

It is made available under a [CC-BY-NC-ND 4.0 International license](#) .

A Data-Driven Framework for Identifying Intensive Care Unit Admissions Colonized with Multidrug-Resistant Organisms

Çağlar Çağlayan, Ph.D.¹,
Sean Barnes, Ph.D.²,
Lisa L. Pineles, M.A.³,
Eili Y. Klein, Ph.D.⁴,
Anthony D. Harris, M.D., M.P.H.³

¹ Johns Hopkins University Applied Physics Laboratory, Laurel, MD

² R.H. Smith School of Business, University of Maryland, College Park, MD

³ Department of Epidemiology and Public Health, University of Maryland School of Medicine, Baltimore, MD

⁴ Department of Emergency Medicine, Johns Hopkins University School of Medicine, Baltimore, MD

Corresponding author: Çağlar Çağlayan

Next Generation Care Delivery Group

Johns Hopkins University Applied Physics Laboratory (JHU/APL)

11100 Johns Hopkins Road, Laurel, MD 20723

Phone: (240) 709-5184 E-mail: caglar.caglayan@jhuapl.edu

Acknowledgement: This study is supported by the U.S. Centers for Disease Control and Prevention (CDC) Modeling Infectious Disease (MInD) Network prime award number 1U01CK000536. The content is solely the responsibility of the authors and does not necessarily represent the official views of the CDC.

Word Count: Abstract 195 words; Main Text 5,587 words (including section titles)

Tables and Figures: 2 Figure and 3 Tables

Abstract: The rising prevalence of multi-drug resistant organisms (MDROs), such as Methicillin-resistant *Staphylococcus aureus* (MRSA), Vancomycin-resistant *Enterococci* (VRE), and Carbapenem-resistant *Enterobacteriaceae* (CRE), is an increasing concern in healthcare settings. Leveraging electronic healthcare record data, we developed a data-driven framework to predict MRSA, VRE, and CRE colonization upon intensive care unit admission (ICU), and identify the associated socio-demographic and clinical factors using logistic regression (LR), random forest (RF), and XGBoost algorithms. We performed threshold optimization for converting predicted probabilities into binary predictions and identified the cut-off maximizing the sum of sensitivity and specificity. We achieved the following sensitivity and specificity values with the best performing models: 80% and 66% for VRE with LR, 73% and 77% for CRE with XGBoost, 76% and 59% for MRSA with RF, and 82% and 83% for MDRO (i.e., VRE or CRE or MRSA) with RF. Further, we identified several predictors of MDRO colonization, including long-term care facility exposure, current diagnosis of skin/subcutaneous tissue or infectious/parasitic disease, and recent isolation precaution procedures before ICU admission. Our data-driven modeling framework can be used as a clinical decision support tool for timely predictions, identification of high-risk patients, and selective and timely use of infection control measures in ICUs.

Key words: Multidrug-Resistant Organisms (MDROs); Carbapenem-Resistant *Enterobacteriaceae* (CRE); Vancomycin-Resistant *Enterococci* (VRE); Methicillin-Resistant *Staphylococcus Aureus* (MRSA); Healthcare-Associated Infections (HAIs); Machine Learning (ML); Data-Centric Analytics; Predictive Analytics

It is made available under a [CC-BY-NC-ND 4.0 International license](#) .

A Data-Driven Framework for Identifying Intensive Care Unit Admissions Colonized with Multidrug-Resistant Organisms

1. Introduction

The increasing prevalence of multidrug resistant organisms (MDROs), bacteria that are resistant to one or more classes of antibiotics, is an increasingly concerning issue in the community, and in particular, healthcare settings where admitted patients are especially susceptible to developing an infection ([Centers for Disease Control and Prevention \(CDC\) 2013](#)) ([Spellberg et al. 2008](#)) ([Boucher et al. 2009](#)). These organisms pose a significant threat to patient safety in the form of healthcare-associated infections ([Haque et al. 2018](#)), which are associated with considerable morbidity, mortality, and healthcare costs ([Klevens, Edwards, and Richards 2007](#)), and have the potential to spread within the community ([Molton et al. 2013](#)) ([Bassetti, Nicco, and Mikulska 2009](#)).

Two MDROs that are the most prevalent causes of HAIs are Methicillin—resistant *Staphylococcus aureus* (MRSA) and vancomycin—resistant *Enterococcus* (VRE) ([Calfee 2012](#)) ([Harris et al. 2013](#)), which are currently classified as serious threats by the U.S. Centers for Disease Control and Prevention (CDC) ([Centers for Disease Control and Prevention \(CDC\) 2018](#)). MRSA is reported to cause 80,461 infections and 11,285 deaths per year, and VRE causes 20,000 infections and 11,300 deaths per year ([Centers for Disease Control and Prevention \(CDC\) 2013](#)), with both MDROs being associated with poor treatment outcomes following infections ([Cosgrove et al. 2003](#)) ([DiazGranados et al. 2005](#)), longer length of hospitalization, and higher healthcare costs ([Song et al. 2003](#)) ([Cosgrove et al. 2005](#)) ([Maragakis, Perencevich, and Cosgrove 2008](#)).

In recent years, Carbapenem-resistant *Enterobacteriaceae* (CRE)—an MDRO class that is highly resistant to carbapenems and other antibiotics reserved for treatment of severe infections—have reached concerning levels in healthcare facilities in the U.S. ([Centers for Disease Control and Prevention \(CDC\) 2015](#)), and around the world ([World Health Organization 2017](#)). This trend has prompted the CDC to classify CRE as an urgent threat to public health, its highest risk category ([Centers for Disease Control and Prevention \(CDC\) 2013](#)). CRE is less prevalent than MRSA and VRE, causing 9,000 infections and 600 deaths per year ([Centers for Disease Control and Prevention \(CDC\) 2013](#)), but is an immediate public health threat because infections caused by CRE (e.g., pneumonia, urinary tract infections, bloodstream infections and wound infections ([Tischendorf, de Avila, and Safdar 2016](#))) are very difficult to treat ([Jacob et al. 2013](#)) and have been associated with poor treatment outcomes ([Morrill et al. 2015](#)) ([Borer et al. 2009](#); [Papadimitriou-Olivgeris et al. 2013](#); [Schwaber et al. 2008](#)), and high costs ([Bartsch et al. 2017](#)).

Colonized patients carry an MDRO at a detectable level, meaning that a cultured swab sample would test positive but the patient would not show clinical indications of illness caused by an MDRO. Harboring MDROs, these patients are at a risk for subsequent infection, as a significant fraction of MDRO colonization will eventually cause clinically apparent infections that are difficult to treat ([Coello et al. 1997](#)) ([Diekmann, Heesterbeek, and Britton 2012](#)). They also pose a threat to other patients, as healthcare workers who interact with these patients can become contaminated with the organism and transmit it to other patients. As a result, it is important to rapidly identify and then monitor colonized

It is made available under a [CC-BY-NC-ND 4.0 International license](#) .

A Data-Driven Framework for Identifying Intensive Care Unit Admissions Colonized with Multidrug-Resistant Organisms

patients so that the colonization rate is kept under control in hospitals, and the risk of transmission and subsequent infection is minimized (Furuno et al. 2004).

The importation of MDROs into hospitals and other healthcare settings is a major determinant for transmission and outbreak (Tacconelli 2006) (D'agata, Horn, and Webb 2002) (Ziakas et al. 2013). Among hospital departments, intensive care units (ICUs) are the wards where the prevalence of MDROs is particularly high (Zhanet et al. 2008) (Hanberger et al. 2009). Further, patients admitted to ICUs are most vulnerable to develop infections due to these organisms (Vincent et al. 2009; Vincent 2003). Accordingly, ICUs have become a central point of focus for the control and prevention of MDRO colonization and infection within hospitals (Strich and Palmore 2017).

A variety of interventions have been proposed and implemented in order to prevent the transmission of MDROs in ICUs. Effective and commonly utilized interventions include (i) hand hygiene, especially when healthcare workers contact colonized or infected patients (Boyce et al. 2009), (ii) contact precautions (e.g., wearing gloves and gowns) when caring for colonized or infected patients (Siegel, Rhinehart, Jackson, and Chiarello 2007), and (iii) isolation or cohorting of colonized or infected patients (Landelle, Pagani, and Harbarth 2013). Despite their effectiveness, however, these preventive measures are often not applied in a timely manner due to imperfect compliance and the delay (or even failure) to detect patients colonized with an MDRO (Harris et al. 2013).

Surveillance for MDRO colonization is an instrumental practice for detecting patients who may require an intervention (Humphreys 2014) (Pofahl et al. 2009; Schwaber and Carmeli 2013). Yet, the implementation and cost-effectiveness of universal (i.e., active) surveillance strategies, such as screening of all newly admitted ICU patients, has been a controversial topic (Edmond and Wenzel 2013). Some critics argue that the costs associated with universal screening, including the opportunity costs of the human and physical resources being utilized, are likely to outweigh the benefits of active surveillance (Wenzel and Edmond 2010). Accordingly, universal surveillance of all patients may not be feasible to implement in many healthcare facilities due to resource constraints (Tacconelli and Cataldo 2008) (Roth et al. 2016) (Lapointe-Shaw et al. 2017). Instead, targeted surveillance strategies, which offer a cost-effective compromise for detecting asymptomatic colonization, have been advocated by national guidelines (Muto et al. 2003) (Siegel, Rhinehart, Jackson, Chiarello, et al. 2007), (Weber et al. 2007) when a sufficiently accurate method for identifying high-risk individuals is available. Accordingly, rapid identification of patients who are at high risk for MDRO colonization is critical for timely and targeted implementation of screening and other preventive measures, as well as administration of appropriate treatment (e.g., avoiding the misuse of antibiotics).

Given the aforementioned challenges, a system that facilitates timely and reliable identification of newly admitted patients who are likely to be colonized with an MDRO would be quite useful to improve patient safety and effective utilization of critical hospital resources (Delerue et al. 2019). By accurately identifying significant risk factors, this system can help define high-risk subpopulations and hence, could enable the

It is made available under a [CC-BY-NC-ND 4.0 International license](https://creativecommons.org/licenses/by-nc-nd/4.0/) .

