# Early Prediction of MRSA Infections using Electronic Health Records

Thomas Hartvigsen[1], Cansu Sen[1], Sarah Brownell[2], Erin Teeple[1],
Xiangnan Kong[1] and Elke Rundensteiner[1]
[1]*Worcester Polytechnic Institute, 100 Institute Rd., Worcester, MA, U.S.A.*
[2]*Simmons College, 300 Fenway, Boston, MA, U.S.A.*

Keywords: MRSA, Healthcare-associated Infections, Risk Stratification, Machine Learning, Electronic Health Records.

Abstract: Despite eradication efforts, Methicillin-resistant Staphylococcus aureus (MRSA) remains a common cause of serious hospital-acquired infections (HAI) in the United States. Electronic Health Record (EHR) systems capture MRSA infection events along with detailed patient information preceding diagnosis. In this work, we design and apply machine learning methods to support early recognition of MRSA infection by estimating risk at several time points during hospitalization. We use EHR data including on-admission and throughout-stay patient information. On-admission features capture clinical and non-clinical information while throughout-stay features include vital signs, medications, laboratory studies, and other clinical assessments. We evaluate prediction accuracy achieved by core Machine Learning methods, namely Logistic Regression, Support Vector Machine, and Random Forest classifiers, when mining these different types of EHR features to detect patterns predictive of MRSA infection. We evaluate classification performance using MIMIC III – a critical care data set comprised of 12 years of patient records from the Beth Israel Deaconess Medical Center Intensive Care Unit in Boston, MA. Our methods can achieve near-perfect MRSA prediction accuracies one day before documented clinical diagnosis. Also, they perform well for early MRSA prediction many days in advance of diagnosis. These findings underscore the potential clinical applicability of machine learning techniques.



## 1 INTRODUCTION

### 1.1 Antibiotic Resistance and MRSA

The antibiotic resistance crisis presents a formidable global health threat for the 21st century. The discovery of antibiotics to treat bacterial infections transformed medicine and saved millions of lives (Ventola, 2015). Over time, however, the use of antibiotics has resulted in the selection and spread of antibiotic-resistant strains. Infections caused by organisms resistant to traditional antibiotics are more difficult to treat and may require the use of more expensive and potentially more toxic alternative therapies, if any are available (Neu, 1992).

Staphylococcus aureus is one of the most common causes of Hospital-Acquired Infections (HAIs), accounting for an estimated 12% of HAIs between 2011-2014 and causing over 80,000 infections in the United States in 2011 alone (Weiner et al., 2016; Dantes et al., 2013). Methicillin-resistant Staphylococcus aureus (MRSA) is one antibiotic-resistant strain of this bacteria. MRSA infections may result in serious complications including sepsis and death. Unfortunately, hospitals are known to be high-risk zones for spread of MRSA because contamination may go undetected. Also, many hospitalized patients are at increased risk of infection (Maree et al., 2007).

### 1.2 Leveraging EHR Systems for MRSA Infection Prediction

The construction of intelligent infection prediction systems using machine learning presents one important opportunity for confronting the challenges of antibiotic resistance and the spread of infections such as MRSA in healthcare environments (Sintchenko et al., 2008). Infection prediction systems have shown to be successfully identify early signals for other infections, such as *Clostridium difficile* (Sen et al., 2017). An overview of such an approach is depicted in Figure 1. Before caregivers recognize or test for MRSA, machine learning algorithms have the potential to identify likely MRSA cases in advance based on patterns learned from the medical information of previous cases. Such early detection would facilitate **(1)** early



Hartvigsen, T., Sen, C., Brownell, S., Teeple, E., Kong, X. and Rundensteiner, E.
Early Prediction of MRSA Infections using Electronic Health Records.
DOI: 10.5220/0006599601560167
In *Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2018) - Volume 5: HEALTHINF*, pages 156-167
ISBN: 978-989-758-281-3

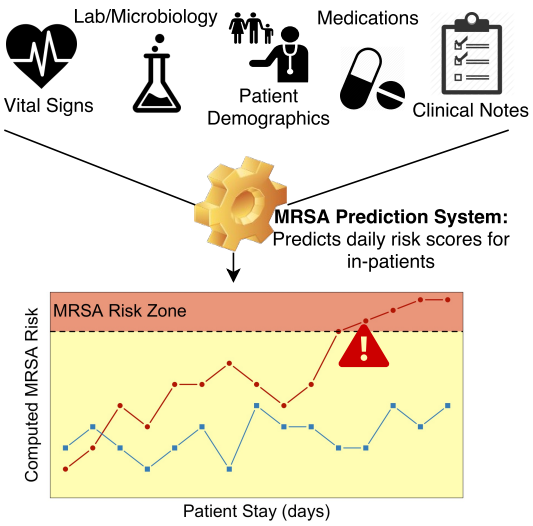


Figure 1: Overview of Intelligent Infection Prediction.

isolation to reduce spread of resistant strains within the healthcare facility; **(2)** judicious, more precisely targeted antibiotic usage; and **(3)** earlier initiation of optimal treatments to improve patient outcomes. Using insights created through intelligent prediction systems, healthcare professionals could decide to initiate precautionary measures or alter treatment plans. For example, at the point of admission, high-risk patients could benefit from *contact precautions*, such as gloving, gowning, and/or environment alterations, or *patient placement precautions*, such as patient assignment decisions based on risk factors (Shang et al., 2000). Machine learning-guided patient assessment thus offers an enhanced range of intervention points to arrest the spread of MRSA.

Electronic Health Records (EHR) systems have been universally adopted by medical facilities across the United States as a result of the Health Information Technology for Economic and Clinical Health (HITECH) Act (Congress, 2009) and the Centers for Medicare and Medicaid EHR Incentive Programs (CMS, 2011). To date, however, data accumulated within these systems has been largely underutilized for predictive analytics (Celi et al., 2013). The widespread digitalization of health records presents a unique opportunity for healthcare innovation (Jensen et al., 2012; Raghupathi and Raghupathi, 2014). It is evident that there are signals embedded in these complex patient data that could indicate the likelihood of an evolving infection or other medically important conditions (Jensen et al., 2012; Raghupathi and Raghupathi, 2014). Thus, in this work we focus on one such application, namely the prediction of hospital-acquired MRSA infections using machine learning methods.

