention. Proceedings of the 2013 Joint Ifsa World Congress and Nafips Annual Meeting (Ifsa/Nafips). 2013:251-6.

139. Luo Y, Xin Y, Joshi R, Celi L, Szolovits P, editors. Predicting ICU mortality risk by grouping temporal trends from a multivariate panel of physiologic measurements. Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence; 2016 02/12/2016: AAAI Press.

140. Curto S, Carvalho JP, Salgado C, Vieira SM, Sousa JMC, editors. Predicting ICU readmissions based on bedside medical text notes. 2016 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE); 2016 24-29 July 2016.

141. Miller JM, Binnicker MJ, Campbell S, Carroll KC, Chapin KC, Gilligan PH, et al. A Guide to Utilization of the Microbiology Laboratory for Diagnosis of Infectious Diseases: 2018 Update by the Infectious Diseases Society of America and the American Society for Microbiology. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America. 2018;67(6):e1-e94.

142. Rhoads DD, Sintchenko V, Rauch CA, Pantanowitz L. Clinical microbiology informatics. Clin Microbiol Rev. 2014;27(4):1025-47.

143. Graham PL, 3rd, San Gabriel P, Lutwick S, Haas J, Saiman L. Validation of a multicenter computer-based surveillance system for hospital-acquired bloodstream infections in neonatal intensive care departments. Am J Infect Control. 2004;32(4):232-4.

144. Bellini C, Petignat C, Francioli P, Wenger A, Bille J, Klopotov A, et al. Comparison of automated strategies for surveillance of nosocomial bacteremia. Infect Control Hosp Epidemiol. 2007;28(9):1030-5.

145. 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.

146. Sanchez-Pinto LN, Stroup EK, 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.

147. Vuokko R, Makela-Bengs P, Hypponen H, Lindqvist M, Doupi P. Impacts of structuring the electronic health record: Results of a systematic literature review from the perspective of secondary use of patient data. International journal of medical informatics. 2017;97:293-303.

148. Chaitram JM, Jevitt LA, Lary S, Tenover FC, Group WHOAR. The World Health Organization's External Quality Assurance System Proficiency Testing Program has improved the accuracy of antimicrobial susceptibility testing and reporting among participating laboratories using NCCLS methods. J Clin Microbiol. 2003;41(6):2372-7.

149. Wikipedia c. List of clinically important bacteria: Wikipedia, The Free Encyclopedia; updated 12 August 2021 05:32 UTC. [Available from:

https://en.wikipedia.org/w/index.php?title=List_of_clinically_important_bacteria&oldid=1038375238].

150. Moinat M, Schuemie M, Rijnbeek P. Usagi GitHub: Observational Health Data Sciences and Informatics (OHDSI); 2016 [Available from: https://github.com/OHDSI/Usagi#readme].

151. Jones M, DuVall SL, Spuhl J, Samore MH, Nielson C, Rubin M. Identification of methicillin-resistant Staphylococcus aureus within the nation's Veterans Affairs medical centers using natural language processing. BMC Med Inform Decis Mak. 2012;12:34.

152. Yim WW, Evans HL, Yetisgen M. Structuring Free-text Microbiology Culture Reports For Secondary Use. AMIA Jt Summits Transl Sci Proc. 2015;2015:471-5.

153. Matheny ME, Fitzhenry F, Speroff T, Hathaway J, Murff HJ, Brown SH, et al. Detection of blood culture bacterial contamination using natural language processing. AMIA Annual Symposium proceedings AMIA Symposium. 2009;2009:411-5.

154. Evans L, Rhodes A, Alhazzani W, Antonelli M, Coopersmith CM, French C, et al. Surviving Sepsis Campaign: International Guidelines for Management of Sepsis and Septic Shock 2021. Critical care medicine. 2021;49(11):e1063-e143.

155. Klann JG, Estiri H, Weber GM, Moal B, Avillach P, Hong C, et al. Validation of an internationally derived patient severity phenotype to support COVID-19 analytics from electronic health record data. Journal of the American Medical Informatics Association. 2021;28(7):1411-20.

156. Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Medical Research Methodology. 2014;14(1):40.

157. Ramspek CL, Jager KJ, Dekker FW, Zoccali C, van Diepen M. External validation of prognostic models: what, why, how, when and where? Clinical Kidney Journal. 2020;14(1):49-58.

158. Riley RD, Ensor J, Snell KIE, Debray TPA, Altman DG, Moons KGM, 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.

159. 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):279-89.

160. 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):623-4.

161. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, Bossuyt P, et al. Calibration: the Achilles heel of predictive analytics. BMC Medicine. 2019;17(1):230.

162. 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):971-80.

163. 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).

164. ATC classification index with DDDs. In: Methodology WCCfDS, editor. Oslo, Norway2019.

165. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in P ython. Journal of Machine Learning Research. 2011;12:2825--30.

166. Platt J. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. Adv Large Margin Classif. 2000;10.

167. Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. Proceedings of the 22nd international conference on Machine learning; Bonn, Germany: Association for Computing Machinery; 2005. p. 625–32.

168. 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 Medical Informatics and Decision Making. 2022;22(1):142.

169. 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):837-45.

170. Sun X, Xu W. Fast Implementation of DeLong's Algorithm for Comparing the Areas Under Correlated Receiver Operating Characteristic Curves. IEEE Signal Processing Letters. 2014;21(11):1389-93.

171. Huang Y, Li W, Macheret F, Gabriel RA, Ohno-Machado L. A tutorial on calibration measurements and calibration models for clinical prediction models. Journal of the American Medical Informatics Association. 2020;27(4):621-33.

172. 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. Journal of Clinical Epidemiology. 2016;74:167-76.

173. Debray TPA, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KGM. A new framework to enhance the interpretation of external validation studies of clinical prediction models. Journal of Clinical Epidemiology. 2015;68(3):279-89.

174. Ioannidis JPA. The Proposal to Lower P Value Thresholds to .005. JAMA. 2018;319(14):1429-30.

175. Nieboer D, van der Ploeg T, Steyerberg EW. Assessing Discriminative Performance at External Validation of Clinical Prediction Models. PLoS One. 2016;11(2):e0148820.

176. Khilnani GC, Zirpe K, Hadda V, Mehta Y, Madan K, Kulkarni A, et al. Guidelines for Antibiotic Prescription in Intensive Care Unit. Indian J Crit Care Med. 2019;23(Suppl 1):S1-s63.

177. Singh N, Yu VL. Rational empiric antibiotic prescription in the ICU. Chest. 2000;117(5):1496-9.

178. Paxton C, Niculescu-Mizil A, Saria S. Developing predictive models using electronic medical records: challenges and pitfalls. AMIA Annual Symposium proceedings AMIA Symposium. 2013;2013:1109-15.

179. Corey KM, Lorenzi E, Balu S, Sendak M, editors. Model Ensembling vs Data Pooling: Alternative ways to merge hospital information across sites. Machine Learning for Healthcare; 2019 10/18/22; University of Michigan.