A Data-Driven Framework for Identifying Intensive Care Unit Admissions Colonized with Multidrug-Resistant Organisms

implementation of a cost-effective targeted screening program. Moreover, if highly predictive, it can further be used to immediately initiate clinical interventions, such as contact precautions, as soon as a high-risk individual is admitted to the ICU. Such a real-time system would be particularly useful in ICUs because, currently, identification of colonized patients relies on clinical lab results that usually require at least 1-2 days to process and can delay subsequent action to prevent and control the spread of MDROs.

A particular challenge for the design of a reliable prediction framework is the class imbalance problem that is commonly observed in clinical datasets. Clinical datasets are often not balanced in their class labels, where the predictors and/or prediction outcomes do not make up an equal portion of the data. The imbalance can be particularly large when the prediction outcomes are MDROs, as their prevalence is usually $< 15\%$ and can be as low as $< 2\%$ as observed in our data. Given that ignoring the class imbalance, especially when it is large, yields poor predictions, it is necessary to consider and address this challenge while developing a prediction framework for accurate and reliable results.

In this study, we developed a data-driven framework to identify patients who are likely to be colonized with VRE, CRE, or MRSA upon ICU admission, leveraging two years of electronic health record (EHR) data from a large academic medical center. Our objective was to develop a modeling framework that can cope with significant class imbalance, commonly observed in clinical datasets, and can be used to (1) to generate timely and accurate predictions for newly admitted ICU patients, and (2) to identify the key socio-demographic and clinical factors affecting the incidence of MDRO colonization. Our framework relied on three supervised machine learning algorithms, specifically regularized logistic regression, random forest, and XGBoost, which were trained on the EHR data and facilitate real-time predictions for newly admitted patients to the ICU.

We achieved the following results for the primary MDRO colonization outcomes: 80% sensitivity and 66% specificity for VRE, 73% and 77% for CRE, 76% and 59% for MRSA, and 82% and 83% for colonization with any MDRO (i.e., VRE, CRE, or MRSA). Moreover, our modeling approach identified long-term care facility exposure, current diagnosis of skin/subcutaneous tissue conditions or infectious/parasitic disease, and recent isolation precaution procedures before ICU admission as key predictors. We were able to detect over 80% of positive MDRO cases upon ICU admission with less than a 20% false-positive rate, which would enable timely and targeted implementation of preventive measures for infection control in ICUs.

The remainder of this article is organized as follows: In Section 2, we present our data and describe our methodology. In particular, in Section 2.1, we introduce our data and describe the clinical and socio-demographic predictors included in our models. Then, in Section 2.2, we introduce the predictive models and then describe the techniques we utilize to improve prediction accuracy and address class imbalance for our application. In Section 3, we present our prediction results and report the key predictors for MDRO colonization in our data set. In Section 4, we summarize our results, discuss the policy implications of our approach and findings, propose directions for future research, and conclude our study.

A Data-Driven Framework for Identifying Intensive Care Unit Admissions Colonized with Multidrug-Resistant Organisms

2. Materials and Methods

In this section, we first describe our data source, in Section 2.1, and present the variables and prediction outcomes in our dataset. Then, in Section 2.2, we introduce our modeling framework and describe our methods. In particular, first, we introduce the prediction models we used, and then, discuss our model specification (training) and performance evaluation (testing) stages, describing how we performed hyperparameter tuning, stratified cross-validation, threshold optimization, and finally, out-of-sample evaluations.

2.1 Data Description

In this study, we used electronic healthcare record (EHR) data from the University of Maryland Medical Center (UMMC), an academic teaching hospital located in Baltimore, Maryland. Our dataset contained records for 3,958 patients admitted to a surgical or medical ICU in 2017 or 2018. In total, we observed 4,670 individual admissions. Our dataset included the following variables: (1) hospital admission source and type, (2) age, (3) sex, (4) race and ethnicity, (5) region/state of residency, (6) total time admitted during prior ICU and hospital inpatient stays within the previous year, (7) prior antibiotic prescription, (8) diagnoses for prior hospital and/or ICU stays within the previous year, (9) diagnoses for current hospital stay that are present before ICU admission, (10) surgical and medical procedures conducted during prior hospital and/or ICU stays within the previous year, and (11) recent procedures conducted for current hospital stay prior to ICU admission. We treated all predictors utilized in the models as categorical. Descriptive statistics regarding these variables and their categories can be found in [Appendix A](#).

In the dataset, all prior and current diagnoses were coded using the International Statistical Classification of Diseases and Related Health Problems (ICD)-10 codification. We used the Agency for Healthcare Research and Quality's Clinical Classifications Software (CCS) to further categorize the prior and current diagnoses that were present on admission (PoA). The CCS is a diagnosis and procedure categorization catalog (<https://www.hcup-us.ahrq.gov/toolssoftware/ccs10/ccs10.jsp>), mapping the ICD-10 diagnosis codes into 18 categories: (1) Infectious and parasitic diseases, (2) Neoplasms, (3) Endocrine, nutritional, and metabolic diseases and immunity disorders, (4) Diseases of the blood and blood-forming organs, (5) Mental illness, (6) Diseases of the nervous system and sense organs, (7) Diseases of the circulatory system, (8) Diseases of the respiratory system, (9) Diseases of the digestive system, (10) Diseases of the genitourinary system, (11) Complications of pregnancy, childbirth, and the puerperium, (12) Diseases of the skin and subcutaneous tissue, (13) Diseases of the musculoskeletal system and connective tissue, (14) Congenital anomalies, (15) Certain conditions originating in the perinatal period, (16) Injury and poisoning, (17) Symptoms, signs, and ill-defined conditions and factors influencing health status, and (18) Residual or unclassified codes.

It is made available under a [CC-BY-NC-ND 4.0 International license](https://creativecommons.org/licenses/by-nc-nd/4.0/) .

A Data-Driven Framework for Identifying Intensive Care Unit Admissions Colonized with Multidrug-Resistant Organisms

We labeled a procedure as recent if it was performed during the current hospital stay. We recorded all recent procedures performed in the hospital inpatient settings prior to the current ICU admission with respect to the ICD-10 Procedure Coding System (PCS), for which each character has a categorical indication. Using the first character of the ICD-10 PCS codes, we classify the recent procedures into eight categories as follows: (i) Medical and Surgical (“0”), (ii) Placement (“2”), (iii) Administration (“3”), (iv) Measurement and Monitoring (“4”), (v) Extracorporeal or Systemic Procedures (“5” and “6”), (vi) Other Procedures (“8”), (vii) Imaging (“B”), and (viii) Other/Miscellaneous (“1”, “7”, “9”, “C”, “D”, “F”, “G”, and “X”). Further, using the first two characters of the ICD-10 PCS codes, we also map the recent procedures into 44 categories (See [Appendix A](#)). In our analysis, we include both the single- and double-character based categorizations so that our algorithms can learn which specifications are most important for predicting our MDRO outcomes. We classified prior hospital procedures having the ICD-10 PCS codes in a similar manner as the recent procedures.

Prior outpatient procedures were recorded using the Current Procedural Terminology (CPT) system (<https://www.ama-assn.org/amaone/cpt-current-procedural-terminology>), which we classified into 6 categories: (i) Evaluation and Management, (ii) Anesthesia (iii) Medicine (iv) Radiology (v) Pathology and Laboratory, and (vi) Surgery. The CPT codes for surgery include 18 sub-types, enabling us to construct a more detailed categorization with 23 classes. We used both the 6-class and 23-class CPT codes for our analysis.

The prediction outcomes were colonization with VRE, CRE, or MRSA upon ICU admission, both separately and as an aggregate (union) outcome. Conducting active surveillance in the ICUs, UMMC screened newly admitted patients for colonization with these organisms upon admission and periodically during their stay. At UMMC, active surveillance involves taking routine peri-rectal cultures for VRE and nasal cultures for MRSA on all patients admitted to an ICU at the time of admission, weekly and upon discharge. CRE detection was also primarily done via perirectal swabs and also included clinical cultures (e.g., blood, urine, wound cultures). We identified the positive (i.e., colonized) and negative (i.e., uncolonized) results based on the laboratory tests conducted within two days (i.e., both before and after) of ICU admissions. We limited the time window for the screening results within two days ([Paling, Wolkewitz, Bode, et al. 2017](#); [Paling, Wolkewitz, Depuydt, et al. 2017](#)) in an attempt to avoid inclusion of acquisition cases, for which initially susceptible patients acquire an MDRO during their ICU stay. In our data set, 13.03% of patients admitted to the ICU tested positive for VRE, 1.45% for CRE, 7.47% for MRSA, and 17.59% for any MDRO.

2.2 Prediction Models, Model Training and Validation, and Threshold Optimization

A variety of techniques have been utilized to analyze complex disease dynamics and quantify its parameters (e.g., the estimation of transmission rate), identify risk factors, and estimate the impact of infection control strategies ([van Kleef et al. 2013](#)). These approaches include prediction modeling, computational simulation, and analytic-formula based models such as decision trees ([Goodman et al. 2016](#)), artificial neural network

It is made available under a [CC-BY-NC-ND 4.0 International license](#) .

A Data-Driven Framework for Identifying Intensive Care Unit Admissions Colonized with Multidrug-Resistant Organisms

(Chang et al. 2011), agent-based simulation for a hospital ward (Barnes, Golden, and Wasil 2010b; Codella et al. 2015) or healthcare system (Lee et al. 2013), dynamic patient and healthcare worker networks (Barnes, Golden, and Wasil 2010a; Cusumano-Towner et al. 2013; Ueno and Masuda 2008), compartmental systems dynamics models (based on ordinary differential equations) (D'Agata et al. 2012; de Cellès et al. 2013), (approximate) Bayesian (computation) techniques (Cooper et al. 2008), and Markov chain based approaches (Kastner and Shachtman 1982 ; Bootsma et al. 2007). Among these techniques, data-driven prediction models, such as the ones we use in this study, are particularly valuable tools for generating real-time predictions, identifying the significant risk factors, and quantifying their impact on the outcomes under study (Wiens and Shenoy 2017). In addition to these modeling-based approaches, there is also rich clinical literature studying MDRO colonization. See Appendix B for a summary of the clinical studies that assessed the risk factors associated with MDRO colonization, and developed simple clinical prediction rules based on the identified predictors.