## 1.3 Methods for MRSA Prediction

Previous research efforts have begun to explore the application of intelligent systems to EHR data for HAI prediction. In one such investigation, EHR data was analyzed to generate predictions for HAI occurrence without specific determination of infection type (Chang et al., 2011). This method used only 16 patient characteristics recorded at the beginning of the hospital stay and classified patients using Logistic Regression and Artificial Neural Networks. Even using this limited set of patient variables, the authors reported high predictive accuracy. In another study (Nseir et al., 2010), admission information was used to identify multi-drug resistant bacterial infections using Logistic Regression methods. For MRSA infection prediction, (Dutta and Dutta, 2006) used Bayes' Theorem and a Maximum Probability Rule to predict MRSA cases from medical-sensor data with up to 99.83% accuracy. However, these authors did not focus on early detection. Instead, they used data available right up to microbiological MRSA confirmation. Last-minute detection has limited clinical value. Another study by (Shang et al., 2000) used EHR data collected at the time of admission to diagnose community-acquired MRSA using Logistic Regression and simple Artificial Neural Networks, but this approach did not incorporate information obtained throughout the hospital stay.

## 1.4 Scope of this Work

In this investigation, we use Logistic Regression for consistency with earlier diagnostic prediction studies (Chang et al., 2011; Shang et al., 2000), and we also include Support Vector Machines and Random Forests due to their previous use in detecting other infections (Lebedev et al., 2014; Khalilia et al., 2011; Wiens et al., 2012; Wu et al., 2010). All of these algorithms are also known to be easily interpretable. In contrast to earlier work, however, we focus on developing clinically translatable models. Previous studies have either used data from immediately before diagnosis to achieve high predictive accuracy or used only data collected at admission, which prevents the identification of conditions that evolve during hospitalization. Our objective in this investigation is to balance early prediction with high accuracy to achieve clinically translatable MRSA detection methods.

We evaluate our techniques using the MIMIC III database, a publicly-available critical-care data set collected over 12 years from the Beth Israel Deaconness Hospital Intensive Care Unit in Boston, MA (Johnson et al., 2016). Our findings confirm that machine learning is a highly effective technology for

early prediction using EHR data. Core machine learning methods are shown to effectively identify high-risk MRSA patients. We report AUC scores of over 0.97 one day before diagnosis and 0.93 to 0.96 using only the first day of EHR data for each patient. These findings underscore the potential for machine learning techniques to generate early warnings of infections.

# 2 METHODOLOGY

## 2.1 Objectives

There are many steps and possible options when extracting features from EHR databases. This is complicated further when considering predictions for specific infections. In this work, data processing and classification decisions were structured to answer the following questions:

- Whether or not known risk factors alone can generate accurate predictions
- How many days should be used to make predictions for each patient
- How early can accurate predictions be generated

## 2.2 The Dataset

The Medical Information Mart for Intensive Care III (MIMIC III) is a publicly available critical care database collected from the Beth Israel Deaconess Medical Center Intensive Care Unit (ICU) between 2001 and 2012 (Johnson et al., 2016). It contains 58,000 admissions comprised of:

- Billing: Coded data recorded for billing and administrative purposes (CPT, DRG, ICD codes).
- Descriptive: Demographic detail, admission and discharge times, and dates of death.
- Interventions: Procedures such as dialysis.
- Laboratory: Blood chemistry, hematology, urine analysis, and microbiology test results.
- Medications: Administration records of intravenous medications and medication orders.
- Notes: Free-form text notes such as provider progress notes and hospital discharge summaries.
- Physiologic: Nurse-verified vital signs such as heart rate and blood pressure.

Contained within these items are all known risk factors for MRSA infections. We display these features and their availability in Table 2 (Aureden et al., 2010).

To identify MRSA patients, we extract the microbiology test associated with the organism **80293** (MRSA), found in the *Microbiology Events* table. We use the microbiology test, as opposed to the ICD9 code, to extract the time of diagnosis. The presence of this test in a patient's record indicates a positive result. Therefore we extract all 1,304 patients who have a record of this test as our MRSA-positive population. As the vast majority of MIMIC consists of patients who do not contract MRSA, the dataset is imbalanced. To handle this, as we experiment with different subsets of MRSA-positive patients, we randomly subsample 1,304 patients who have no record of a test for organism 80293, obtaining equally-sized groups of positive and negative examples.

## 2.3 Feature Engineering

### 2.3.1 On-admission Features

Certain patient information is known at the time of admission and does not change during a patient's stay. We refer to this as *on-admission*, or *static*, data. The only known on-admission risk factor accessible in the MIMIC III database is *age*. We extract a set of features from the on-admission data and classify them into two groups:

- ***Demographic features*** are immutable patient features. These include age, gender, ethnicity, marital status, and religion.
- ***Stay-specific features*** describe a patient's admission such as admission location, allowing inference on the patient's condition. Stay-specific data could be different for the same patient upon readmission. We extracted 3 such features: admission type (e.g., Emergency), admission location (e.g., Transfer from another hospital), and insurance (e.g., Medicaid).

We extracted a total of 9 on-admission features and display the 4 that best contrast the MRSA-positive and MRSA-negative patients in Table 1.

### 2.3.2 Throughout-stay Features

Throughout the hospital stay of a patient, observations such as laboratory results and vital signs are recorded continuously. This results in *throughout-stay* data. Additionally, for each day of a patient's stay, we generate multiple binary features flagging the use of certain types of medication groups, such as antibiotics. We extracted 80 throughout-stay features, as summarized in Table 3.