180. Ashley EA, Dance DAB, Turner P. Grading antimicrobial susceptibility data quality: room for improvement. The Lancet Infectious Diseases. 2018;18(6):603-4.

181. Garnacho-Montero J, Garcia-Garmendia JL, Barrero-Almodovar A, Jimenez-Jimenez FJ, Perez-Paredes C, Ortiz-Leyba C. Impact of adequate empirical antibiotic therapy on the outcome of patients admitted to the intensive care unit with sepsis*. Critical care medicine. 2003;31(12):2742-51.

182. Ibrahim EH, Sherman G, Ward S, Fraser VJ, Kollef MH. The Influence of Inadequate Antimicrobial Treatment of Bloodstream Infections on Patient Outcomes in the ICU Setting. Chest. 2000;118(1):146-55.

183. Kollef MH, Morrow LE, Niederman MS, Leeper KV, Anzueto A, Benz-Scott L, et al. Clinical characteristics and treatment patterns among patients with ventilator-associated pneumonia. Chest. 2006;129(5):1210-8.

184. Garnacho-Montero J, Gutiérrez-Pizarraya A, Escoresca-Ortega A, Corcia-Palomo Y, Fernández-Delgado E, Herrera-Melero I, et al. De-escalation of empirical therapy is associated with lower mortality in patients with severe sepsis and septic shock. Intensive care medicine. 2014;40(1):32-40.

185. Rello J, Vidaur L, Sandiumenge A, Rodríguez A, Gualis B, Boque C, et al. De-escalation therapy in ventilator-associated pneumonia*. Critical care medicine. 2004;32(11):2183-90.

186. Schuts EC, Hulscher M, Mouton JW, Verduin CM, Stuart J, Overdiek H, et al. Current evidence on hospital antimicrobial stewardship objectives: a systematic review and meta-analysis. The Lancet Infectious diseases. 2016;16(7):847-56.

187. Guo Y, Gao W, Yang H, Ma C, Sui S. De-escalation of empiric antibiotics in patients with severe sepsis or septic shock: A meta-analysis. Heart Lung. 2016;45(5):454-9.

188. Lambregts MMC, Wijnakker R, Bernards AT, Visser LG, Cessie SL, Boer MGJ. Mortality after Delay of Adequate Empiric Antimicrobial Treatment of Bloodstream Infection. J Clin Med. 2020;9(5).

189. Ohnuma T, Chihara S, Costin B, Treggiari MM, Bartz RR, Raghunathan K, et al. Association of Appropriate Empirical Antimicrobial Therapy With In-Hospital Mortality in Patients With Bloodstream Infections in the US. JAMA Network Open. 2023;6(1):e2249353-e.

190. Retamar P, Portillo MM, López-Prieto MD, Rodríguez-López F, de Cueto M, García MV, et al. Impact of inadequate empirical therapy on the mortality of patients with bloodstream infections: a propensity score-based analysis. Antimicrob Agents Chemother. 2012;56(1):472-8.

191. González AL, Leal AL, Cortés JA, Sánchez R, Barrero LI, Castillo JS, et al. Effect of adequate initial antimicrobial therapy on mortality in critical patients with Pseudomonas aeruginosa bacteremia. Biomedica. 2014;34 Suppl 1:58-66.

192. Yoon YK, Park DW, Sohn JW, Kim HY, Kim Y-S, Lee C-S, et al. Effects of inappropriate empirical antibiotic therapy on mortality in patients with healthcare-associated methicillin-resistant Staphylococcus aureus bacteremia: a propensity-matched analysis. BMC Infectious Diseases. 2016;16(1):331.

193. Rodríguez-Baño J, Millán AB, Domínguez MA, Borraz C, González MP, Almirante B, et al. Impact of inappropriate empirical therapy for sepsis due to health care-associated methicillin-resistant Staphylococcus aureus. Journal of Infection. 2009;58(2):131-7.

194. Mortensen EM, Restrepo MI, Anzueto A, Pugh JA. Antibiotic Therapy and 48-Hour Mortality for Patients with Pneumonia. The American Journal of Medicine. 2006;119(10):859-64.

195. Sibbald B, Roland M. Understanding controlled trials: Why are randomised controlled trials important? BMJ. 1998;316(7126):201.

196. Eikenboom AM, Le Cessie S, Waernbaum I, Groenwold RHH, de Boer MGJ. Quality of Conduct and Reporting of Propensity Score Methods in Studies Investigating the Effectiveness of Antimicrobial Therapy. Open Forum Infectious Diseases. 2022;9(4).

197. Sibbald B, Roland M. Understanding controlled trials. Why are randomised controlled trials important? BMJ: British Medical Journal. 1998;316(7126):201.

198. Benedetto U, Head SJ, Angelini GD, Blackstone EH. Statistical primer: propensity score matching and its alternatives. Eur J Cardiothorac Surg. 2018;53(6):1112-7.

199. Austin PC, Xin Yu AY, Vyas MV, Kapral MK. Applying Propensity Score Methods in Clinical Research in Neurology. Neurology. 2021;97(18):856-63.

200. Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate Behav Res. 2011;46(3):399-424.

201. McMurry TL, Hu Y, Blackstone EH, Kozower BD. Propensity scores: Methods, considerations, and applications in the Journal of Thoracic and Cardiovascular Surgery. J Thorac Cardiovasc Surg. 2015;150(1):14-9.

202. Rosenbaum PR, Rubin DB. The Central Role of the Propensity Score in Observational Studies for Causal Effects. Biometrika. 1983;70(1):41-55.

203. Rosenbaum PR. Propensity Score. Encyclopedia of Biostatistics2005.

204. Yao XI, Wang X, Speicher PJ, Hwang ES, Cheng P, Harpole DH, et al. Reporting and Guidelines in Propensity Score Analysis: A Systematic Review of Cancer and Cancer Surgical Studies. J Natl Cancer Inst. 2017;109(8).

205. Kline A, Luo Y. PsmPy: A Package for Retrospective Cohort Matching in Python. Annu Int Conf IEEE Eng Med Biol Soc. 2022;2022:1354-7.

206. Austin PC. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. Pharm Stat. 2011;10(2):150-61.

207. Staffa SJ, Zurakowski D. Five Steps to Successfully Implement and Evaluate Propensity Score Matching in Clinical Research Studies. Anesthesia & Analgesia. 2018;127(4):1066-73.

208. Austin PC, Grootendorst P, Anderson GM. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. Stat Med. 2007;26(4):734-53.

209. Mayr VD, Dünser MW, Greil V, Jochberger S, Luckner G, Ulmer H, et al. Causes of death and determinants of outcome in critically ill patients. Crit Care. 2006;10(6):R154.

210. Vincent JL, Marshall JC, Namendys-Silva SA, Francois B, Martin-Loeches I, Lipman J, et al. Assessment of the worldwide burden of critical illness: the intensive care over nations (ICON) audit. Lancet Respiratory Medicine. 2014;2(2213-2619 (Electronic)).