We utilized three supervised machine learning (ML) algorithms to predict colonized patients upon ICU admission and to identify significant clinical and socio-demographic factors associated with our outcomes of interest: (1) logistic regression (Reed and Berkson 1929) (Berkson 1944), (2) random forest (Breiman 2001), and (3) XGBoost (Chen and Guestrin 2016). To perform regularization and feature selection for our logistic regression models, we used least absolute shrinkage and selection operator (LASSO), which was originally developed for linear regression (Tibshirani 1996) and then applied to other algorithms including logistic regression (Roth 2004).

For each model, we split the data into an 80% subset for training and cross-validation and a 20% subset for out-of-sample evaluation. We used a 10-fold stratified cross validation scheme for both the tuning of hyperparameters for the algorithm and the optimization of the threshold for classification of the predictions (see Figure 1). We selected the 10-fold due to the relatively small sample size of our data, in an effort to preserve as much data as possible for model training. We selected the stratified scheme to account for the class imbalance in our data, which preserves a similar proportion of the positive outcome for each fold as the complete data set.

We defined a grid search for a core set of hyperparameters for each algorithm (see Appendix D), and used the area under the receiver operating characteristic curve (AUC) as the objective function to maximize (out-of-sample) model performance. We selected the hyperparameters achieving the highest mean AUC across the 10 folds for model training.

After choosing the hyperparameters, the next step of the model specification was to identify the ideal cut-off value (i.e., threshold optimization) for converting predicted probabilities into binary predictions. As an initial output, the ML algorithms generates predicted probabilities for the training instances, indicating how likely each patient to be colonized with an MDRO. These predicted probabilities are then translated into binary prediction outcomes using a threshold value. Specifically, observations for which the predicted probabilities are greater than this threshold, τ , are classified as positive (i.e.,

It is made available under a [CC-BY-NC-ND 4.0 International license](#) .

A Data-Driven Framework for Identifying Intensive Care Unit Admissions Colonized with Multidrug-Resistant Organisms

colonized), otherwise, the patient is assigned to the negative (i.e., susceptible) class. Given the class imbalance observed in our dataset, the default threshold value of 0.5 was unlikely to be effective for our study (see [Figure 2](#)). Consequently, we performed an optimization ([Sheng and Ling 2006](#)) to search for the best threshold that classifies the predicted probabilities while maximizing the Youden Index (i.e., sensitivity + specificity - 1) for out-of-sample predictions ([Youden 1950](#)).

We performed the threshold optimization out-of-sample using the same 10-fold stratified cross validation scheme used for the hyperparameter tuning. We determined the optimal threshold for each fold using the in-sample predicted probabilities from the 90% subset of training data, and then evaluated the performance (i.e., Youden's index) of this threshold over the 10% subset of out-of-sample data. We repeated this process for each fold, and selected the mean of these 10 optimal thresholds as our optimized value. We used a bounded numerical search algorithm to solve the optimization problem ([Nocedal and Wright 2006](#)), using a lower bound of zero and varying the upper bound for each algorithm to ensure an effective threshold is found. It is noteworthy to emphasize that the upper bound values we considered for each specific outcome were different because the prevalence of the colonized (i.e., positive) instances among VRE, CRE, MRSA, and MDRO were different and directly affected the outcome of the threshold optimization procedure.

Model specification was completed when we determined the hyperparameters, chose the threshold value (for each model), and re-trained the models on the full (80%) training set. Next, we evaluated the performance of the trained models on the (20%) test sets, reporting the AUC, sensitivity, and specificity values obtained. For each MDRO, we conducted a systematic numerical experiment with a range of upper bound values for threshold optimization, and obtained predictions with varying sensitivity and specificity values for VRE, CRE, MRSA, and MDRO (the aggregate prediction outcome). We provide these results in Section 4 for each outcome (e.g., VRE) and algorithm (e.g., XGBoost), and separately, discuss the best performing models for each MDRO.

We also use our modeling framework to identify the key socio-demographic and clinical factors for predicting colonization with VRE, CRE, and MRSA separately and in aggregate. For the LR models, we use odds ratios (ORs), which quantify the associated increase (for values greater than 1) or decrease (for values less than 1) in the likelihood of colonization. For the tree-based models (i.e., RF and XGBoost), we use feature importance (FI), which quantifies the relative frequency that each factor is used to construct the ensemble. Using these two metrics (i.e., OR and FI), we order the identified predictors for each MDRO and report the top five key predictors that are highly ranked across all of the best performing ML models, calculated by the average ranking across the best models.

3. Results

We separately predict VRE, CRE, and MRSA colonization upon ICU admission. In addition, combining these three antibiotic-resistant bacteria, we also predict colonization

It is made available under a [CC-BY-NC-ND 4.0 International license](#) .

A Data-Driven Framework for Identifying Intensive Care Unit Admissions Colonized with Multidrug-Resistant Organisms

with any of these MDROs (i.e., VRE, CRE, or MRSA) upon ICU admission without specifying the particular organism. As a result, we generate predictions for four cases (namely, VRE, CRE, MRSA, and MDRO) using logistic regression (with LASSO regularization, LR), random forest (RF), and XGBoost algorithms. In [Table 1](#), we summarize the model results for these four outcomes for different upper bound values for the threshold optimization.

After considering all of the models that we trained for each outcome, we selected the ones with the highest (out-of-sample) Youden index, which we summarize in [Table 2](#). For VRE, the best performing model generated a Youden index of 0.46, achieved via the LR model. By comparison, the RF and XGBoost models generated Youden index values of 0.41 and 0.39, respectively. For CRE, the XGBoost algorithm generate the highest Youden index (0.50), followed by LR (0.45) and RF (0.42). The performance for MRSA is noticeably lower than the other outcomes, for which RF achieved the highest Youden index (0.34). Finally, the prediction models for the aggregate MDRO outcome produced the highest Youden index values when compared to the individual MDRO outcomes, with the RF model (0.65) outperforming the XGBoost (0.57) and LR models (0.30). We note here that the tree-based models performed significantly better than the linear LR model for this aggregated outcome, which is likely due to the former's natural ability to capture interactions. In an effort to provide support for this hypothesis, we also show the performance of a single classification tree ([Breiman et al. 1984](#)) (0.54), which also performs significantly better than the LR model for this particular outcome. On the other hand, for separate VRE, CRE, and MRSA predictions, the single tree models are always dominated by (at least one of) the other algorithms, and hence, not presented in [Table 1](#).

For each model presented in [Table 2](#), the difference between the (out-of-sample) AUC for the (cross-validated) training and testing sets are typically small, suggesting well-trained models. The LR and RF models for CRE demonstrate larger gaps, suggesting that these models might be slightly less robust than others; however, this volatility is likely explained by the extremely low prevalence of positive cases on which to train the models. The best predictions for VRE colonization upon ICU admission are generated by the LR model, which achieves 80% sensitivity and 66% specificity. For CRE, XGBoost produces the best model, having 73% sensitivity and 77% specificity. For MRSA, the RF model performs best, yielding 76% sensitivity and 59% specificity. Finally, the most effective model for the aggregate MDRO outcome is the random forest model, which is capable of detecting 82% of colonized patients with 83% specificity.

In addition to generating predictions, we also use our modeling framework to identify the key predictors for separate and aggregate VRE, CRE, and MRSA colonization. In [Table 3](#), we summarize the top five predictors for the models reported in [Table 2](#), and provide their ranking in the corresponding models as indicated by OR and FI. See [Appendix C](#) for the OR and FI values of the factors presented in [Table 3](#).

Among the recent ICD-10 procedures that were performed during the current hospital stay before ICU admission, procedures categorized as “Other Procedures” in the ICD-10 PCS are among the top five predictors for VRE, CRE, MRSA, and MDRO. In our

It is made available under a [CC-BY-NC-ND 4.0 International license](#) .

A Data-Driven Framework for Identifying Intensive Care Unit Admissions Colonized with Multidrug-Resistant Organisms

dataset, a significant proportion of these procedures are “8E0ZXY6”, an ICD-10 code designated for isolation precautions. The patients having a history of a prior colonization or infection for a given MDRO (or are at risk for another indication) are flagged with this code upon admission to the hospital so that they are closely monitored (and if needed, isolated) during their hospital stay. Our results presented in [Table 3](#) show that these patients are at a higher risk for being colonized with an MDRO at ICU admission regardless of the specific indication for which the close monitoring and isolation precautions are put in place.

Another key predictor for VRE, CRE, MRSA, and MDRO colonization is the CCS-based diagnosis category “skin and subcutaneous tissue disease” that is PoA ([Table 3](#)). The diagnoses that fall under this CCS category are determined for the current hospital admission and include rash, cellulitis, cutaneous abscess, pressure ulcer, non-pressure chronic ulcer, and other skin conditions. Our finding resonates with the clinical literature and practice as skin and soft tissue infections are amongst the most common bacterial infections, are mostly treated with antibiotics that might cause antimicrobial resistance ([Eckmann and Dryden 2010](#)). Further, skin and soft tissue infections are the most frequently reported clinical manifestations of community-acquired MRSA ([Fridkin et al. 2005](#)).

For MDRO and in particular MRSA, the CCS-based current diagnosis category “infectious and parasitic diseases” is one of the critical factors that increase the risk of colonization. This category includes diseases such as chronic viral hepatitis C, bacteremia, human immunodeficiency virus (HIV), and sepsis. Patients with these diseases might be at higher risk for MDRO, and in particular MRSA, colonization due to a compromised immune system.

For VRE and CRE, having a prior long-term care facility (LTCF) stay is one of the key predictors for colonization upon ICU admission. This association between VRE or CRE colonization and a previous LTCF stay has been reported by other studies ([Prabaker et al. 2012](#); [Tacconelli et al. 2004](#)) (also see [Appendix B](#)). High rates of MDRO colonization, debilitating diseases, and the receipt of multiple antibiotics among LTCF residents are likely to be the primary causes of this association both for VRE and CRE colonization ([Elizaga, Weinstein, and Hayden 2002](#)) .

