CREST - Risk Prediction for Clostridium Difficile Infection Using Multimodal Data Mining

Cansu Sen[1](✉), Thomas Hartvigsen[1], Elke Rundensteiner[1], and Kajal Claypool[2]

[1] Worcester Polytechnic Institute, Worcester, MA, USA
[2] Harvard Medical School, Boston, MA, USA
{csen,twhartvigsen,rundenst}@wpi.edu, kajal_claypool@hms.harvard.edu

Abstract. Clostridium difficile infection (CDI) is a common hospital acquired infection with a $1B annual price tag that resulted in ∼30,000 deaths in 2011. Studies have shown that early detection of CDI significantly improves the prognosis for the individual patient and reduces the overall mortality rates and associated medical costs. In this paper, we present CREST: **C**DI **R**isk **Est**imation, a data-driven framework for *early* and *continuous* detection of CDI in hospitalized patients. CREST uses a three-pronged approach for high accuracy risk prediction. First, CREST builds a rich set of highly predictive features from Electronic Health Records. These features include clinical and non-clinical phenotypes, key biomarkers from the patient's laboratory tests, synopsis features processed from time series vital signs, and medical history mined from clinical notes. Given the inherent multimodality of clinical data, CREST bins these features into three sets: time-invariant, time-variant, and temporal synopsis features. CREST then learns classifiers for each set of features, evaluating their relative effectiveness. Lastly, CREST employs a second-order meta learning process to ensemble these classifiers for optimized estimation of the risk scores. We evaluate the CREST framework using publicly available critical care data collected for over 12 years from Beth Israel Deaconess Medical Center, Boston. Our results demonstrate that CREST predicts the probability of a patient acquiring CDI with an AUC of 0.76 five days prior to diagnosis. This value increases to 0.80 and even 0.82 for prediction two days and one day prior to diagnosis, respectively.

Keywords: Clostridium difficile, Risk stratification, Multimodal data mining, Multivariate time series classification, Electronic Health Records

1 Introduction

Motivation. Clostridium difficile infection (CDI) is a common hospital acquired infection resulting in gastrointestinal illness with substantial impact on morbidity and mortality. In 2011, nearly half a million CDI infections were identified in

the US resulting in 29,000 patient deaths [1, 2]. Despite well-known risk factors and the availability of mature clinical practice guidelines [5], the infection and mortality rates of CDI continue to rise with an estimated $1 billion annual price tag [6]. Early detection of CDI has been shown to be significantly correlated with a successful resolution of the infection within a few days, and is projected to save $3.8 billion in medical costs over a period of 5 years [7]. In current practice, a diagnostic test is usually ordered as a confirmation of a highly-suspect case, only after appearance of symptoms[3]. This points to a tremendous opportunity for employing machine learning techniques to develop intelligent systems for early detection of CDI to eradicate this medical crisis.

State-of-the-art. Our literature review shows that there have been some initial efforts to apply machine learning techniques to develop risk score estimation models for CDI. These efforts largely exploit two approaches. The first, a *moment-in-time approach*, uses only the data from one single moment in patient's stay. This moment can be the admission time [18] or the most recent snapshot data at the time of risk estimation [17]. The second, an *independent-days approach*, uses the complete hospital stay, but treats the days of a patient's stay as independent from each other [14, 15]. The *complete physiological state* of the patient, *changes in the physiological state*, and *clinical notes* containing past medical information have been left out of the risk prediction process.

Challenges. To fill this gap, the following challenges must be addressed:

Varying lengths of patient stays. Stay-lengths vary between patients, complicating the application of learning algorithms. Thus, we must design a fixed-length representation of time series patient-stay data. This requires temporal summarization of data such that the most relevant information for the classification task is preserved.

Incorporating clinical notes. Clinical notes from a patient's EHR contain vital information (e.g., co-morbidities and prior medications). These are often taken in short-hand and largely abbreviated. Mining and analysis of clinical notes is an open research problem, but some application of current techniques is necessary to transform them into a format usable for machine learning algorithms.