211. Demirdal T, Sen P, Nemli A, Kizilkaya M. The Diagnostic Value of Scoring Systems in Predicting Bacteremia and the Mortality Rate of Patients With the Signs of Infection in Intensive Care Units. Open Forum Infectious Diseases. 2016;3(suppl_1).

212. Kaji DA, Zech JR, Kim JS, Cho SK, Dangayach NS, Costa AB, et al. An attention based deep learning model of clinical events in the intensive care unit. PLoS ONE. 2019;14(2).

213. Choi MH, Kim D, Choi EJ, Jung YJ, Choi YJ, Cho JH, et al. Mortality prediction of patients in intensive care units using machine learning algorithms based on electronic health records. Scientific Reports. 2022;12(1):7180.

214. Vincent JL, Moreno R. Clinical review: scoring systems in the critically ill. Crit Care. 2010;14(2):207.

215. Inverse solutions of the Severinghaus and Thomas equations which allow Po2 to be derived directly from So22011.

216. Imai K, King G, Stuart EA. Misunderstandings between experimentalists and observationalists about causal inference. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2008;171(2):481-502.

217. Austin PC. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. Stat Med. 2008;27(12):2037-49.

218. Cohen J. A power primer. Psychol Bull. 1992;112(1):155-9.

219. Ali MS, Groenwold RH, Belitser SV, Pestman WR, Hoes AW, Roes KC, et al. Reporting of covariate selection and balance assessment in propensity score analysis is suboptimal: a systematic review. Journal of clinical epidemiology. 2015;68(2):122-31.

220. Ali MS, Groenwold RH, Klungel OH. Best (but oft-forgotten) practices: propensity score methods in clinical nutrition research. The American Journal of Clinical Nutrition. 2016;104(2):247-58.

221. Austin PC. Comparing paired vs non-paired statistical methods of analyses when making inferences about absolute risk reductions in propensity-score matched samples. Stat Med. 2011;30(11):1292-301.

222. Fay MP, Lumbard K. Confidence intervals for difference in proportions for matched pairs compatible with exact McNemar's or sign tests. Statistics in Medicine. 2021;40(5):1147-59.

223. Yasunaga H. Introduction to Applied Statistics—Chapter 1 Propensity Score Analysis. Annals of Clinical Epidemiology. 2020;2(2):33-7.

224. Chen JW, Maldonado DR, Kowalski BL, Miecznikowski KB, Kyin C, Gornbein JA, et al. Best Practice Guidelines for Propensity Score Methods in Medical Research: Consideration on Theory, Implementation, and Reporting. A Review. Arthroscopy. 2022;38(2):632-42.

225. Rosenbaum PR. Overt Bias in Observational Studies. In: Rosenbaum PR, editor. Observational Studies. New York, NY: Springer New York; 2002. p. 71-104.

226. King BM, Rosopa PJ, Minium EW. Statistical reasoning in the behavioral sciences: John Wiley & Sons; 2018.

227. Mangiafico S. rcompanion: Functions to Support Extension Education Program Evaluation. 2.4.21 ed2023.

228. Mathur MB, Ding P, Riddell CA, VanderWeele TJ. Web Site and R Package for Computing E-values. Epidemiology. 2018;29(5):e45-e7.

229. VanderWeele TJ, Ding P. Sensitivity Analysis in Observational Research: Introducing the E-Value. Annals of internal medicine. 2017;167(4):268-74.

230. Rudolph KE, Stuart EA. Using Sensitivity Analyses for Unobserved Confounding to Address Covariate Measurement Error in Propensity Score Methods. American Journal of Epidemiology. 2017;187(3):604-13.

231. Liu W, Kuramoto SJ, Stuart EA. An introduction to sensitivity analysis for unobserved confounding in nonexperimental prevention research. Prev Sci. 2013;14(6):570-80.

232. Li L, Shen C, Wu AC, Li X. Propensity score-based sensitivity analysis method for uncontrolled confounding. Am J Epidemiol. 2011;174(3):345-53.

233. Zhao Q-Y, Luo J-C, Su Y, Zhang Y-J, Tu G-W, Luo Z. Propensity score matching with R: conventional methods and new features. Annals of Translational Medicine. 2021;9(9):812.

234. Ho D, Imai K, King G, Stuart E. Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. Political Analysis. 2007;15:199–236.

235. Zhang Z, Kim HJ, Lonjon G, Zhu Y. Balance diagnostics after propensity score matching. Ann Transl Med. 2019;7(1):16.

236. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Stat Med. 2009;28(25):3083-107.

237. Emsley R, Lunt M, Pickles A, Dunn G. Implementing double-robust estimators of causal effects. The Stata Journal. 2008;8(3):334-53.

238. Leone M, Bechis C, Baumstarck K, Lefrant JY, Albanèse J, Jaber S, et al. De-escalation versus continuation of empirical antimicrobial treatment in severe sepsis: a multicenter non-blinded randomized noninferiority trial. Intensive care medicine. 2014;40(10):1399-408.

239. Takahashi N, Imaeda T, Nakada Ta, Oami T, Abe T, Yamao Y, et al. Short- versus long-course antibiotic therapy for sepsis: a post hoc analysis of the nationwide cohort study. Journal of Intensive Care. 2022;10(1):49.

240. Marquet K, Liesenborgs A, Bergs J, Vleugels A, Claes N. Incidence and outcome of inappropriate in-hospital empiric antibiotics for severe infection: a systematic review and meta-analysis. Crit Care. 2015;19(1):63.

241. Fraser A, Paul M, Almanasreh N, Tacconelli E, Frank U, Cauda R, et al. Benefit of Appropriate Empirical Antibiotic Treatment: Thirty-day Mortality and Duration of Hospital Stay. The American Journal of Medicine. 2006;119(11):970-6.

242. Kappen TH, van Klei WA, van Wolfswinkel L, Kalkman CJ, Vergouwe Y, Moons KGM. Evaluating the impact of prediction models: lessons learned, challenges, and recommendations. Diagnostic and Prognostic Research. 2018;2(1):11.