Other key predictors for VRE were recent procedures “administration circulatory” (ICD-10-PCS '30'), such as transfusion, and “medical and surgical anatomical regions, general” ('0W'), such as drainage, insertion, removal, and transplantation procedures. For CRE, a prior ICU stay longer than 20 days and a total number of diagnoses PoA (i.e., current diagnoses) greater than 30 were two critical factors increasing the risk of colonization. For MRSA, the current diagnosis for “injury and poisoning”, mostly consisting of procedural injuries such as accidental puncture or dural laceration during a procedure, is associated with an increased colonization risk. On the contrary, the recent procedure code for “administration” (i.e., ICD-10 PCS codes with first character “3”) was found to lower the risk of colonization. Finally, female sex and the “mental illness” category for current diagnosis, including diagnosis for cocaine abuse, opioid abuse, poisoning by heroin and

It is made available under a [CC-BY-NC-ND 4.0 International license](#) .

A Data-Driven Framework for Identifying Intensive Care Unit Admissions Colonized with Multidrug-Resistant Organisms

psychological disorders, were two other key factors associated with an increased risk for MDRO colonization. Patients in this category (i.e., the “mental illness”) are at higher risk for using injections and causing damage to their skin, which might explain the increased risk for MDRO colonization.

4. Discussions, Future Work, and Conclusion

Leveraging a rich EHR data set and supervised ML algorithms, we developed an accurate and interpretable framework for predicting VRE, CRE, and MRSA colonization upon ICU admission. We achieved the following sensitivity and specificity values for VRE, CRE, and MRSA colonization: 80% and 66% for VRE with LR, 73% and 77% for CRE with XGBoost, and 76% and 59% for MRSA with RF. Further, we predicted MDRO (i.e., VRE, CRE, or MRSA) colonization as an aggregate outcome with 82% sensitivity and 83% specificity for MDRO using RF.

Our results indicate that predicting MDRO colonization in aggregate, rather than separately predicting VRE, CRE, and MRSA, achieved the highest prediction accuracy in terms of both AUC and Youden’s index. On the one hand, predicting a specific MDRO would be preferable, as it would enable more customized interventions such as tailored antibiotic therapy. On the other hand, accurately predicting MDRO colonization without specifying whether it is VRE, CRE, or MRSA is still quite important for clinical practice as key interventions for these MDROs are similar such as contact precautions and enhanced environmental cleaning. Accordingly, many infection control measures can be implemented rapidly upon ICU admission for the patients who are suspected to be colonized, and treatment strategies and more advanced interventions can be tailored later as more information becomes available.

In addition to producing timely predictions for newly admitted ICU patients, our ML-based modeling framework can also be utilized to identify the key predictors for VRE, CRE, and MRSA colonization upon ICU admission. We identified several important predictors of MDRO colonization, including long-term care facility exposure, a current diagnosis of skin/subcutaneous tissue or infectious/parasitic disease, and a recent ICD-10 procedure “Other Procedures”, including isolation precaution procedures, as the key predictors for MDRO colonization upon ICU admission. These predictors can help identify ICU patients at high-risk for MDRO colonization and hence, facilitate timely implementation of infection control measures such as selective use of contact precautions, targeted surveillance, and tailored antibiotic therapy.

The primary limitation of our study is that we do not utilize any data on patient medical history outside of UMMC. For example, we do not take into account any antibiotics consumption outside of UMMC or during outpatient visits. Similarly, we do not have information about patients who could have been admitted elsewhere, thus censoring any information about whether they received or underwent additional treatments and

It is made available under a [CC-BY-NC-ND 4.0 International license](#) .

A Data-Driven Framework for Identifying Intensive Care Unit Admissions Colonized with Multidrug-Resistant Organisms

procedures in other healthcare facilities. As we utilize administrative data for procedures and diagnoses, which are primarily used for billing, we do not have full access to exact clinical conditions and we do not know the exact reason why a specific procedure was performed or diagnosis was established. Our discussions with clinicians shed some light into these uncertainties but we could not determine the exact details for each individual patient other than what the data conveys. Finally, our data is derived from a single source and we are only able to observe the performance of our modeling framework on an out-of-sample subset from the same facility.

It is noteworthy to emphasize that our study, which focused on predicting MDRO colonization for newly admitted ICU patients, would not prevent the importation of VRE, CRE, and MRSA into the ICU setting. However, by producing reliable predictions and identifying key risk factors for colonization, our approach could enable early detection of colonized patients and facilitate timely and targeted implementation of preventive measures on asymptomatic MDRO carriers. This approach could help reduce transmission of these so-called "superbugs" in ICUs, which we plan to further analyze via a comprehensive simulation model in future efforts, and would particularly be useful for healthcare settings where active surveillance is not performed.

Traditionally, many prediction rules, developed as a decision support tool for clinicians, are designed to be very simple, relying on only a small number of variables, for practicality. Yet, with the increasing availability of electronic healthcare record (EHR) data and the expansion of modern database and software systems, the use of data-driven prediction models and other analytical and computational methods for the identification, control, and prevention of MDROs and other HAIs has been increasing ([van Kleef et al. 2013](#)). As a result, a growing number of healthcare facilities are capable of generating more complex prediction models in an automated fashion. Accordingly, taking advantage of the advances in computational and data recording technologies, many healthcare organizations can use our data-driven prediction framework to produce real-time predictions and identify the high-risk patients for MDRO colonization.

There are three research directions that we plan to pursue in near future: First, we will study the acquisition outcomes, where we focus on the ICU patients who were initially colonization-free but acquired VRE, CRE, or MRSA colonization during their ICU stay. Second, we will develop a comprehensive agent-based simulation model to analyze MDRO colonization and infection in ICUs and assess the impact of commonly utilized prevention and control measures on MDRO transmission. Finally, we plan to acquire more data from another major healthcare facility and conduct a similar study by leveraging this additional dataset. This will not only enable us to enlarge the size our dataset, leading to more accurate predictions, but will also give us an opportunity to assess the generalizability of our findings and help us develop more robust predictions.

To summarize, in this study, we proposed a data-centric modeling framework to predict VRE, CRE, and MRSA colonization upon ICU admission and identify the associated risk factors. We achieved the highest prediction accuracy, measured by Youden's index, when we predicted VRE, CRE, and MRSA colonization combined as an aggregate outcome.

It is made available under a [CC-BY-NC-ND 4.0 International license](#) .

A Data-Driven Framework for Identifying Intensive Care Unit Admissions Colonized with Multidrug-Resistant Organisms

Capable of coping with significant class imbalance, a feature commonly observed in clinical datasets, our framework can be used as a clinical decision support tool to provide accurate on-time predictions especially if it is regularly updated and trained off-line as additional (i.e., more recent) data become available. It can further be used to identify the key risk factors and define high-risk populations, for which targeted interventions can be implemented rapidly to reduce transmission of MDROs in ICUs.

Acknowledgement: This study is supported by the U.S. Centers for Disease Control and Prevention (CDC) Modeling Infectious Disease (MInD) Network prime award number 1U01CK000536. The content is solely the responsibility of the authors and does not necessarily represent the official views of the CDC.

Conflict of Interest Disclosure: No potential conflict of interest was reported by the authors.

A Data-Driven Framework for Identifying Intensive Care Unit Admissions Colonized with Multidrug-Resistant Organisms

References:

- Barnes, S., B. Golden, and E. Wasil. 2010a. 'A dynamic patient network model of hospital-acquired infections', in *Proceedings of the Winter Simulation Conference*: 2249-60.
- . 2010b. 'MRSA transmission reduction using agent-based modeling and simulation', *INFORMS Journal on Computing*, 22: 635-46.
- Bartsch, S.M., J.A. McKinnell, L.E. Mueller, L.G. Miller, S.K. Gohil, S.S. Huang, and B.Y. Lee. 2017. 'Potential economic burden of carbapenem-resistant Enterobacteriaceae (CRE) in the United States', *Clinical Microbiology and Infection*, 23: 48-e9.
- Bassetti, M., E. Nicco, and M. Mikulska. 2009. 'Why is community-associated MRSA spreading across the world and how will it change clinical practice?', *International journal of antimicrobial agents*, 34: S15-S19.
- Berkson, J. 1944. 'Application of the logistic function to bio-assay', *Journal of the American Statistical Association*, 39: 357-65.
- Bootsma, M.C.J., M.J.M. Bonten, S. Nijssen, A.C. Fluit, and O. Diekmann. 2007. 'An algorithm to estimate the importance of bacterial acquisition routes in hospital settings', *American journal of epidemiology*, 166: 841-51.
- Borer, Abraham, Lisa Saidel-Odes, Klaris Riesenber, Seada Eskira, Nejama Peled, Ronit Nativ, Francise Schlaeffer, and Michael Sherf. 2009. 'Attributable mortality rate for carbapenem-resistant *Klebsiella pneumoniae* bacteremia.', *Infection Control & Hospital Epidemiology*, 30: 972-76.
- Boucher, H.W., G.H. Talbot, J.S. Bradley, J.E. Edwards, D. Gilbert, L.B. Rice, M. Scheld, B. Spellberg, and J. Bartlett. 2009. 'Bad bugs, no drugs: no ESKAPE! An update from the Infectious Diseases Society of America', *Clinical infectious diseases*, 48: 1-12.
- Boyce, J., Y. Chartier, M. Chraiti, B. Cookson, N. Damani, and S. Dharan. 2009. 'WHO guidelines on hand hygiene in health care. Geneva: World Health Organization'. <https://www.who.int/gpsc/5may/tools/9789241597906/en/>.
- Breiman, L. 2001. 'Random forests', *Machine learning*, 45: 5-32.
- Breiman, L., H. Freidman, A. Olshen, and C.J. Stone. 1984. *Classification and regression trees* (Wadsworth & Brooks/Cole Advanced Books & Software: Monterey CA).
- Calfee, D.P. 2012. 'Methicillin-resistant *Staphylococcus aureus* and vancomycin-resistant enterococci, and other Gram-positives in healthcare', *Current opinion in infectious diseases*, 25: 385-94.