Combining multimodal data. EHR data is typically multimodal, including text, static data and time series data, that require transformation and normalization prior to use in machine learning. The choices made when transforming the data may have significant impact on classification accuracy if key transformations are not appropriate for the domain.

Our Proposed CREST System. CREST: **C**DI **R**isk **Est**imation is a novel framework that addresses these challenges and estimates the risk of a patient contracting CDI. Figure 1 gives an overview of CREST. CREST extracts highly predictive features capturing both time-invariant and time-variant aspects of patient histories from multimodal input data (i.e., consisting of clinical and non-clinical phenotypes, biomarkers from lab tests, time series vital signs, and clinical

[3] The authors would like to thank Elizabeth Claypool, RN, Coordinator of Patient Safety at U. Colorado Health for the valuable information she provided.

notes) while maintaining temporal characteristics. Feature selection methods are applied to select the features with the highest predictive power. Feeding these selected features into the classification pipeline, multiple models are fit ranging from primary classifiers to meta-learners. Once trained, CREST continuously generates daily risk scores to aid medical professionals by flagging at-risk patients for improved prognoses.

Contributions. In summary, our contributions include:

1. Time-alignment of time series data. We design two time-alignment methods that solve the varying length of patient's stay problem. This enables us to bring a *multiple-moments-in-time* approach to the task of predicting patient infections.

2. Multimodal feature combination. To our knowledge, CREST is the first work to combine clinical notes and multivariate time series data to perform classification for CDI risk prediction. We show that synopsis temporal features from patient time-series data significantly improve classification performance, while achieving interpretable results.

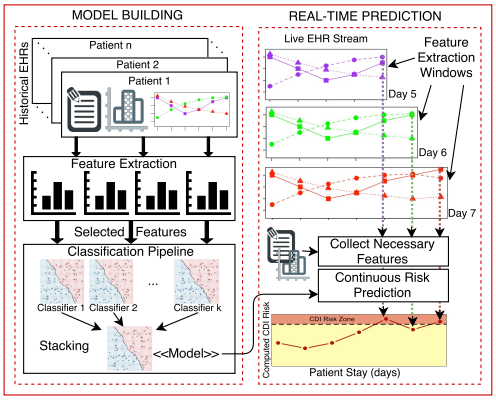


Fig. 1. Overview of CREST framework

3. Early detection of the infection. We evaluate our system with publicly-available critical-care data collected at the Beth Israel Deaconness Intensive Care Unit in Boston, MA [8]. Our evaluation shows that CREST improves the accuracy of predicting high-risk CDI patients by 0.22 one day before and 0.16 five days before the actual diagnosis compared to risk estimated using only admission time data.

2 Predictive Features of CREST

We categorize patient EHR information into three feature sets: time-invariant, time-variant, and temporal synopsis. An overview of our feature extraction process is depicted in Figure 2.

2.1 Time-Invariant and Time-Variant Properties of EHR Data

Time-Invariant Properties. These represent all data for a patient known at the time of admission which does not change throughout the patient's stay. A number of known CDI risk factors are represented in this data (e.g. age, prior antibiotic usage). To capture these, we extract a set of time-invariant features. ***Demographic features*** are immutable patient features such as age, gender, and

ethnicity. ***Stay-specific features*** describe a patient's admission such as admission location and insurance type, allowing inference on the patient's condition. These data could be different for the same patient upon readmission. ***Medical history features*** model historical patient co-morbidities (e.g., diabetes, kidney disease) and medications (e.g., antibiotics, proton-pump inhibitors) associated with increased CDI risk. These are extracted from clinical notes (free-form text files) using text mining. Using the Systematized Nomenclature of Medicine Clinical Terms dictionary (SNOMED CT), synonyms for these diseases and medications are identified to facilitate extraction of said factors from a patient's history.