# 8. APPENDICES

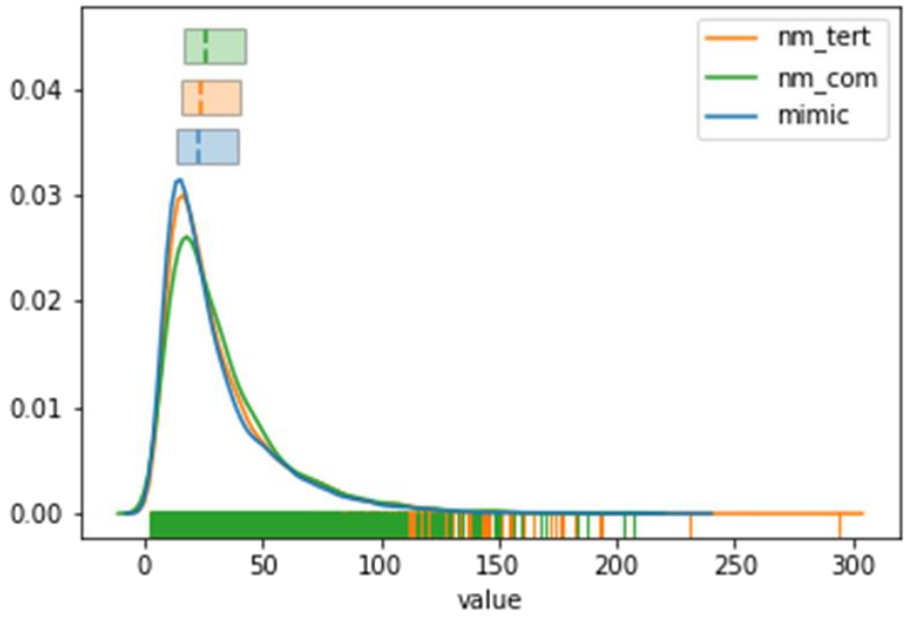


*Figure S 1. Kernel Density Estimation of Blood Urea Nitrogen (mg/dL) across MIMIC, NM-T and NM-C cohorts.*

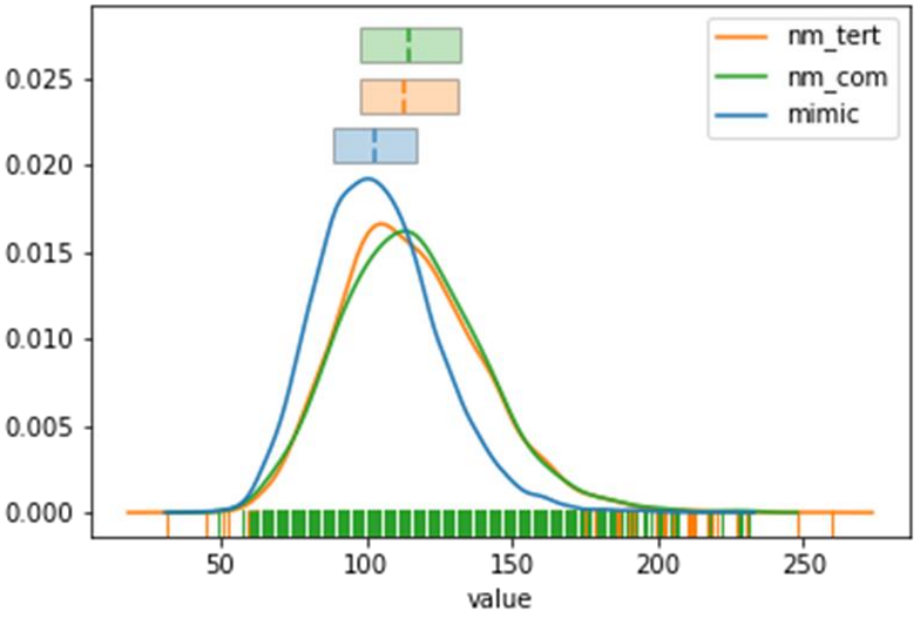


*Figure S 2. Kernel Density Estimation of Heart rate (beats/min) across MIMIC, NM-T and NM-C cohorts.*

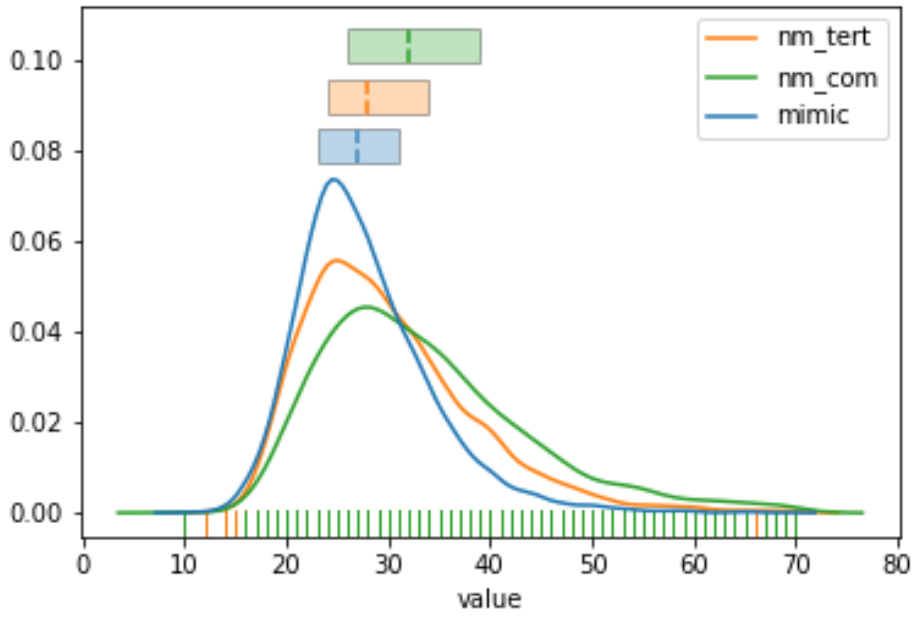


*Figure S 3. Kernel Density Estimation of Respiration rate (breath/min) across MIMIC, NM-T and NM-C cohorts.*

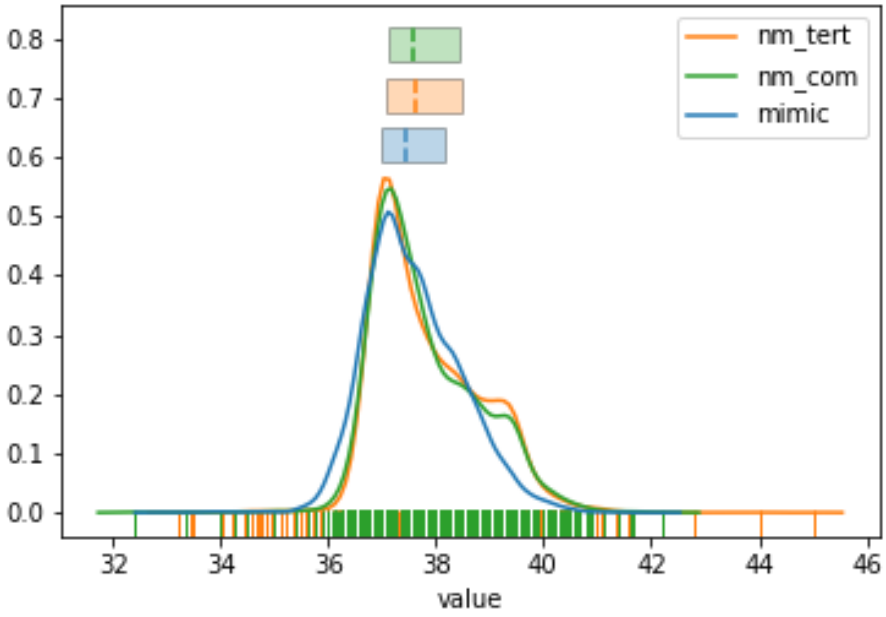


*Figure S 4. Kernel Density Estimation of Temperature (Celsius) across MIMIC, NM-T and NM-C cohorts.*