A Data-Driven Framework for Identifying Intensive Care Unit Admissions Colonized with Multidrug-Resistant Organisms

- Centers for Disease Control and Prevention (CDC). 2013. "Antibiotic resistance threats in the United States, 2013." <https://www.cdc.gov/drugresistance/pdf/antibiotic-threats-2013-508.pdf>." In. Atlanta, GA, USA.
- . 2015. 'Facility guidance for control of carbapenem-resistant Enterobacteriaceae (CRE)—November 2015 update CRE toolkit.' <https://www.cdc.gov/hai/pdfs/cre/CRE-guidance-508.pdf>.
- . 2018. 'Antibiotic / Antimicrobial Resistance - Biggest Threats and Data'. <https://www.cdc.gov/drugresistance/biggest-threats.html>.
- Chang, Y.J., M.L. Yeh, Y.C. Li, C.Y. Hsu, C.C. Lin, M.S. Hsu, and W.T. Chiu. 2011. 'Predicting hospital-acquired infections by scoring system with simple parameters', *PloS one*, 6: e231-37.
- Chen, T., and C. Guestrin. 2016. 'Xgboost: A scalable tree boosting system', *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, ACM: 785-94.
- Codella, J., N. Safdar, R. Heffernan, and O. Alagoz. 2015. 'An agent-based simulation model for Clostridium difficile infection control', *Medical Decision Making*, 35: 211-29.
- Coello, R., J.R. Glynn, C. Gaspar, J.J. Picazo, and J. Fereres. 1997. 'Risk factors for developing clinical infection with methicillin-resistant Staphylococcus aureus (MRSA) amongst hospital patients initially only colonized with MRSA', *Journal of Hospital Infection*, 37: 39-46.
- Cooper, B.S., G.F. Medley, S.J. Bradley, and G.M. Scott. 2008. 'An augmented data method for the analysis of nosocomial infection data', *American journal of epidemiology*, 168: 548-57.
- Cosgrove, S.E., Y. Qi, K.S. Kaye, S. Harbarth, A.W. Karchmer, and Y. Carmeli. 2005. 'The impact of methicillin resistance in Staphylococcus aureus bacteremia on patient outcomes: mortality, length of stay, and hospital charges', *Infection Control & Hospital Epidemiology*, 26: 166-74.
- Cosgrove, S.E., G. Sakoulas, E.N. Perencevich, M.J. Schwaber, A.W. Karchmer, and Y. Carmeli. 2003. 'Comparison of mortality associated with methicillin-resistant and methicillin-susceptible Staphylococcus aureus bacteremia: a meta-analysis.', *Clinical infectious diseases*, 36: 53-59.
- Cusumano-Towner, M., D.Y. Li, S. Tuo, G. Krishnan, and D.M. Maslove. 2013. 'A social network of hospital acquired infection built from electronic medical record data', *Journal of the American Medical Informatics Association*, 20: 427-34.
- D'Agata, E.M., M.A. Horn, S. Ruan, G.F. Webb, and J.R. Wares. 2012. 'Efficacy of infection control interventions in reducing the spread of multidrug-resistant organisms in the hospital setting', *PloS one*, 7: e30170.
- D'agata, E.M., M.A. Horn, and G.F. Webb. 2002. 'The impact of persistent gastrointestinal colonization on the transmission dynamics of vancomycin-resistant enterococci', *The Journal of infectious diseases*, 185: 766-73.
- de Cellès, M.D., J.R. Zahar, V. Abadie, and D. Guillemot. 2013. 'Limits of patient isolation measures to control extended-spectrum beta-lactamase-producing Enterobacteriaceae: model-based analysis of clinical data in a pediatric ward', *BMC infectious diseases*, 13: 187.

It is made available under a [CC-BY-NC-ND 4.0 International license](https://creativecommons.org/licenses/by-nc-nd/4.0/) .

A Data-Driven Framework for Identifying Intensive Care Unit Admissions Colonized with Multidrug-Resistant Organisms

- Delerue, T., H. Cordel, J. Figoni, S. Dziri, T. Billard-Pomares, O. Bouchaud, E. Carbonnelle, and J.R. Zahar. 2019. 'Prediction of methicillin-resistant *Staphylococcus aureus* bloodstream infection: do we need rapid diagnostic tests?', *European Journal of Clinical Microbiology & Infectious Diseases*: 1-8.
- DiazGranados, C.A., S.M. Zimmer, K. Mitchel, and J.A. Jernigan. 2005. 'Comparison of mortality associated with vancomycin-resistant and vancomycin-susceptible enterococcal bloodstream infections: a meta-analysis', *Clinical infectious diseases*, 41: 327-33.
- Diekmann, O., H. Heesterbeek, and T. Britton. 2012. 'Chapter 14. Data-driven modeling of hospital infections.' in, *Mathematical tools for understanding infectious disease dynamics* (Princeton University Press).
- Eckmann, C., and M. Dryden. 2010. 'Treatment of complicated skin and soft-tissue infections caused by resistant bacteria: value of linezolid, tigecycline, daptomycin and vancomycin.', *European journal of medical research*, 15: 554.
- Edmond, M.B., and R.P. Wenzel. 2013. 'Screening inpatients for MRSA-case closed', *The New England journal of medicine*, 368: 2314.
- Elizaga, M.L., R.A. Weinstein, and M.K. Hayden. 2002. 'Patients in long-term care facilities: a reservoir for vancomycin-resistant enterococci', *Clinical infectious diseases*, 34: 441-46.
- Fridkin, S.K., J.C. Hageman, M. Morrison, L.T. Sanza, K. Como-Sabetti, J.A. Jernigan, K. Harriman, L.H. Harrison, R. Lynfield, and M.M. Farley. 2005. 'Methicillin-resistant *Staphylococcus aureus* disease in three communities', *New England Journal of Medicine*, 352: 1436-44.
- Furuno, J.P., A.D. Harris, M.O. Wright, J.C. McGregor, R.A. Venezia, J. Zhu, and E.N. Perencevich. 2004. 'Prediction rules to identify patients with methicillin-resistant *Staphylococcus aureus* and vancomycin-resistant enterococci upon hospital admission', *American journal of infection control*, 32: 436-40.
- Goodman, K.E., J. Lessler, S.E. Cosgrove, A.D. Harris, E. Lautenbach, J.H. Han, A.M. Milstone, C.J. Massey, and P.D. Tamma. 2016. 'A clinical decision tree to predict whether a bacteremic patient is infected with an extended-spectrum β -lactamase-producing organism.', *Clinical infectious diseases*, 63: 896-903.
- Hanberger, H., D. Arman, H. Gill, V. Jindrák, S. Kalenic, A. Kurcz, M. Licker, P. Naaber, E.A. Scicluna, V. Vanis, and S.M. Walther. 2009. 'Surveillance of microbial resistance in European Intensive Care Units: a first report from the Care-ICU programme for improved infection control. ', *Intensive care medicine*, 35: 91-100.
- Haque, M., M. Sartelli, J. McKimm, and M.A. Bakar. 2018. 'Health care-associated infections—an overview', *Infection and drug resistance*, 11: 2321.
- Harris, A.D., L. Pineles, B. Belton, J.K. Johnson, M. Shardell, M. Loeb, R. Newhouse, L. Dembry, B. Braun, E.N. Perencevich, and K.K. Hall. 2013. 'Universal glove and gown use and acquisition of antibiotic-resistant bacteria in the ICU: a randomized trial', *Jama*, 310: 1571-80.
- Humphreys, H. 2014. 'Controlling the spread of vancomycin-resistant enterococci. Is active screening worthwhile?', *Journal of Hospital Infection*, 88: 191-98.

A Data-Driven Framework for Identifying Intensive Care Unit Admissions Colonized with Multidrug-Resistant Organisms

- Jacob, J.T., G.E. Klein, R. Laxminarayan, Z. Beldavs, R. Lynfield, A.J. Kallen, P. Ricks, J. Edwards, A. Srinivasan, S. Fridkin, and K.J. Rasheed. 2013. 'Vital signs: carbapenem-resistant Enterobacteriaceae', *Morbidity and mortality weekly report*, 62: 165-69.
- Kastner, G.T., and R.H. Shachtman. 1982 'A stochastic model to measure patient effects stemming from hospital-acquired infections', *Operations research*, 30: 1105-33.
- Klevens, R.M., J.R. Edwards, and C.L. Richards. 2007. 'Estimating health care-associated infections and deaths in U.S. hospitals, 2002. ', *Public Health Reports*, 122: 160-66.
- Landelle, C., L. Pagani, and S. Harbarth. 2013. 'Is patient isolation the single most important measure to prevent the spread of multidrug-resistant pathogens?', *Virulence*, 4: 163-71.
- Lapointe-Shaw, L., T. Voruganti, P. Kohler, H.H. Thein, B. Sander, and A. McGeer. 2017. 'Cost-effectiveness analysis of universal screening for carbapenemase-producing Enterobacteriaceae in hospital inpatients', *European Journal of Clinical Microbiology & Infectious Disease*, 36: 1047-55.
- Lee, B.Y., K.F. Wong, S.M. Bartsch, S.L. Yilmaz, T.R. Avery, S.T. Brown, Y. Song, A. Singh, D.S. Kim, and S.S. Huang. 2013. 'The Regional Healthcare Ecosystem Analyst (RHEA): a simulation modeling tool to assist infectious disease control in a health system', *Journal of the American Medical Informatics Association*, 20: e139-e46.
- Maragakis, L.L., E.N. Perencevich, and S.E. Cosgrove. 2008. 'Clinical and economic burden of antimicrobial resistance', *Expert review of anti-infective therapy*, 6: 751-63.
- Molton, J.S., P.A. Tambyah, B.S. Ang, M.L. Ling, and D.A. Fisher. 2013. 'The global spread of healthcare-associated multidrug-resistant bacteria: a perspective from Asia', *Clinical infectious diseases*, 56: 1310-18.
- Morrill, H.J., J.M. Pogue, K.S. Kaye, and K.L. LaPlante. 2015. 'Treatment options for carbapenem-resistant Enterobacteriaceae infections', *In Open forum infectious diseases*, 2: Oxford University Press.
- Muto, C.A., J.A. Jernigan, B.E. Ostrowsky, H.M. Richet, W.R. Jarvis, J.M. Boyce, and B.M. Farr. 2003. 'SHEA guideline for preventing nosocomial transmission of multidrug-resistant strains of Staphylococcus aureus and enterococcus', *Infection Control & Hospital Epidemiology*, 24: 362-86.
- Nocedal, J., and S.J. Wright. 2006. *Theory of Constrained Optimization. Numerical Optimization* (Springer).
- Paling, F.P., M. Wolkewitz, L.G. Bode, P.K. Klouwenberg, D.S.Y. Ong, P. Depuydt, L. De Bus, F. Sifakis, M.J.M. Bonten, and J.A.J.W. Kluytmans. 2017. 'Staphylococcus aureus colonization at ICU admission as a risk factor for developing S. aureus ICU pneumonia', *Clinical Microbiology and Infection*, 23: 49-e9.
- Paling, F.P., M. Wolkewitz, P. Depuydt, L. De Bus, F. Sifakis, M.J. Bonten, and J.A. Kluytmans. 2017. 'P. aeruginosa colonization at ICU admission as a risk factor for developing P. aeruginosa ICU pneumonia', *Antimicrobial Resistance & Infection Control*, 6: 38.