One challenge is that each patient's stay is recorded as a series of clinical observations that tend

Table 1: Distributions of on-admission features for MRSA and non-MRSA patients (in percent) in the database. We only display variables that are notably different between these two patient sets.

| **Variables** | **MRSA (%)** | **non-MRSA (%)** |
|---|---|---|
| Gender | Male: 57.5<br>Female: 42.5 | Male: 55.1<br>Female: 44.9 |
| Insurance | Medicare: 68.7<br>Private: 20.1<br>Medicaid: 9.7<br>Other: 1.5 | Medicare: 52.8<br>Private: 35.1<br>Medicaid: 9.3<br>Other; 2.8 |
| Admiss. Type | Emergency: 92.3<br>Elective: 6.8<br>Newborn: 0<br>Urgent: 0.8 | Emergency: 74<br>Elective: 14.4<br>Newborn: 8.6<br>Urgent: 3.0 |
| Age (av $\pm$ std) | $68.5 \pm 16.6$ | $59.0 \pm 24.2$ |

Table 2: **Known Risk Factors** for MRSA (Fukuta et al., 2012). **Available** column indicates if we can extract this information, and **source** column indicates the table in MIMIC.

| **Risk Factors** | **Available** | **Source** |
|---|---|---|
| Old Age | Yes | Admission |
| Nursing Home Residence | Unknown | Unknown |
| Receipt of Transfusion | Yes | Services |
| Placing of Central Line | Yes | Chart |
| Respiratory Failure | Yes | Chart |
| Open Wounds | Unknown | Unknown |
| Severe Bacteremia | Yes | Lab Tests |
| Organ Impairment | Yes | Services |
| Other health conditions | Yes | Services |
| Previous Hospital Stay | Yes | Admission |
| Treatment with antibiotics | Yes | Medications |

to be *irregularly spaced*. The frequency at which these measurements are taken varies between patients (e.g. once a day vs. multiple times a day). This variation is a function of **(1)** the observation (lab tests may be taken only once a day while vital signs may be measured multiple times a day); **(2)** each patient's condition (more severely ill patients must be monitored more closely); and **(3)** the time of the day (nurses are less likely to wake up patients in the middle of the night).

To make these data comparable across patients, cleaning and aggregation are required. Here, we roll up all observations taken more than once a day into evenly sampled averages, resulting in one value per day. If there are no measurements for a day, they are considered missing values. To handle these empty spaces, we compute the median value for each variable and use it to fill in missing values.

A second challenge is that the total number of observations recorded per patient is not only a function of the frequency of observation, but also the length of the patient's stay. After the above described

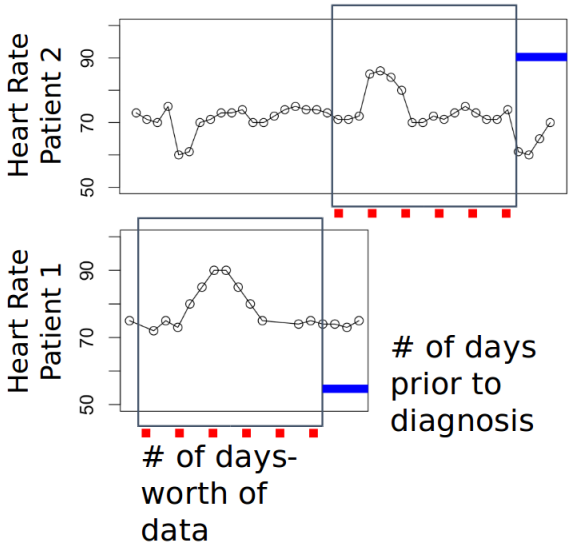


Figure 2: Time alignment strategy. Number of days-worth of data indicates the feature extraction window and the number of days prior to diagnosis indicates the ending-position of the window relative to actual diagnosis.

day-based aggregation, we apply a rectangularization strategy to extract the same number of days for each patient. This is achieved by time-aligning the variable-length feature vectors such that the first days for all patients are lined up with one another.

In our work, different pieces of each patient's records are extracted based on the experimental design. For example, we might extract the first day's worth of throughout-stay data, the first and second days worth of throughout-stay data, and so on. The data can then be directly compared since it has been time-aligned.

Next, we define the feature extraction window for patients. For MRSA patients, it starts on the day of admission and ends $n$ days before the MRSA diagnosis, $n \in \{1, \ldots, 7\}$. For MRSA-negative patients, there are a few alternatives for defining this window. Prior research has used the discharge day as the end of the risk period (Dubberke et al., 2011). However, as the state of a patient can be expected to either improve or decrease drastically near the discharge date, this may lead to deceptive results (Wiens et al., 2012). Instead, for the MRSA-negative patients, we use the halfway point of each stay as the end of the risk period or the minimum possible stay based on the experimental setup, whichever is greater.

Figure 2 shows the feature extraction window and ending position of it for two different patients. The optimal size of the feature extraction window and the number of days prior to prediction are empirically determined for each experiment.

Table 3: Throughout-stay feature descriptions.

| Feature | Explanation | Example |
|---|---|---|
| Lab tests | Daily average results of 20 lab tests | White Blood Cell count, Potassium level |
| Vital signs | Daily average results of 24 vital signs | Heart Rate, Temperature |
| Services | Categorical feature showing patient is on which service that day | Cardiac Medical, Cardiac Surgery |
| Microbiology tests | Daily average results of 13 microbiology tests | Enterococcus Sp., Yeast |
| High-risk antibiotics | Daily binary indicator of high-risk antibiotic prescription | Cephalosporins, Fluoroquinolones |
| Antibiotics | Daily binary indicator of antibiotic prescription | Capreomycin |
| H2 antagonists | Daily binary indicator of H2 antagonist prescription | Nizatidine |
| Proton pump inhibitors | Daily binary indicator proton pump inhibitor prescription | Rabeprazole |

#### 2.3.3 Label Generation

In supervised machine learning, each data object must have an associated label, indicating the outcome. In this work, the outcome is a binary flag indicating the diagnosis of MRSA (1 for MRSA, 0 for non-MRSA), stored in a vector with one label per patient.