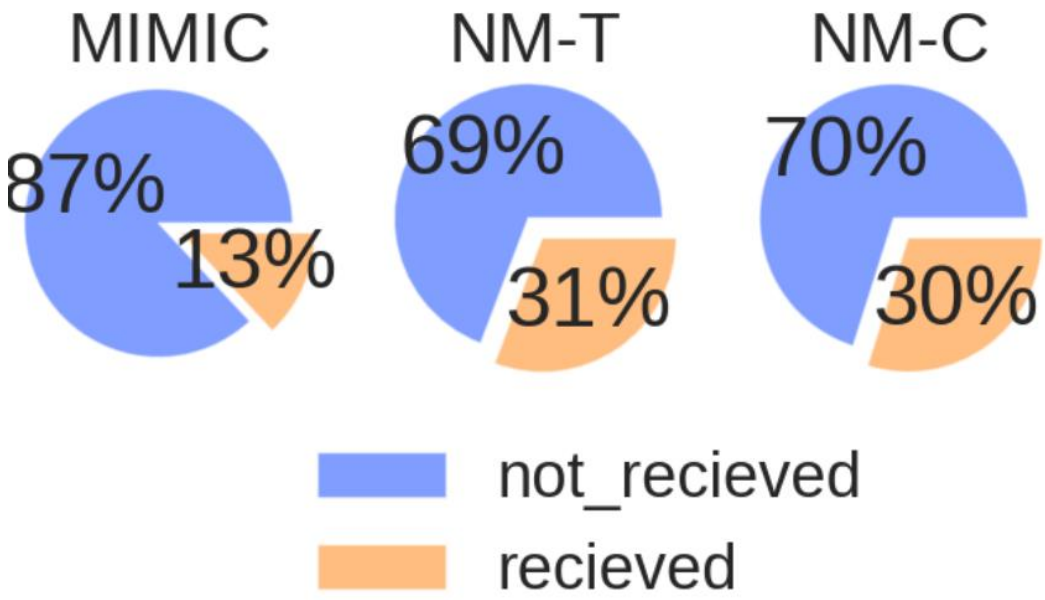


*Figure S 5. Distribution of binary indication for a patient having received norepinephrine across MIMIC, NM-T and NM-C cohorts.*

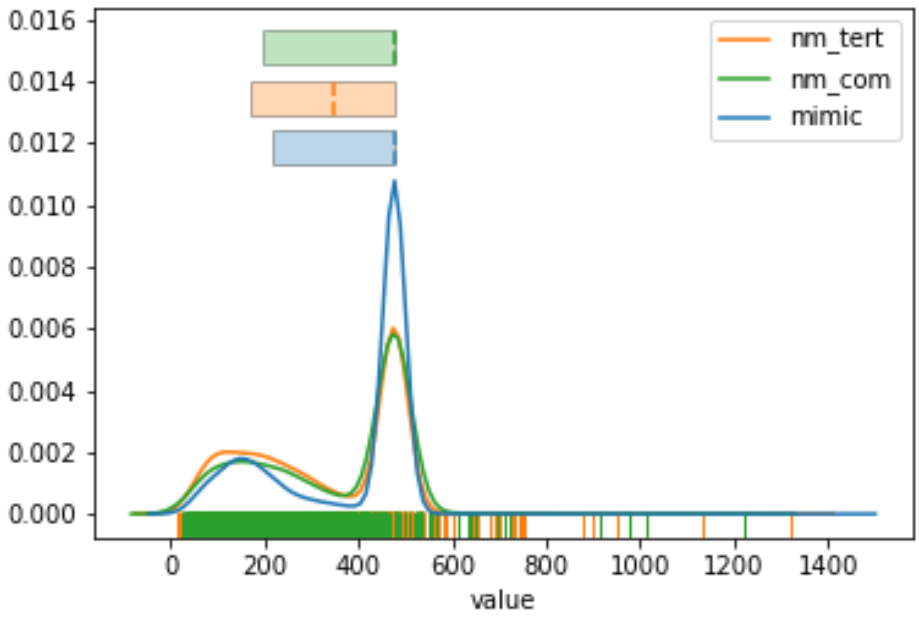


*Figure S 6. Kernel Density Estimation of Partial pressure of arterial oxygen (PaO2)/ Fraction of inspired oxygen (FiO2) ratio (PaO2:FiO2) across MIMIC, NM-T and NM-C cohorts.*

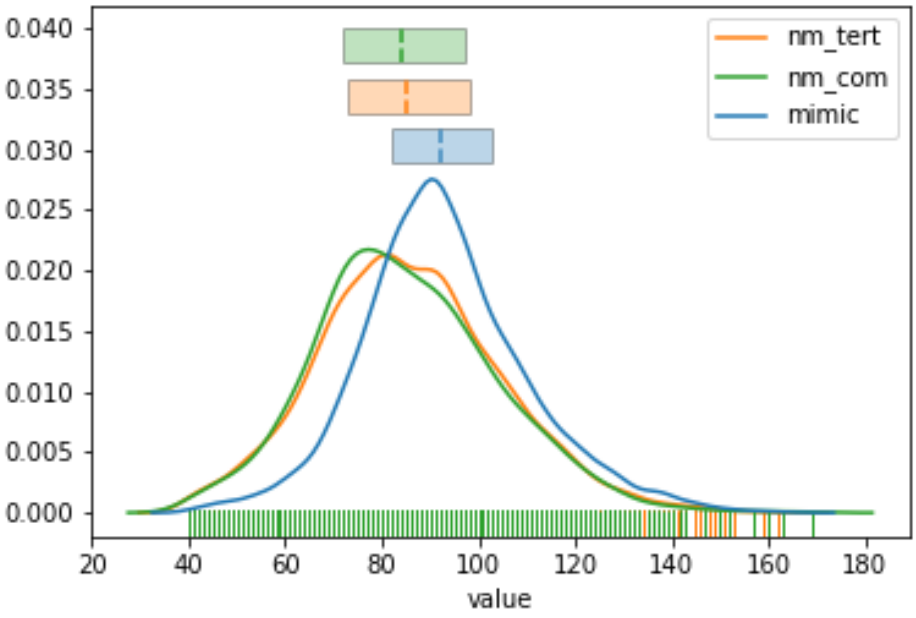


*Figure S 7. Kernel Density Estimation of Systolic Blood Pressure (mmHg) across MIMIC, NM-T and NM-C cohorts.*