A Data-Driven Framework for Identifying Intensive Care Unit Admissions Colonized with Multidrug-Resistant Organisms

- Papadimitriou-Olivgeris, M., M. Marangos, F. Fligou, M. Christofidou, C. Sklavou, S. Vamvakopoulou, E.D. Anastassiou, and K.S. Filos. 2013. 'KPC-producing *Klebsiella pneumoniae* enteric colonization acquired during intensive care unit stay: the significance of risk factors for its development and its impact on mortality.', *Diagnostic microbiology and infectious disease*, 77: 169-73.
- Pofahl, W.E., C.E. Goettler, K.M. Ramsey, M.K. Cochran, D.L. Nobles, and M.F. Rotondo. 2009. 'Active surveillance screening of MRSA and eradication of the carrier state decreases surgical-site infections caused by MRSA', *Journal of the American College of Surgeons*, 208: 981-86.
- Prabaker, K., M.Y. Lin, M. McNally, K. Cherabuddi, S. Ahmed, A. Norris, K. Lolans, R. Odeh, V. Chundi, R.A. Weinstein, and M.K. Hayden. 2012. 'Transfer from high-acuity long-term care facilities is associated with carriage of *Klebsiella pneumoniae* carbapenemase-producing Enterobacteriaceae: a multihospital study', *Infection Control & Hospital Epidemiology*, 33: 1193-99.
- Reed, L.J., and J. Berkson. 1929. 'The application of the logistic function to experimental data', *The Journal of Physical Chemistry*, 33: 760-79.
- Roth, V. 2004. 'The generalized LASSO', *IEEE transactions on neural networks*, 15.
- Roth, V.R., T. Longpre, D. Coyle, K.N. Suh, M. Taljaard, K.A. Muldoon, K. Ramotar, and A. Forster. 2016. 'Cost analysis of universal screening vs. risk factor-based screening for methicillin-resistant *Staphylococcus aureus* (MRSA)', *PloS one*, 11: e0159667.
- Schwaber, M.J., and Y. Carmeli. 2013. 'An ongoing national intervention to contain the spread of carbapenem-resistant Enterobacteriaceae.', *Clinical Infectious Diseases*, 58: 697-703.
- Schwaber, M.J., S. Klarfeld-Lidji, S. Navon-Venezia, D. Schwartz, A. Leavitt, and Y. Carmeli. 2008. 'Predictors of carbapenem-resistant *Klebsiella pneumoniae* acquisition among hospitalized adults and effect of acquisition on mortality', *Antimicrobial Agents and Chemotherapy*, 52: 1028-33.
- Sheng, V.S., and C.X. Ling. 2006. 'Thresholding for making classifiers cost-sensitive', *Association for the Advancement of Artificial Intelligence (AAAI)*: 476-81.
- Siegel, J.D., E. Rhinehart, M. Jackson, and L. Chiarello. 2007. "2007 Guideline for isolation precautions preventing transmission of infectious agents in healthcare settings. Last update: May, 2019." In *AJIC: American Journal of Infection Control*, S65-S164.
- Siegel, J.D., E. Rhinehart, M. Jackson, L. Chiarello, and the Healthcare Infection Control Practices Advisory Committee. 2007. 'Management of multidrug-resistant organisms in health care settings, 2006', *American journal of infection control*, 35: S165-S93.
- Song, X., A. Srinivasan, D. Plaut, and T.M. Perl. 2003. 'Effect of nosocomial vancomycin-resistant enterococcal bacteremia on mortality, length of stay, and costs', *Infection Control & Hospital Epidemiology*, 24: 251-56.
- Spellberg, B., R. Gidos, D. Gilbert, J. Bradley, H.W. Boucher, W.M. Scheld, J.G. Bartlett, J. Edwards Jr, and Infectious Diseases Society of America. 2008. 'The epidemic of antibiotic-resistant infections: a call to action for the medical

It is made available under a [CC-BY-NC-ND 4.0 International license](#) .

A Data-Driven Framework for Identifying Intensive Care Unit Admissions Colonized with Multidrug-Resistant Organisms

- community from the Infectious Diseases Society of America', *Clinical infectious diseases*, 46: 155-64.
- Strich, J.R., and T.N. Palmore. 2017. 'Preventing transmission of multidrug-resistant pathogens in the intensive care unit.', *Infectious Disease Clinics*, 31: 535-50.
- Tacconelli, E. 2006. 'New strategies to identify patients harbouring antibiotic-resistant bacteria at hospital admission', *Clinical Microbiology and Infection*, 12: 102-09.
- Tacconelli, E., and M.A. Cataldo. 2008. 'Vancomycin-resistant enterococci (VRE): transmission and control', *International journal of antimicrobial agents*, 31: 99-106.
- Tacconelli, E., A.W. Karchmer, D. Yokoe, and E.M. D'agata. 2004. 'Preventing the influx of vancomycin-resistant enterococci into health care institutions, by use of a simple validated prediction rule', *Clinical infectious diseases*, 39: 964-70.
- Tibshirani, R. 1996. 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society: Series B (Methodological)*, 58: 267-88.
- Tischendorf, J., R.A. de Avila, and N. Safdar. 2016. 'Risk of infection following colonization with carbapenem-resistant Enterobacteriaceae: A systematic review', *American journal of infection control*, 44: 539-43.
- Ueno, T., and N. Masuda. 2008. 'Controlling nosocomial infection based on structure of hospital social networks', *Journal of Theoretical Biology*, , 254: 655-66.
- van Kleef, E., J.V. Robotham, M. Jit, S.R. Deeny, and W.J. Edmunds. 2013. 'Modelling the transmission of healthcare associated infections: a systematic review', *BMC infectious diseases*, 13: 294.
- Vincent, J.L. 2003. 'Nosocomial infections in adult intensive-care units', *The Lancet*, 361: 2068-77.
- Vincent, J.L., J. Rello, J. Marshall, E. Silva, A. Anzueto, C. D. Martin, R. Moreno, J. Lipman, C. Gomersall, Y. Sakr, and K. Reinhart. 2009. 'International study of the prevalence and outcomes of infection in intensive care units', *Jama*, 302: 2323-29.
- Weber, S.G., S.S. Huang, S. Oriola, W.C. Huskins, G.A. Noskin, K. Harriman, R.N. Olmsted, M. Bonten, T. Lundstrom, M.W. Climo, and M.C. Roghmann. 2007. 'Legislative mandates for use of active surveillance cultures to screen for methicillin-resistant Staphylococcus aureus and vancomycin-resistant enterococci: position statement from the Joint SHEA and APIC Task Force', *Infection Control & Hospital Epidemiology*, 28: 249-60.
- Wenzel, R.P., and M.B. Edmond. 2010. 'Infection control: the case for horizontal rather than vertical interventional programs', *International Journal of Infectious Diseases*, 14: S3-S5.
- Wiens, J., and E.S. Shenoy. 2017. 'Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology', *Clinical infectious diseases*, 66: 149-53.
- World Health Organization. 2017. 'Guidelines for the prevention and control of carbapenem-resistant Enterobacteriaceae, Acinetobacter baumannii and

It is made available under a [CC-BY-NC-ND 4.0 International license](https://creativecommons.org/licenses/by-nc-nd/4.0/) .

A Data-Driven Framework for Identifying Intensive Care Unit Admissions Colonized with Multidrug-Resistant Organisms

- Pseudomonas aeruginosa* in health care facilities'. <https://www.who.int/infection-prevention/publications/guidelines-cre/en/>.
- Youden, W.J. 1950. 'Index for rating diagnostic tests', *Cancer*, 3: 32-35.
- Zhanel, G.G., DeCorby, Laing M., N., , B. Weshnoweski, R. Vashisht, F. Tailor, K.A. Nichol, A. Wierzbowski, P.J. Baudry, J.A. Karlowsky, and P. Lagacé-Wiens. 2008. 'Antimicrobial-resistant pathogens in intensive care units in Canada: results of the Canadian National Intensive Care Unit (CAN-ICU) study, 2005-2006', *Antimicrobial Agents and Chemotherapy*, 52: 1430-37.
- Ziakas, P.D., R. Thapa, L.B. Rice, and E. Mylonakis. 2013. 'Trends and significance of VRE colonization in the ICU: a meta-analysis of published studies', *PloS one*, 8: e75658.

It is made available under a [CC-BY-NC-ND 4.0 International license](https://creativecommons.org/licenses/by-nc-nd/4.0/).