## 2.4 Classification

In classification tasks, the goal is to divide data points into predefined classes. For this work, there are two distinct classes, *MRSA-positive*, labeled as 1, and *MRSA-negative*, labeled as 0. This creates a binary classification task, where we attempt to learn the relationship between each patient's historical EHR and their associated label. Predicted labels were generated using three different machine learning methods: Logistic Regression, Support Vector Machines, and Random Forests. We evaluate these methods using a popular holdout strategy: the algorithms were trained on 80% of the patient records and tested on the remaining 20% of the patient records to ensure that the selected models generalize to unseen patients effectively. Performance estimation and hyper-parameter selection were embedded in 5 cross validation folds across the training set. The raw predictions generated lie between 0 and 1, requiring transformation into exactly 0's or 1's to be directly comparable to the binary label vector. In this work, if a prediction is $\geq 0.5$, then it is converted to 1. Otherwise it is converted to 0.

#### 2.4.1 L2-Regularized Logistic Regression

Logistic Regression is a classic machine learning method based on the odds ratio of how the change in individual features affects the outcome. This algorithm is commonly used for diagnosis prediction (Chang et al., 2011; Shang et al., 2000; Visser et al., 2002; Wu et al., 2010). In our setting, each input is a vector, $x$, containing a patient's historical information which will in turn be weighted by $\theta$, a vector of coefficients, as shown in Equation 1, where $n$ is the number of patients and p is the number of variables. We also use *L2-Regularization*, controlled by parameter $\lambda$ to normalize the values of $\theta$, ensuring direct comparisons between the variable weights. Finally, the difference between the predictions made and the true label vector $y$ is minimized.

$$F(\mathbf{x}) = \frac{1}{1+e^{-\theta^\top \mathbf{x}}} + \lambda \sum_{i=1}^{p} \theta_i^2 \quad (1)$$

The task is to learn the proper coefficients that project positively labeled data close to 1, and negatively labeled data close to 0. In this setting, the value predicted for a patient can be considered their probability of MRSA infection.

#### 2.4.2 Soft-margin Support Vector Machine Classification

Support Vector Machine Classification is another popular solution to binary classification problems, also commonly used for diagnosis (Wiens et al., 2012; Wu et al., 2010). In contrast to Logistic Regression, this algorithm makes classifications based on distances between data instances. In this case, we compute the coefficients for a hyperplane that divides the dataset based on the labels and the distance from the hyperplane to a few select data instances, termed *support vectors*. To accomplish this task, we again tune the elements of a vector $\theta$, which will subsequently be multiplied by the patient vector, $\mathbf{x}$, to divide the data by label. This is accomplished by minimizing Eq. 2, where $n$ is the number of support vec-

tors, $\mathbf{x}_i$ is each support vector in turn, $y_i$ is the corresponding label for each support vector, $\lambda$ is a regularizing parameter, and $b$ is a bias variable. The linear kernel was used for all SVM experiments in this work.

$$\frac{1}{n}\sum_{i=1}^{n}\max\left[0,\ 1-y_i\left(\theta^\top \mathbf{x}_i - b\right)\right] + \lambda\|\theta\|_2^2 \quad (2)$$

#### 2.4.3 Random Forests

Random Forests are the *bootstrap aggregating* implementation of decision trees, a well known and interpretable classification algorithm (Breiman, 2001). They have been shown to be effective in predicting infections (Lebedev et al., 2014; Khalilia et al., 2011) in many domains while allowing easy access to relative variable importances. To generate classifications, random subsets of both data instances and variables are iteratively used to generate decision trees and make predictions on a training set. Then, once a set of decision trees has been generated, testing instances are input into each decision tree and the predictions from each tree are recorded. Finally, the predictions made by each decision tree are combined into one prediction, typically via majority voting. This ensemble learning technique emphasizes high levels of randomness, aiding the generalizability of our models.

### 2.5 Evaluation Criteria

The *Receiver Operating Characteristic (ROC) Curve* quantifies the performance of a binary classifier using the *True Positive Rate* (Equation 3) and *False Positive Rate* (Equation 4).

$$\text{True Positive Rate (TPR)} = \frac{\text{TP}}{\text{TP + FN}} \quad (3)$$

$$\text{False Positive Rate (FPR)} = \frac{\text{FP}}{\text{TN + FP}} \quad (4)$$

When a binary classifier makes a probabilistic prediction between 0 and 1, a decision criterion (a.k.a. probability cutoff) decides which probabilities to assign to which class. For example, setting cutoff $= 0.5$ means that $class = 1$ if probability $> 0.5$ while for smaller probability values, $class = 0$. Based on the decision criterion, a binary prediction can be made and TPR (sensitivity) and FPR (1-specificity) can be calculated. When evaluating the performance of a clinical test, sensitivity quantifies the ability of a test to correctly identify cases and specificity reflects the ability of a test to correctly rule out the condition of interest. An ROC curve is used to examine how TPR and FPR change as the decision criterion varies from 0 to 1. The sum of the Area Under the Curve (AUC) quantifies the ability of a classifier to distinguish between two classes. An AUC score of 0.5 indicates a randomly-guessing classifier, and an AUC score of 1.0 indicates perfect classification.

AUC is widely used in clinical diagnosis prediction and risk stratification tasks due to several advantages it brings (Hajian-Tilaki, 2013; Wiens et al., 2012). First, it quantifies the success of a classifier independent of a decision criterion. Second, sensitivity and specificity can be easily considered together by examining the curve. Finally, for risk prediction tasks, the optimal cut-off value can be determined using ROC curve analysis to determine at-risk patients.

### 2.6 Software and Availability

All preprocessing and machine learning are implemented in Python 3.5. Specifically, Pandas 0.18 and Numpy 1.13 are used for preprocessing, Scikit-Learn 0.18 is used to train machine learning algorithms and Matplotlib 1.5 is used for visualizations. PostgreSQL 9.5 is used for data storage and extraction. The scripts used in this work are available at *https://github.com/wpi-dsrg/MRSA-prediction-healthinf*.

## 3 RESULTS

### 3.1 On-admission Stratification

To evaluate how successfully we can predict likely MRSA-positive patients at the time of admission, we train a set of models based only on admission-time data. We consider two training paradigms employing different feature sets: (1) *Risk-Factor* Models and (2) *Data-driven* Models. Risk-factor models use only known risk factors as their input, whereas data-driven models use all extracted *on-admission features* as discussed in Section 2.3.1. By considering both of these settings, we study the predictive power contained in only known MRSA risk factors and how it compares to the complete set of on-admission features.