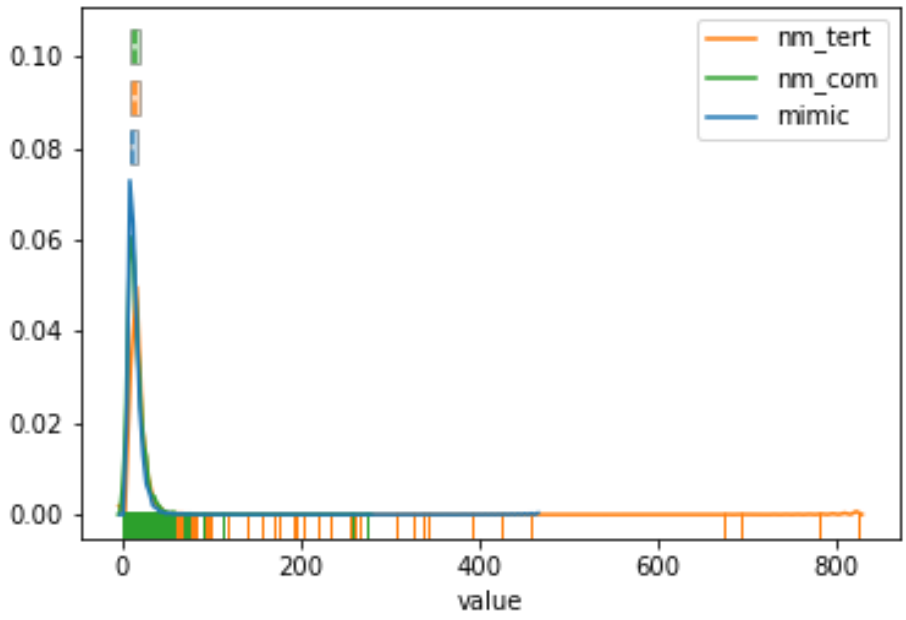


*Figure S 8. Kernel Density Estimation of White Blood Cell Count (K/uL) across MIMIC, NM-T and NM-C cohorts.*

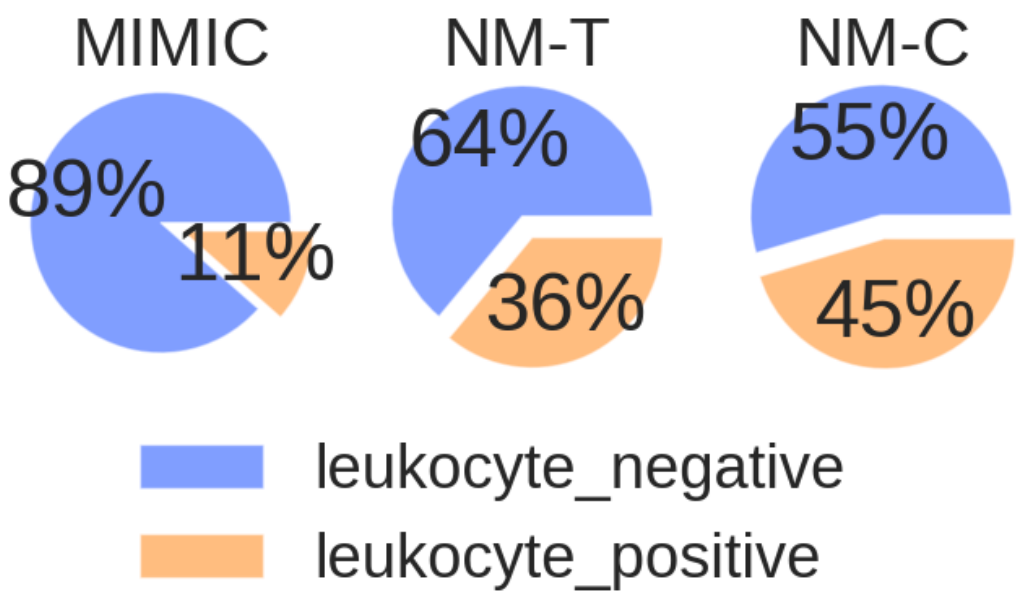


*Figure S 9. Distribution of binary indication of leukocytes in urine across MIMIC, NM-T and NM-C cohorts.*

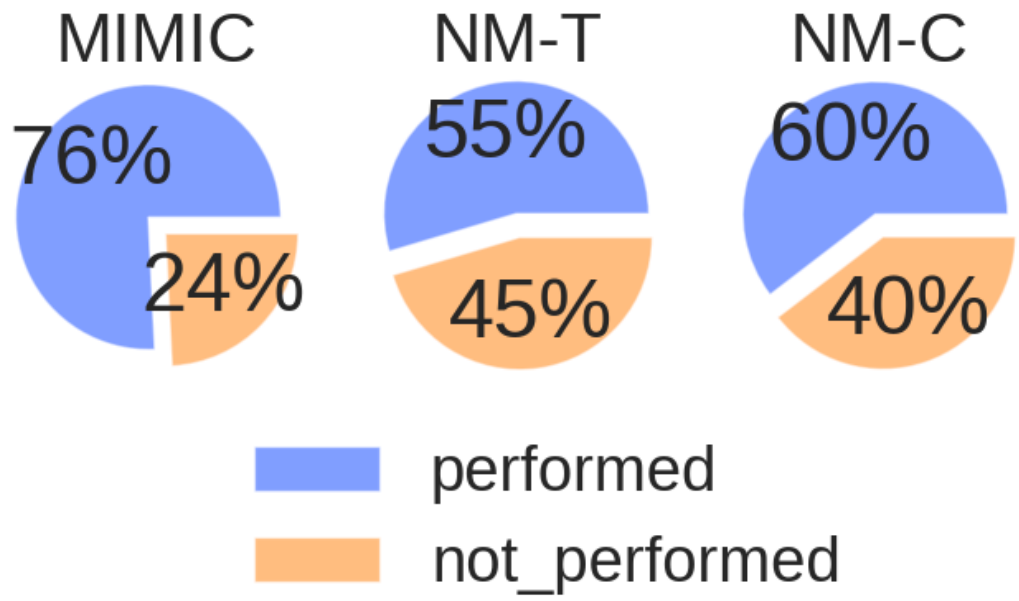


*Figure S 10. Distribution of binary indication for having a blood culture performed across MIMIC, NM-T and NM-C cohorts.*

Patient case review for top N=5 miscalibrated $MIMIC_M$ False Negatives on NM-$C_{val}$.

*Annotation S 1. $Ensemble_M$ False Negatives on NM-$C_{val}$ #1*

**ICUSTAY_ID**= 235672407263389594291076384005368862603

**$Ensemble_M$ Predicted Probability: 0.232 (-0.0.036)**

**$MIMIC_M$ Predicted Probability: 0.317 (+0.043)**

**$Tertiary_M$ Predicted Probability: 0.145 (-0.121)**

**Clinical assessment**: 55 y.o. female, Primary diagnosis opioid withdrawal, secondary diagnosis of E. Coli UTI without signs of serious infection. No signs of serious bacterial infection. Low BI risk appears to be most accurate classification.