A Data-Driven Framework for Identifying Intensive Care Unit Admissions Colonized with Multidrug-Resistant Organisms

Let $X \in \{0,1\}^n$ be the vector of predictions for n data points, $P \in [0,1]^n$ be the predicted probabilities, and $Y \in \{0,1\}^n$ be the vector of true values of the prediction outcome. Let $\tau \in [0,1]$ be a threshold value converting P to X such that $X_i = 0$ if $P_i \leq \tau$ and $X_i = 1$ when $P_i > \tau$ for every data point $i = 1, 2, \dots, n$. Then the threshold optimization problem that identifies the optimal threshold value τ^* for maximizing Youden's index can be formulated as follows:

maximize sensitivity (X, Y) + specificity $(X, Y) - 1$
 $\tau \in [0, 1]$
such that
 $P_i - \tau < X_i \quad i=1, \dots, n$
 $\tau - P_i > X_i \quad i=1, \dots, n$
 $P \in [0, 1]^n \quad X \in \{0, 1\}^n \quad Y \in \{0, 1\}^n$

Figure 1. Threshold Optimization Formulation

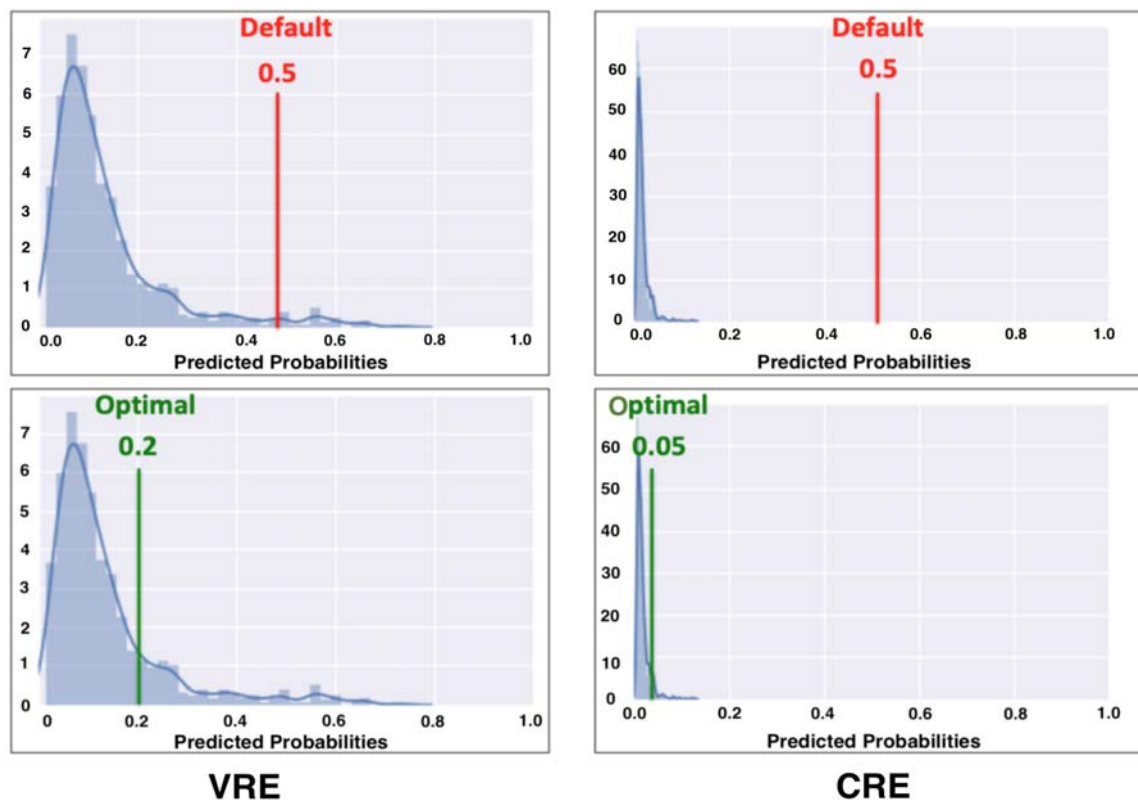


Figure 2. Threshold Value for Converting Predicted Probabilities to Binary Predictions

It is made available under a [CC-BY-NC-ND 4.0 International license](https://creativecommons.org/licenses/by-nc-nd/4.0/).

A Data-Driven Framework for Identifying Intensive Care Unit Admissions Colonized with Multidrug-Resistant Organisms

| VREI503/3860#I3.03% | | | CREI53/3661#I4.5% | | | MRS4I332/4446#I7.47% | | | MDROI792/4508#I7.59% | | |
|---|----------|----------|--|----------|----------|---|----------|----------|------------------------------|----------|----------|
| ThresholdOpt. UpperBound#0.5 | Log.Reg. | Dec.Free | ThresholdOpt. UpperBound#0.040 | Log.Reg. | Dec.Free | ThresholdOpt. UpperBound#0.03 | Log.Reg. | Dec.Free | ThresholdOpt. UpperBound#0.1 | Log.Reg. | Dec.Free |
| TrainingAUC | 0.76 | 0.77 | TrainingAUC | 0.70 | 0.78 | TrainingAUC | 0.65 | 0.66 | TrainingAUC | 0.72 | 0.88 |
| TestingAUC | 0.80 | 0.77 | TestingAUC | 0.78 | 0.71 | TestingAUC | 0.66 | 0.69 | TestingAUC | 0.70 | 0.76 |
| TestingSensitivity | 0.99 | 1.00 | TestingSensitivity | 1.00 | 0.82 | TestingSensitivity | 1.00 | 1.00 | TestingSensitivity | 0.92 | 0.93 |
| TestingSpecificity | 0.09 | 0.33 | TestingSpecificity | 0.31 | 0.37 | TestingSpecificity | 0.02 | 0.00 | TestingSpecificity | 0.30 | 0.51 |
| VREI503/3860#I3.03% | | | CREI53/3661#I4.5% <td colspan="3">MRS4I332/4446#I7.47% <td colspan="3">MDROI792/4508#I7.59% </td></td> | | | MRS4I332/4446#I7.47% <td colspan="3">MDROI792/4508#I7.59% </td> | | | MDROI792/4508#I7.59% | | |
| ThresholdOpt. UpperBound#0.1 | Log.Reg. | Dec.Free | ThresholdOpt. UpperBound#0.015 | Log.Reg. | Dec.Free | ThresholdOpt. UpperBound#0.075 | Log.Reg. | Dec.Free | ThresholdOpt. UpperBound#0.2 | Log.Reg. | Dec.Free |
| TrainingAUC | 0.76 | 0.77 | TrainingAUC | 0.70 | 0.80 | TrainingAUC | 0.65 | 0.67 | TrainingAUC | 0.72 | 0.88 |
| TestingAUC | 0.80 | 0.77 | TestingAUC | 0.78 | 0.71 | TestingAUC | 0.66 | 0.71 | TestingAUC | 0.70 | 0.81 |
| TestingSensitivity | 0.89 | 0.88 | TestingSensitivity | 0.82 | 0.73 | TestingSensitivity | 0.76 | 0.82 | TestingSensitivity | 0.68 | 0.93 |
| TestingSpecificity | 0.49 | 0.57 | TestingSpecificity | 0.57 | 0.51 | TestingSpecificity | 0.45 | 0.45 | TestingSpecificity | 0.59 | 0.58 |
| VREI503/3860#I3.03% <td colspan="3">CREI53/3661#I4.5% <td colspan="3">MRS4I332/4446#I7.47% <td colspan="3">MDROI792/4508#I7.59% </td></td></td> | | | CREI53/3661#I4.5% <td colspan="3">MRS4I332/4446#I7.47% <td colspan="3">MDROI792/4508#I7.59% </td></td> | | | MRS4I332/4446#I7.47% <td colspan="3">MDROI792/4508#I7.59% </td> | | | MDROI792/4508#I7.59% | | |
| ThresholdOpt. UpperBound#0.15 | Log.Reg. | Dec.Free | ThresholdOpt. UpperBound#0.020 | Log.Reg. | Dec.Free | ThresholdOpt. UpperBound#0.1 | Log.Reg. | Dec.Free | ThresholdOpt. UpperBound#0.3 | Log.Reg. | Dec.Free |
| TrainingAUC | 0.76 | 0.77 | TrainingAUC | 0.70 | 0.80 | TrainingAUC | 0.65 | 0.66 | TrainingAUC | 0.72 | 0.76 |
| TestingAUC | 0.80 | 0.77 | TestingAUC | 0.78 | 0.73 | TestingAUC | 0.66 | 0.68 | TestingAUC | 0.70 | 0.81 |
| TestingSensitivity | 0.80 | 0.78 | TestingSensitivity | 0.73 | 0.73 | TestingSensitivity | 0.64 | 0.73 | TestingSensitivity | 0.65 | 0.89 |
| TestingSpecificity | 0.66 | 0.65 | TestingSpecificity | 0.69 | 0.63 | TestingSpecificity | 0.60 | 0.57 | TestingSpecificity | 0.63 | 0.65 |
| VREI503/3860#I3.03% <td colspan="3">CREI53/3661#I4.5% <td colspan="3">MRS4I332/4446#I7.47% <td colspan="3">MDROI792/4508#I7.59% </td></td></td> | | | CREI53/3661#I4.5% <td colspan="3">MRS4I332/4446#I7.47% <td colspan="3">MDROI792/4508#I7.59% </td></td> | | | MRS4I332/4446#I7.47% <td colspan="3">MDROI792/4508#I7.59% </td> | | | MDROI792/4508#I7.59% | | |
| ThresholdOpt. UpperBound#0.2 | Log.Reg. | Dec.Free | ThresholdOpt. UpperBound#0.025 | Log.Reg. | Dec.Free | ThresholdOpt. UpperBound#0.15 | Log.Reg. | Dec.Free | ThresholdOpt. UpperBound#0.4 | Log.Reg. | Dec.Free |
| TrainingAUC | 0.76 | 0.77 | TrainingAUC | 0.70 | 0.79 | TrainingAUC | 0.65 | 0.67 | TrainingAUC | 0.72 | 0.76 |
| TestingAUC | 0.80 | 0.77 | TestingAUC | 0.78 | 0.71 | TestingAUC | 0.66 | 0.69 | TestingAUC | 0.70 | 0.79 |
| TestingSensitivity | 0.66 | 0.63 | TestingSensitivity | 0.73 | 0.64 | TestingSensitivity | 0.48 | 0.71 | TestingSensitivity | 0.57 | 0.79 |
| TestingSpecificity | 0.76 | 0.75 | TestingSpecificity | 0.73 | 0.67 | TestingSpecificity | 0.75 | 0.58 | TestingSpecificity | 0.73 | 0.64 |
| VREI503/3860#I3.03% <td colspan="3">CREI53/3661#I4.5% <td colspan="3">MRS4I332/4446#I7.47% <td colspan="3">MDROI792/4508#I7.59% </td></td></td> | | | CREI53/3661#I4.5% <td colspan="3">MRS4I332/4446#I7.47% <td colspan="3">MDROI792/4508#I7.59% </td></td> | | | MRS4I332/4446#I7.47% <td colspan="3">MDROI792/4508#I7.59% </td> | | | MDROI792/4508#I7.59% | | |
| ThresholdOpt. UpperBound#0.3 | Log.Reg. | Dec.Free | ThresholdOpt. UpperBound#0.030 | Log.Reg. | Dec.Free | ThresholdOpt. UpperBound#0.2 | Log.Reg. | Dec.Free | ThresholdOpt. UpperBound#0.5 | Log.Reg. | Dec.Free |
| TrainingAUC | 0.76 | 0.77 | TrainingAUC | 0.70 | 0.79 | TrainingAUC | 0.65 | 0.66 | TrainingAUC | 0.72 | 0.76 |
| TestingAUC | 0.80 | 0.77 | TestingAUC | 0.78 | 0.71 | TestingAUC | 0.66 | 0.69 | TestingAUC | 0.70 | 0.80 |
| TestingSensitivity | 0.66 | 0.65 | TestingSensitivity | 0.64 | 0.64 | TestingSensitivity | 0.70 | 0.67 | TestingSensitivity | 0.56 | 0.77 |
| TestingSpecificity | 0.78 | 0.74 | TestingSpecificity | 0.78 | 0.74 | TestingSpecificity | 0.55 | 0.59 | TestingSpecificity | 0.75 | 0.70 |
| VREI503/3860#I3.03% <td colspan="3">CREI53/3661#I4.5% <td colspan="3">MRS4I332/4446#I7.47% <td colspan="3">MDROI792/4508#I7.59% </td></td></td> | | | CREI53/3661#I4.5% <td colspan="3">MRS4I332/4446#I7.47% <td colspan="3">MDROI792/4508#I7.59% </td></td> | | | MRS4I332/4446#I7.47% <td colspan="3">MDROI792/4508#I7.59% </td> | | | MDROI792/4508#I7.59% | | |
| ThresholdOpt. UpperBound#0.5 | Log.Reg. | Dec.Free | ThresholdOpt. UpperBound#0.050 | Log.Reg. | Dec.Free | ThresholdOpt. UpperBound#0.3 | Log.Reg. | Dec.Free | ThresholdOpt. UpperBound#0.6 | Log.Reg. | Dec.Free |
| TrainingAUC | 0.76 | 0.77 | TrainingAUC | 0.70 | 0.79 | TrainingAUC | 0.65 | 0.66 | TrainingAUC | 0.72 | 0.77 |
| TestingAUC | 0.80 | 0.77 | TestingAUC | 0.78 | 0.72 | TestingAUC | 0.66 | 0.70 | TestingAUC | 0.70 | 0.80 |
| TestingSensitivity | 0.61 | 0.59 | TestingSensitivity | 0.55 | 0.64 | TestingSensitivity | 0.48 | 0.76 | TestingSensitivity | 0.58 | 0.89 |
| TestingSpecificity | 0.82 | 0.77 | TestingSpecificity | 0.82 | 0.79 | TestingSpecificity | 0.75 | 0.59 | TestingSpecificity | 0.72 | 0.62 |