Current clinical practice emphasizes assessment of MRSA risk factors and observation of signs of infection. From the on-admission data, the only known risk factor as per CDC that we can capture in the EHR data set is *age*. To understand its effect on MRSA diagnoses, we build classifiers using only *age* and compare them to classifiers built using all on-admission features, including *age*. To train these classifiers, we use all 1,308 MRSA-positive patients and randomly sample 1,308 MRSA-negative patients. We then split these 2,616 patients into 80% training (2,093 patients)

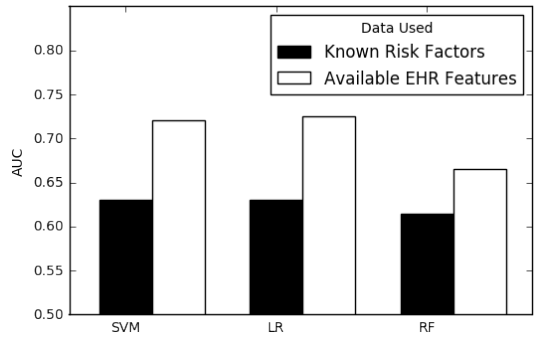


Figure 3: Predictions using On-admission data only comparing known risk factors versus all available on-admission data. Black bars indicate predictions made using only on-admission risk factors (*Age*), white bars use all on-admission data. Models were tuned on 2,093 training patients and tested on 523 unseen testing patients.

and 20% testing (523 patients). We train the algorithms on the 2,093 training patients. Finally, we generate predictions for the 523 unseen patients to understand how well our machine learning models generalize.

We show in Figure 3 that using the only on-admission risk factor *age* leads to non-random predictions (AUC $>$ 0.5) with AUC scores over 0.7 using Logistic Regression and SVM. However, they are significantly less accurate than predictions made using all on-admission features. Support Vector Classifiers and Logistic Regression also outperform Random Forests when considering all on-admission features. We conclude that while *age* contains significant predictive power, combining all features leads to our best admission time MRSA predictions. The AUC scores achieved here will also serve as a baseline for our next experiments, as using only these features is the minimum amount of information to base predictions upon.

## 3.2 Throughout-stay Stratification

Throughout each patient's stay, data are recorded that quantify a patient's condition. This may in turn relate to the risk of acquiring MRSA. To capture the predictability of MRSA based on the current patient state, we trained machine learning models with *throughout-stay* features collected throughout each patient's hospital stay.

To this end, we first define a *baseline patient set*. This dataset consists of patients who have at least 5 days-worth of data to have a significant amount of throughout-stay features, while not excluding many patients. In the MIMIC III database, 998 of the total 1,308 MRSA-positive patients qualify. We randomly subsample 998 MRSA-negative patients, creating a balanced dataset of 1996 patients. Finally this dataset is shuffled and split again into 80% training (1596 patients) and 20% testing subsets (400 patients). For the following, we use this *baseline dataset* of 1996 patients and use 5-fold cross validation over the training patients, reporting the AUC scores for each fold. Finally we average these scores over all 5 folds to choose hyperparameters. We then validate the chosen models on the unseen 400 testing patients.

### 3.2.1 Throughout-stay Risk-factor Model

Similar to the Risk-Factor model we train using on-admission data, we also train a Risk-Factor model on throughout-stay data. There are several known risk factors for MRSA, and many of these are found in the data recorded during inpatient hospitalization (Aureden et al., 2010). These risk factors include *receipt of transfusion, placing of a central line, respiratory failure, bacteremia, organ impairment, and antibiotic use* (Table 2). We expect that these provide significant predictive power for MRSA. To evaluate this, we extract all known risk factors. We then train a set of classifiers using only these risk factors as features.

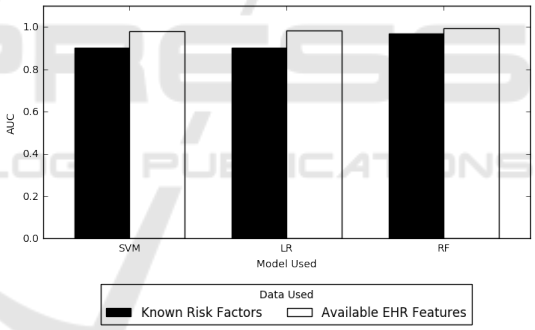


Figure 4: Predictions using throughout-stay data comparing known risk factors only versus all available throughout-stay data. Black bars indicate predictions from known throughout-stay risk factors (Table 2), while white bars indicate predictions using available throughout-stay features. Predictions are made on 400 previously unseen test cases.

Our Risk-Factor model trained on the throughout-stay data achieve an average AUC of 0.94 as shown in Figure 4. This is significantly higher than both the Data-Driven models (mean AUC 0.70) and the Risk-Factor models (mean AUC 0.62) trained on on-admission data. This shows that throughout-stay data is much more telling of a patient's MRSA risk. However, there is more data available in the EHR records.

### 3.2.2 Throughout-stay Feature Groups

As discussed in Section 2.3.2, we propose that non-risk factor features may contain strong predictive

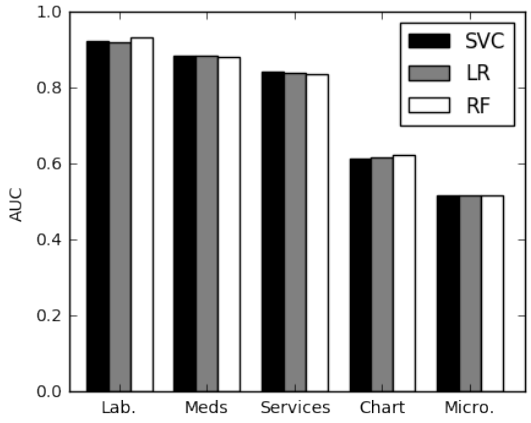


Figure 5: Throughout-stay clinical data subcategory predictions. Dataset used consists of 5 days worth of throughout-stay features for the 1996 patients who stayed over 5 days.

power. The MIMIC III database is categorized by the type of data contained (e.g., Lab and Chart events). These groups may each contain different predictive power. We train predictive models on each group of throughout-stay features to understand their respective predictive powers.