*Annotation S 2. EnsembleM False Negatives on NM-Cval #2*

**ICUSTAY_ID**= 179403945252634728011685320581670748825

**$Ensemble_M$ Predicted Probability: 0.195 (-0.072)**

**$MIMIC_M$ Predicted Probability: 0.284 (+0.010)**

**$Tertiary_M$ Predicted Probability: 0.105 (-0.162)**

**Clinical assessment**: 84 y.o. female currently receiving oncology treatment, primary diagnosis of viral pneumonia with possible secondary bacterial pneumonia with neutropenia. Appears to have received antibacterial treatments prior to ICU admission. Unclear BI risk.

*Annotation S 3. EnsembleM False Negatives on NM-Cval #3*

**ICUSTAY_ID**= 57473130153343061619339355741308641456

**Ensemble$_M$ Predicted Probability: 0.185 (-0.082)**

**MIMIC$_M$ Predicted Probability: 0.275 (+0.001)**

**Tertiary$_M$ Predicted Probability: 0.096 (-0.172)**

**Clinical assessment**: 64 y.o. female presented to ICU with DKA and hyperglycemia, found to have a UTI as a secondary diagnosis. No signs of serious bacterial infection. Low BI risk appears to be most accurate classification.

*Annotation S 4. EnsembleM False Negatives on NM-Cval #4*

**ICUSTAY_ID**= 33275200732028899365779470215493954091 9

**Ensemble$_M$ Predicted Probability: 0.204 (-0.063)**

**MIMIC$_M$ Predicted Probability: 0.302 (+0.029)**

**Tertiary$_M$ Predicted Probability: 0.106 (-0.161)**

**Clinical assessment**: 84 y.o. male with h/o coronary artery disease, heart failure and s/p aortic valve replacement admitted to ICU with multiple medical problems and was transferred because of weakness and low BPs. Patient developed strep salivarius bacteremia. This case is challenging to label BI status and classify due to complicated timings of antibiotics/culture and laboratory tests. High BI risk appears to be most accurate classification for this case.

*Annotation S 5. EnsembleM False Negatives on NM-Cval #5*

**ICUSTAY_ID**= 164657114322571042035219888797857470968

**Ensemble$_M$ Predicted Probability: 0.214 (-0.054)**

**MIMIC$_M$ Predicted Probability: 0.288 (+0.014)**

**Tertiary$_M$ Predicted Probability: 0.139 (-0.128)**

**Clinical assessment**: 60 y.o. male with history of COPD and coronary artery disease presenting with acute on chronic hypoxemia of unclear etiology. Found to have coagulase negative staph bacteremia. Patient transferred from an outside ED where they were started on antibiotics (though not detailed in NMH prescription information). High BI risk appears to be most accurate classification for this case.

*Annotation S 6. MIMIC$_M$ False Negatives on NM-C$_{val}$ #1*

**ICUSTAY_ID**= 232950261229633663346209012108970814851

**Ensemble$_M$ Predicted Probability: 0.279 (+0.011)**

**MIMIC$_M$ Predicted Probability: 0.217 (-0.057)**

**Tertiary$_M$ Predicted Probability: 0.339 (+0.073)**

**Clinical assessment**: 86 y.o. male with indwelling urinary catheter diagnosed with a possible UTI and axillary cellulitis. No signs of serious bacterial infection. Low BI risk appears to be most accurate classification.

*Annotation S 7. MIMICM False Negatives on NM-Cval #2*

**ICUSTAY_ID**= 26573127152078424055522007612823850500O

**Ensemble$_M$ Predicted Probability: 0.254 (+0.013)**

**MIMIC$_M$ Predicted Probability: 0.209 (-0.065)**

**Tertiary$_M$ Predicted Probability: 0.299 (+0.031)**

**Clinical assessment**: 66 y.o. female with chronic respiratory failure and heart failure admitted for dyspnea, found to have a E. Coli + urine culture. Low BI risk appears to be most accurate classification.

*Annotation S 8. MIMICM False Negatives on NM-Cval #3*

**ICUSTAY_ID**= 331837907160240515755118455837560318O3

**Ensemble$_M$ Predicted Probability: 0.279 (+0.012)**

**MIMIC$_M$ Predicted Probability: 0.239 (-0.035)**

**Tertiary$_M$ Predicted Probability: 0.318 (+0.051)**

**Clinical assessment**: 77 y.o. female patient admitted for GI bleed/ hernia repair complications. Found to have a culture-positive UTI. No signs of serious bacterial infection. Low BI risk appears to be most accurate classification.

*Annotation S 9. MIMICM False Negatives on NM-Cval #4*

**ICUSTAY_ID**= 116605758089851103555130508100449988636

**Ensemble$_M$ Predicted Probability: 0.282 (+0.015)**

**MIMIC$_M$ Predicted Probability: 0.230 (-0.045)**

**Tertiary$_M$ Predicted Probability: 0.335 (+0.068)**

**Clinical assessment**: 70 y.o female with complex medical history, admitted for toxic metabolic encephalopathy found to have a UTI. Low BI risk appears to be most accurate classification.

*Annotation S 10. MIMICM False Negatives on NM-Cval #5*

**ICUSTAY_ID**= 211923585744204536793083163485841394844

**Ensemble$_M$ Predicted Probability: 0.301 (+0.035)**

**MIMIC$_M$ Predicted Probability: 0.237 (-0.037)**

**Tertiary$_M$ Predicted Probability: 0.366 (+0.099)**

**Clinical assessment**: 89 y.o. male with complex medical history and traumatic subdural hematoma with urinary obstruction secondary to indwelling catheter possibly in the setting of a UTI. No signs of serious bacterial infection. Low BI risk appears to be most accurate classification.

# 9. CURRICULUM VITAE


## Garrett Eickelberg

Ph.D. Candidate, Biomedical Informatics
Driskill Graduate Program
Northwestern University Feinberg School of Medicine


---

## Education

2008 – 2012 **Bachelor of Arts (Chemistry), Willamette University**

2015 – 2017 **Post-Bachelor degree courses, Portland State University & Portland Community College**

2017 – 2023 **PhD in Biomedical Informatics, Northwestern University**

Developed, optimized, and evaluated open-source modeling workflow to predict patient-level risk of bacterial infection and guide antibiotic therapy decisions for critically ill adults in the MIMIC-III database.

Created and externally validated an open-source python package to extract bacterial infection status concepts from free-text microbiology reports.

Extracted and harmonized data from Northwestern's electronic data warehouse. Data were used to externally validate previously developed bacterial infection model performance and transportability across multiple unique critical care environments.

Measured the association between outcomes and prolonged antibiotic exposure in critically ill adults with a low predicted risk of bacterial infection using a propensity score matching approach.