Table 1. Performance Summary for VRE, CRE, MRSA, and MDRO Colonization

Note: In each sub-table, we summarize the out-of-sample AUC for the training set, along with the out-of-sample AUC, sensitivity, and specificity for the test set. For each outcome, we include results for different upper bounds for the threshold optimization.

It is made available under a [CC-BY-NC-ND 4.0 International license](https://creativecommons.org/licenses/by-nc-nd/4.0/).

A Data-Driven Framework for Identifying Intensive Care Unit Admissions Colonized with Multidrug-Resistant Organisms

| Models with the Best Youden Index | VRE (503/3860 (13.03%)) | | | Models with the Best Youden Index | MRSA (332/4446 (7.47%)) | | |
|-----------------------------------|-------------------------|---------|--------------|-----------------------------------|-------------------------|---------|--------------|
| | Log. Reg. | XGBoost | Rand. Forest | | Log. Reg. | XGBoost | Rand. Forest |
| Training AUC | 0.76 | 0.77 | 0.77 | Training AUC | 0.65 | 0.66 | 0.66 |
| Testing AUC | 0.80 | 0.77 | 0.77 | Testing AUC | 0.66 | 0.66 | 0.70 |
| Testing Sensitivity | 0.80 | 0.73 | 0.75 | Testing Sensitivity | 0.70 | 0.67 | 0.76 |
| Testing Specificity | 0.66 | 0.65 | 0.66 | Testing Specificity | 0.55 | 0.60 | 0.59 |
| Youden Index | 0.46 | 0.39 | 0.41 | Youden Index | 0.24 | 0.27 | 0.34 |
| Threshold Opt. Bound | 0.15 | 0.15 | 0.20 | Threshold Opt. Bound | 0.20 | 0.10 | 0.30 |

| Models with the Best Youden Index | CRE (53/3661 (1.45%)) | | | Models with the Best Youden Index | MDRO (792/4503 (17.59%)) | | | |
|-----------------------------------|-----------------------|---------|--------------|-----------------------------------|--------------------------|---------|--------------|-----------|
| | Log. Reg. | XGBoost | Rand. Forest | | Log. Reg. | XGBoost | Rand. Forest | Dec. Tree |
| Training AUC | 0.70 | 0.76 | 0.79 | Training AUC | 0.72 | 0.86 | 0.87 | 0.76 |
| Testing AUC | 0.78 | 0.72 | 0.72 | Testing AUC | 0.70 | 0.87 | 0.89 | 0.81 |
| Testing Sensitivity | 0.73 | 0.73 | 0.64 | Testing Sensitivity | 0.56 | 0.75 | 0.82 | 0.89 |
| Testing Specificity | 0.73 | 0.77 | 0.79 | Testing Specificity | 0.75 | 0.82 | 0.83 | 0.65 |
| Youden Index | 0.45 | 0.50 | 0.42 | Youden Index | 0.30 | 0.57 | 0.65 | 0.54 |
| Threshold Opt. Bound | 0.025 | 0.015 | 0.05 | Threshold Opt. Bound | 0.50 | 0.30 | 0.40 | 0.30 |

Table 2. Performance Summary of the Models with the Highest Youden’s Index

| VRE Colonization upon ICU Admission | | Relative Ranking | | |
|-------------------------------------|-----------------------------|------------------|---------|--------------|
| Factors | Features | Log. Reg. | XGBoost | Rand. Forest |
| Long-term Care Facility Stay | Yes | 1 | 3 | 1 |
| Recent 1-Digit CD10 Procedure | Other Procedures | 2 | 1 | 2 |
| Current Diagnosis CCS Class | Skin and Subcutaneous Issue | 3 | 2 | 3 |
| Recent 2-Digit CD10 Procedure | Medical/Surgical Anatomical | 6 | 5 | 8 |
| Recent 2-Digit CD10 Procedure | Administration Circulatory | 8 | 4 | 6 |

Top Five Common Predictors for VRE Colonization upon ICU Admission

| CRE Colonization upon ICU Admission | | Relative Ranking | | |
|-------------------------------------|-----------------------------|------------------|---------|--------------|
| Factors | Features | Log. Reg. | XGBoost | Rand. Forest |
| Current Diagnosis CCS Class | Skin and Subcutaneous Issue | 2 | 2 | 1 |
| Recent 1-Digit CD10 Procedure | Other Procedures | 3 | 3 | 2 |
| Prior ICU Stay | > 20 Days | 4 | 6 | 5 |
| Long-term Care Facility Stay | Yes | 5 | 6 | 6 |
| Number of Current Diagnosis PoA | > 30 and ≤ 50 | 6 | 8 | 3 |

Top Five Common Predictors for CRE Colonization upon ICU Admission

| MRSA Colonization upon ICU Admission | | Relative Ranking | | |
|--------------------------------------|-----------------------------|------------------|---------|--------------|
| Factors | Features | Log. Reg. | XGBoost | Rand. Forest |
| Recent 1-Digit CD10 Procedure | Other Procedures | 1 | 2 | 1 |
| Current Diagnosis CCS Class | Skin and Subcutaneous Issue | 2 | 9 | 2 |
| Current Diagnosis CCS Class | Injury and Poisoning | 7 | 1 | 8 |
| Current Diagnosis CCS Class | Infectious and Parasitic | 9 | 8 | 5 |
| Recent 1-Digit CD10 Procedure | Administration | -3 | 3 | 14 |

Top Five Common Predictors for MRSA Colonization upon ICU Admission

| MDRO Colonization upon ICU Admission | | Relative Ranking | | | |
|--------------------------------------|-----------------------------|------------------|---------|--------------|-----------|
| Factors | Features | Log. Reg. | XGBoost | Rand. Forest | Dec. Tree |
| Recent 1-Digit CD10 Procedure | Other Procedures | 7 | 1 | 1 | 2 |
| Current Diagnosis CCS Class | Skin and Subcutaneous Issue | 16 | 2 | 2 | 14 |
| Current Diagnosis CCS Class | Mental Illness | 35 | 6 | 3 | 12 |
| Current Diagnosis CCS Class | Infectious and Parasitic | 57 | 12 | 4 | 16 |
| Sex | Female | 89 | 3 | 5 | 9 |

Top Five Common Predictors for MDRO Colonization upon ICU Admission

Table 3. Top Five Common Predictors for VRE, CRE, MRSA, and MDRO Colonization
Note: The “-” sign indicates ranking among the factors that decrease the risk (OR < 1).

It is made available under a [CC-BY-NC-ND 4.0 International license](#) .

A Data-Driven Framework for Identifying Intensive Care Unit Admissions Colonized with Multidrug-Resistant Organisms

Images: Figures and Tables

Figure 1. Threshold Optimization Formulation

Figure 2. Threshold Value for Converting Predicted Probabilities to Binary Predictions

Table 1. Performance Summary for VRE, CRE, MRSA, and MDRO Colonization

Table 2. Performance Summary of the Models with the Highest Youden's Index

Table 3. Top Five Common Predictors for VRE, CRE, MRSA, and MDRO Colonization