We see that there is a stark contrast between the efficacy of these throughout-stay variable subsets, as shown in Figure 5. The predictions made using the *Laboratory Events, Medication*, or *Services* features are far more accurate than those made using only *Chart Events* or *Microbiology Events*. This indicates that certain throughout-stay features should receive particular attention in clinical practice.

#### 3.2.3 Data-driven Model

After seeing that there is high predictive power contained in non-risk factor data, we train a third set of models using all throughout-stay features for the same *baseline* patient cohort. We then compare the relative effectiveness of only leveraging known risk factors versus all available EHR data.

As seen in Figure 4, embracing a data-driven approach, our models achieve an even higher AUC than the Risk-Factor model on the throughout-stay data. These results indicate that even though there exist well known risk factors for MRSA, machine learning algorithms still benefit from additional data available in EHR systems.

## 3.3 Rectangularization Strategies on Throughout-stay Data

We have shown in Section 3.2.3 that throughout-stay features contain more predictive power than only on-admission features. However, throughout-stay data is not as straightforward to use as on-admission data in terms of feature extraction. Feature extraction from throughout-stay data inherently creates one problem: it requires a tabular representation, i.e., rectangularization of data. Patients staying in the hospital for different number of days create different amount of data, hence different number of features we extract for each patient.

To choose the rectangularization method, we consider two parameters in these experiments: (1) how *far ahead* from diagnosis to attempt to generate predictions and (2) which *days* to use to make these predictions. Varying these *time-slice extraction* parameters has the potential to dramatically alter classifier performance.

#### 3.3.1 Time-slice Extraction

To understand how each of these parameters affects classification accuracy, we first extract 1402 patients who either get MRSA after their 7th day in the hospital, or who are MRSA-negative but stay longer than 7 days. This way we have a large patient set, all of whom have significant amounts of days spend in the hospital. We use these patients for the following experiments. First, we hold the number of days-worth of data extracted constant, and iteratively make predictions using data from earlier and earlier in each patient's stay. This way, we can understand the relationship between accuracy and the time of prediction. Next, we repeat this for different numbers of days-worth of data depending on the number of days available. For instance, if we extract 3 days-worth of data, we make predictions up to 5 days prior to diagnosis.

Two charts from these experiments are shown in Figure 6. These charts depict results of the experiments where we use 1 day worth of data and make predictions for $\{1,\ldots,7\}$ days prior to diagnosis and use 4-days-worth of data and make predictions for $\{1,\ldots,4\}$ days prior to diagnosis, respectively. In both figures, we see a decrease in the AUC's as we achieve predictions farther in advance. As expected, best predictions are made 1 day prior to diagnosis while predictions made on the farthest day were still respectable. However, AUC values remain high even for the farthest predictions. This implies that signals of MRSA are present in the data far in advance of the actual diagnosis. Also, since predictions using 1 day of data were on average lower than those made using 3 days of data, using more data tended to improve the overall AUC values.

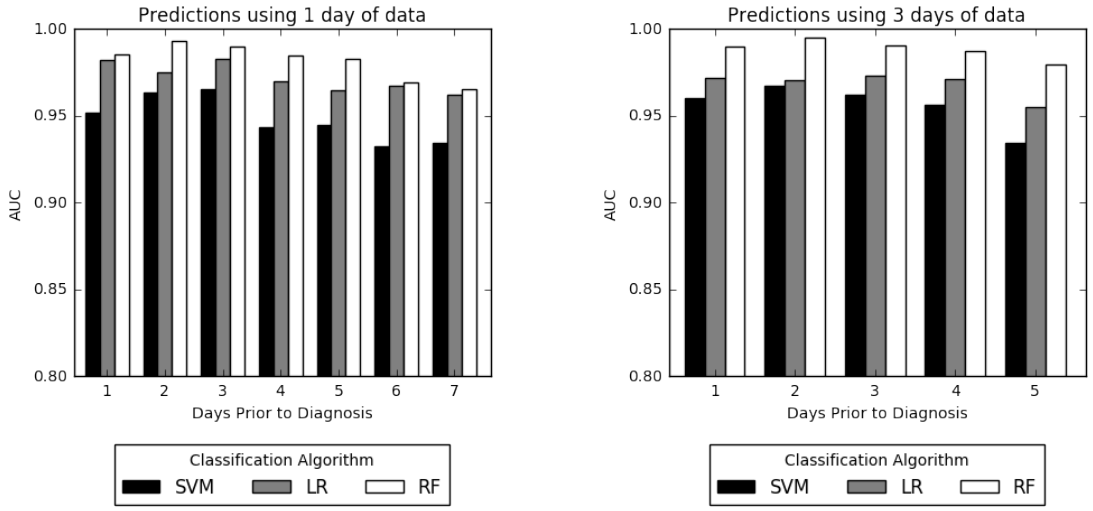


Figure 6: Prediction using days up until diagnosis. We vary the number of days before MRSA diagnosis the risk scores are generated. Experiments shown are for 1402 patients who stayed more than 7 days in the ICU.

### 3.4 Optimizing Rectangularization for Early MRSA Prediction

A recurring challenge in this domain is that a prediction made the day before diagnosis will likely be accurate but not particularly clinically actionable. On the other hand, a prediction made at the time when a patient is admitted will likely be inaccurate, but could potentially lead to the most effective clinical actions. Therefore, we next attempt to optimize our classifiers to strike a balance between these two possibilities by exploring predictive models trained in both situations by considering the optimal prediction strategy: generating risk scores based only on the beginning of each patient's stay. In this setting, if a prediction is made far enough in advance, caregivers can modify their actions to prevent the spread of infections. Here, the task is to uncover how predictive the beginning of patient stays are, and how few days we can use to make adequate predictions. Ideally, using only the first few days worth of data for a patient will lead to an accurate prediction of their likelihood for MRSA.

To train these models, we first extract all patients who stayed 7 or more days in the hospital. Thus, we have many patients (1402) who have significantly long stays. We then generate predictions using only their first day of data and record the AUC scores. We then repeat this process using first 2, 3, and 4 days of these patients' stays. We stopped at 4 days as the goal in these experiments to predict the infection early and there are patients in our cohort who are diagnosed on the 8th day.