## Training/Internship

2009 & 2010 **Student Summer Researcher, Oregon Health & Science University**

2018 & 2019 **Biomedical Data-Driven Discovery (BD3) Training, Northwestern University** Advanced data science coursework offered by the Masters of Science in Analytics (MSIA) program in the McCormick School of Engineering and Applied Science

2021 **Management for Scientists and Engineers Certification, Northwestern University** Eight-week training program in business and leadership skills offered by the Kellogg School of Business.

## Employment

2012 – 2017 **Research Assistant II, Knight Diagnostic Lab**

Responsible for the conceptualization, design, and completion of research projects

Organized and optimized data extraction and cleaning from clinical and laboratory databases

Assembled multiple datasets consisting of cross-referenced clinical outcomes data and laboratory data

Site manager for FDA clinical trials involving: QuantideX® BCR-ABL assay (Asuragen), JAK2 Muta*Quant*® Kit (Ipsogen), Xpert® BCR-ABL Monitor (Cephid)

## Awards

2010 – 2011 Nancy K. Detering Waechter Scholarship Award

2019 Best Poster (Day 1): 2019 NLM Informatics Training Conference

## Extramural Service

2012 – 2013 **Tattoo Removal Volunteer, Outside-In [100 hours]**

2012 – 2015 **Needle Exchange Volunteer, Outside-In [700 hours]**

2015 – 2016 **Internal Medicine Volunteer, Oregon Health & Science University [150 hours]**

2019 – 2023 **Student Assisted Mentoring Program (StAMP) [100 hours]**

2020 **SARS-CoV2 Simulation Researcher, Northwestern University/Illinois Department of Public Health [100 hours]**

2020 **2020 National Library of Medicine Training Grant Conference Student Planning Committee [10 hours]**

## Teaching

2009 – 2012 **Chemistry Tutor, Willamette University**

2019 – 2020 **Teaching Assistant, Northwestern University**

2019 – 2023 **Guest Lecturer, Northwestern University**

## Publications

**Eickelberg G**, Fisher AJ. Environmental regulation of plant gene expression: An RT-qPCR laboratory project for an upper-level undergraduate biochemistry or molecular biology course. *Biochemistry and Molecular Biology Education* **2013.**

Press RD, **Eickelberg G**, McDonald TJ, et al. Highly accurate molecular genetic testing for HFE hereditary hemochromatosis: results from 10 years of blinded proficiency surveys by the College of American Pathologists. *Genet Med* **2016**.

Ma L, Boucher JI, Paulsen J, Eide C, **Eickelberg G**, et al. CRISPR-Cas9-mediated saturated mutagenesis screen predicts clinical drug resistance with improved accuracy. *Proc Natl Acad Sci U S A*. **2017.**

Press RD, **Eickelberg G**, Froman A, et al. Next-generation sequencing-defined minimal residual disease before stem cell transplantation predicts acute myeloid leukemia relapse. *Am J Hematol*. **2019**.

**Eickelberg G**, Sanchez-Pinto N, Luo L. Predictive Modeling of Bacterial Infections and Antibiotic Therapy Needs in Critically Ill Adults. *Journal of Biomedical Informatics* **2020**

Christopher A. Eide, Brian J. Druker, [et al, including **Eickelberg G].** Characterization of the Genomic Landscape of BCR-ABL1 Kinase-Independent Mechanisms of Resistance to ABL1 Tyrosine Kinase Inhibitors in Chronic Myeloid Leukemia. *Blood* 2021

**Eickelberg G**, Sanchez-Pinto N, Luo L. Development and Validation of MicrobEx: an open-source package for microbiology culture concept extraction. **2022** *JAMIA Open, Volume 5, Issue 2*

Rosenberg MW, Savage SL, Eide CA, Reister Schultz A, Cook RJ, Press RD, Rempfer C, **Eickelberg G**, Wilmot B, McWeeney SK, Tyner JW, Druker BJ, Tognon CE. Comprehensive molecular characterization of a rare case of Philadelphia chromosome-positive acute myeloid leukemia. *Cold Spring Harb Mol Case Stud*. 2022

Pacheco J, Rasmussen L, [et al, including **Eickelberg G**]. Evaluation of the Portability of Computable Phenotypes with Natural Language Processing in the eMERGE Network. *Sci Rep* **13**, 1971 (2023). https://doi.org/10.1038/s41598-023-27481-y

**Eickelberg G**, Sanchez-Pinto N, Kline A, Luo L. Transportability of Bacterial Infection Prediction Models for Critically Ill Patients. **[Submitted for review *NPJ Digital Medicine* on 02/2023]**

**Eickelberg G**, Luo L, Kline A, Sanchez-Pinto N. Empiric Antimicrobial Treatment Duration and Mortality in Critically Ill Patients with Low Predicted Risk of Bacterial Infection – A Propensity Score matched analysis. **[Preparing to submit 04/2023]**

## Talks/Posters

"Ethylene Promotes the Floral Transition in *Arabidopsis thaliana* Through the Attenuation of *FLOWERING LOCUS C.*" **Eickelberg G**, Fisher A. Talk presented at Murdoc College Science Research Conference in Seattle, WA. November 15, 2012.

"BCR-ABL1 molecular responses at 12-18 months predict long-term event-free survival in patients with tyrosine kinase inhibitor (TKI)-treated chronic myelogenous leukemia (CML)." **Eickelberg G,** Watt C, Press R et al. Poster defended at Association of Molecular Pathology conference in Charlotte, NC. November 11, 2016.

"Novel MIMICiii Data Pipeline for Patients with Serious Bacterial Infections" **Eickelberg G**, Talk presented at 2019 NLM Informatics Training Conference in Indianapolis, IN. June 24-25, 2019. *(Best Poster Day 1)

"From MIMIC to Model: Building a workflow to model raw retrospective clinical data." **Eickelberg G**, Workshop presented at 2020 Biomedical Data Science Day in Chicago, IL. February 4, 2020.

"Predictive Modeling of Bacterial Infections and Antibiotic Therapy Needs in Critically Ill Adults" **Eickelberg G**, Plenary to be presented at 2020 NLM Informatics Training Conference in Portland, OR. June 23-24, 2020.

## Grant Awards

2018 – 2019 BD2K Predoctoral T32 Training Grant [5T32LM012203]

## Languages

Advanced: Python, R, SQL (sql server, postgres, and mysql)

Intermediate: Bash, Java

## Skills

Predictive Analytics, Clinical Model Validation, Data Mining, Data Harmonization, Deep Learning, Data Visualization, Natural Language Processing, Microsoft Office (Word, Excel, PowerPoint, Outlook), Redcap, Biomedical ontologies, >10 years healthcare data experience

## References

Yuan Luo, Northwestern University [Dissertation Advisor]
yuan.luo@northwestern.edu
312-503-5742

L. Nelson Sanchez-Pinto, Northwestern University/Lurie Children's Hospital
lsanchezpinto@luriechildrens.org

Richard Press, Oregon Health & Science University
pressr@ohsu.edu
503-494-2317

Alison Fisher, Willamette University
ajfisher@willamette.edu
503-370-6793