The results from these experiments are displayed in Figure 7. We see that with only 1 day's worth of

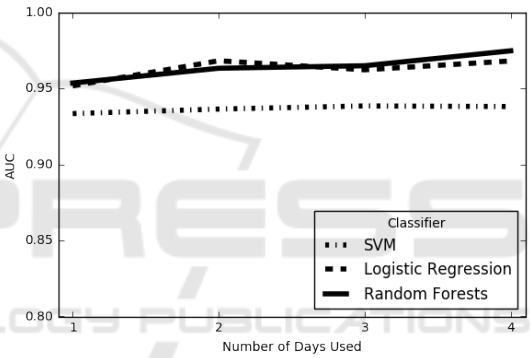


Figure 7: Predictions using increasing number of days starting at the beginning of each patient's stay.

time-variant data added to the on-admission data, we achieve an AUC of over 0.95 with Random Forests, indicating very strong performance from the binary classifiers far in advance of any MRSA diagnoses. As we use more days of data, we achieve a nearly perfect AUC. However, we note that since these classifiers are making predictions for all patients who have 7 days of data, some of the 4-day predictions are predicting only 3 days in advance. These experiments indicate that we can make accurate early predictions for MRSA, which is more actionable than making strong predictions close to the date of diagnosis.

## 4 DISCUSSION

By accurately assessing MRSA infection risk using information recorded long before the time of clinical diagnosis, we have shown that there are strong signals

in EHR data permitting early MRSA infection prediction. The establishment of a MRSA infection risk prediction system using the methods presented in this paper could offer new data-driven insights to inform clinical decisions and enhance patient outcomes. The predictions generated by our method permit healthcare providers to identify patients as either likely or unlikely to develop a MRSA infection at the time of admission and later throughout their hospital stay. While these methods can make an almost perfect prediction the day before diagnosis, it is even more clinically impactful to be able to deliver an accurate prediction many days before diagnosis. In this study, we show that some patients can be reliably classified using only on-admission data, including demographic and stay-specific information (e.g. admission location). For other patients, more data is required, but typically within the first few days of a patient's stay, an accurate risk stratification can be generated, up to an AUC score of over 0.95.

The goal of an intelligent MRSA infection prediction system is to support clinical decision-making and inform infection control efforts by leveraging information contained within EHR data. Currently, most hospitals in the United States have EHR systems in place, with the most popular system being Epic (Jones et al., 2010). In practical use, MRSA risk scores could be communicated to healthcare providers through reporting systems integrated with hospital EHR systems, thus allowing for ongoing system training and easy provider access. Integrating additional information technology with existing EHR systems is now well-recognized as a key strategy for improving patient care while saving costs (Murdoch and Detsky, 2013).

Predictive analysis using data directly retrieved from EHR systems can be integrated into healthcare work processes in several ways. At the individual patient level, warning reports generated when a patient's *MRSA-acquisition risk score* exceeds a set warning threshold can provide useful information for physicians, nurses, and other healthcare providers. Based on clinical context, providers can decide if additional studies or labs are indicated, if isolation precautions should be instituted, or if empirical antibiotic therapy should be started. Alternatively, intelligent systems might also supply information supportive of conservative care choices, such as continued observation. While the use of signals detected within EHR data for patient risk stratification and diagnosis requires further clinical validation, this strategy offers great promise for developing cohesive systems in which retrieval, analysis, and reporting of data would be contained within the EHR software in clinical use. Integration of health records systems with advanced signal detection functionality could then permit not only the recognition of specific medical diagnoses but also the active identification of risk factors and prognostic indicators within facilities and among unique populations.

A limitation of the current study is that our data come from intensive care units in one hospital in the United States. As such, these patients do not equally represent the conditions of general hospitalized populations or the demographics of other regions. The stay-specific data (See Table 1) indicate that while there are diverse groups within the data set, the majority of patients are ethnically white and the gender is predominantly male. In the future, we intend to evaluate the performance of our MRSA risk prediction system using multiple EHR datasets and to ensure generalizability through transfer learning techniques. We also plan to expand these models to predict multiple HAIs concurrently to better serve current hospital needs. An all-encompassing prediction system is the ultimate future goal of this research.

# 5 CONCLUSION

Early-warning systems can be used in real time for risk stratification as well as early HAI detection. In this study, a prediction system was designed to generate MRSA risk scores from easily available EHR data. Clinical time series data, mixed with data collected upon admission, contain strong predictive power for MRSA infection, even for risk scores generated far in advance of MRSA diagnosis dates. Three binary classification algorithms were trained using historical EHR data, leading to highly accurate predictions (Mean AUC = 0.98) on the day before diagnosis. We maintained high performance (Mean AUC $>$ 0.95 ) even when forcing early predictions by using only the first few days of patients' stays. Both of these classification settings lead to results far superior to our baseline classifiers trained using only on-admission data (Accuracy = 0.725, AUC = 0.665). We also considered the predictive power contained in different types of clinical data, concluding that known MRSA risk factors are not sufficient when generating predictions and that the Laboratory, Medication, and Service-related variables are the most indicative of MRSA. We successfully trained machine learning algorithms to detect MRSA far in advance of MRSA diagnosis dates by using on-admission data mixed with the first few days of throughout-admission data. This led to reliable predictions. The evidence indicates that an early warning system could be implemented for hos-

pital patients, to be updated with stay progression, generating reliable daily risk scores to aid clinical decision-making and facilitate preventive measures.

## ACKNOWLEDGEMENTS

Thomas Hartvigsen thanks the US Department of Education for supporting his PhD studies via the grant P200A150306 on "GAANN Fellowships to Support Data-Driven Computing Research", while Cansu Sen thanks WPI for granting her the Arvid Anderson Fellowship (2015-2016) to pursue her PhD studies. Sarah Brownell thanks the National Science Foundation for undergraduate research funding for Summer 2017 through the NSF REU grant #1560229 entitled "REU SITE: Data Science Research for Safe, Sustainable and Healthy Communities". We also thank the DSRG and Data Science Community at WPI for their continued support and feedback.

## REFERENCES

Aureden, K., Arias, K., Burns, L., et al. (2010). Guide to the elimination of methicillin-resistant staphylococcus aureus (mrsa): Transmission in hospital settings. washington, dc. *APIC*.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Celi, L. A., Mark, R. G., Stone, D. J., and Montgomery, R. A. (2013). "big data" in the intensive care unit. closing the data loop. *American Journal of Respiratory and Critical Care Medicine*, 187(11):1157–1160.

Chang, Y., Yeh, M., Li, Y., Hsu, C., Lin, C., Hsu, M., and Chiu, W. (2011). Predicting hospital-acquired infections by scoring system with simple parameters. *PloS One*, 6(8):e23137.

CMS (2011). Electronic health records (ehr) incentive programs. https://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/index.html.

Congress, U. S. (2009). American recovery and reinvestment act. www.healthit.gov/policy-researchers-implementers/health-it-legislation.

Dantes, R., Mu, Y., Belflower, R., Aragon, D., Dumyati, G., Harrison, L. H., Lessa, F. C., Lynfield, R., Nadle, J., Petit, S., et al. (2013). National burden of invasive methicillin-resistant staphylococcus aureus infections, united states, 2011. *JAMA Internal Medicine*, 173(21):1970–1978.

Dubberke, E. R., Yan, Y., Reske, K., Butler, A., Doherty, J., Pham, V., and Fraser, V. (2011). Development and validation of a clostridium difficile infection risk prediction model. *Infection Control & Hospital Epidemiology*, 33(4):360–366.

Dutta, R. and Dutta, R. (2006). Maximum probability rule based classification of mrsa infections in hospital environment: Using electronic nose. *Sensors and Actuators B: Chemical*, 120(1):156–165.

Fukuta, Y., Cunningham, C. A., Harris, P. L., Wagener, M. M., and Muder, R. R. (2012). Identifying the risk factors for hospital-acquired methicillin-resistant staphylococcus aureus (mrsa) infection among patients colonized with mrsa on admission. *Infection Control & Hospital Epidemiology*, 33(12):1219–1225.

Hajian-Tilaki, K. (2013). Receiver operating characteristic (roc) curve analysis for medical diagnostic test evaluation. *Caspian Journal of Internal Medicine*, 4(2):627.

Jensen, P. B., Jensen, L. J., and Brunak, S. (2012). Mining electronic health records: Towards better research applications and clinical care. *Nature Reviews. Genetics*, 13(6):395.

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. (2016). Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3.

Jones, D. A., Shipman, J. P., Plaut, D. A., and Selden, C. R. (2010). Characteristics of personal health records: Findings of the medical library association/national library of medicine joint electronic personal health record task force. *JMLA: Journal of the Medical Library Association*, 98(3):243.

Khalilia, M., Chakraborty, S., and Popescu, M. (2011). Predicting disease risks from highly imbalanced data using random forest. *BMC Medical Informatics and Decision Making*, 11(1):51.

Lebedev, A., Westman, E., Van Westen, G., Kramberger, M., Lundervold, A., Aarsland, D., Soininen, H., Kłoszewska, I., Mecocci, P., Tsolaki, M., et al. (2014). Random forest ensembles for detection and prediction of alzheimer's disease with a good between-cohort robustness. *NeuroImage: Clinical*, 6:115–125.

Maree, C., Daum, R., Boyle-Vavra, S., Matayoshi, K., and Miller, L. (2007). Community-associated methicillin-resistant staphylococcus aureus isolates and healthcare-associated infections. *Emerging Infectious Diseases*, 13(2):236.

Murdoch, T. and Detsky, A. (2013). The inevitable application of big data to health care. *Jama*, 309(13):1351–1352.

Neu, H. C. (1992). The crisis in antibiotic resistance. *Science*, 257(5073):1064–1074.

Nseir, S., Grailles, G., Soury-Lavergne, A., Minacori, F., Alves, I., and Durocher, A. (2010). Accuracy of american thoracic society/infectious diseases society of america criteria in predicting infection or colonization with multidrug-resistant bacteria at intensive-care unit admission. *Clinical Microbiology and Infection*, 16(7):902–908.

Raghupathi, W. and Raghupathi, V. (2014). Big data analytics in healthcare: Promise and potential. *Health Information Science and Systems*, 2(1):3.

Sen, C., Hartvigsen, T., Rundensteiner, E., and Claypool, K. (2017). Crest - risk prediction for clostridium difficile infection using multimodal data mining. Lecture Notes in Computer Science, pages 49–60. Springer.

Shang, J. S., Lin, Y. E., and Goetz, A. M. (2000). Diagnosis of mrsa with neural networks and logistic regression approach. *Health Care Management Science*, 3(4):287.

Sintchenko, V., Coiera, E., and Gilbert, G. L. (2008). Decision support systems for antibiotic prescribing. *Current Opinion in Infectious Diseases*, 21(6):573–579.

Ventola, C. L. (2015). The antibiotic resistance crisis: Part 1: Causes and threats. *Pharmacy and Therapeutics*, 40(4):277.

Visser, H., le Cessie, S., Vos, K., Breedveld, F. C., and Hazes, J. M. (2002). How to diagnose rheumatoid arthritis early: a prediction model for persistent (erosive) arthritis. *Arthritis & Rheumatology*, 46(2):357–365.

Weiner, L., Webb, A., Limbago, B., Dudeck, M., Patel, J., Kallen, A., Edwards, J., and Sievert, D. (2016). Antimicrobial-resistant pathogens associated with healthcare-associated infections: summary of data reported to the national healthcare safety network at the centers for disease control and prevention, 2011–2014. *Infection Control & Hospital Epidemiology*, 37(11):1288–1301.

Wiens, J., Guttag, J., and Horvitz, E. (2012). Learning evolving patient risk processes for c. diff. colonization. In *ICML Workshop on Machine Learning from Clinical Data*.

Wu, J., Roy, J., and Stewart, W. F. (2010). Prediction modeling using ehr data: challenges, strategies, and a comparison of machine learning approaches. *Medical Care*, 48(6):S106–S113.