Time-Variant Properties. Throughout the hospital stay of a patient, many observations are recorded continuously such as laboratory results and vital signs, resulting in a collection of time series. A data-driven approach is leveraged to model all of this data as time-variant features. Additionally, for each day of a patient's stay, we generate multiple binary features flagging the use of antibiotics, H2 antagonists, and proton pump inhibitors, all of which are known to have causal relationships with CDI. Particularly high risk antibiotics, namely Cephalosporins, Fluoroguinolones, Macrolides, Penicillins, Sulfonamides, and Tetracyclines [16], are captured by another binary feature flagging the presence of high-risk antibiotics in a patient's body. Using a binary feature avoids one-hot encoding, a method known to dramatically increases dimensionality and sparseness.

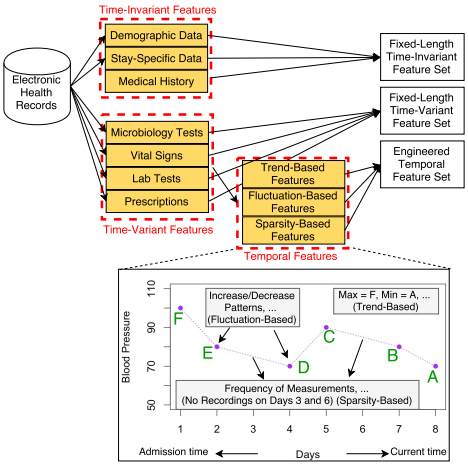
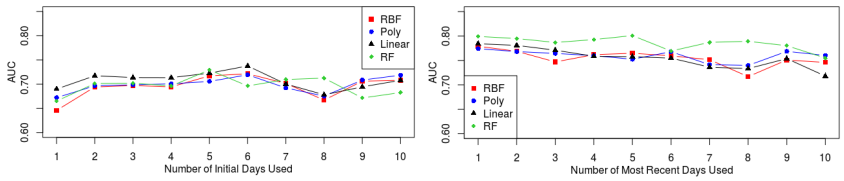


Fig. 2. Feature extraction process

2.2 Two Strategies for Modeling Variable-Length Time-Series Data

Time-Alignment for Time-Series Clinical Data. A patient's stay is recorded as a series of clinical observations that is often characterized as *irregularly spaced time series*. These measurements vary in the frequency at which they are taken (once a day, multiple times a day, etc.). This variation is a function of (a) the observation (a lab test can be taken once a day while a vital sign is measured multiple times), (b) the severity of the patient's condition (patients in more severe conditions must be monitored more closely), and (c) the time of the day (nurses are less likely to wake up patients in the middle of the night). To unify this, we roll up all observations taken more than once a day into evenly sampled averages at the granularity of one day. If there are no measurements for a day, these are considered as missing values and are filled with the median value.

The total number of observations recorded per patient is a function of not only the frequency of observation, but also the length of a patient's stay. After day-based aggregation of continuous data, we produce a fixed-length feature representation by time-aligning the variable-length feature vectors. This time-alignment can be done by either using the same number of *initial days* since admission or the same number of *most recent days* of each patient's hospital stay. We empirically determine the optimal time-alignment window of the patient's stay by evaluating the AUC of the initial days and the most recent days using Random Forests on only time-aligned data. Our results show that AUC using the most recent days was much higher than using the initial days of a patient's stay. We validate our results using SVMs, as shown in Figure 3. Based on these results, we conclude that when predicting CDI risk on day p, the most recent 5 days of the patient stay (i.e. days $p-5$ to $p-1$) capture the most critical information. This is consistent with and validated by the incubation period of CDI (<7 days with a median of 3 days [19, 5]). In CREST, we thus use only the most recent 5 days of each patient's stay as our approach to represent patient vital signs and lab/microbiology tests as continuous numerical feature vectors.



(a) Initial days (b) Most recent days

Fig. 3. AUC results using initial and most-recent days of patient stays shows that using the most recent 5 days contains the most information about the CDI risk.

Computing Temporal Synopsis Features. Time-variant features (e.g., temperature), while capturing the state of the patient for each day of their stay, falsely treat days to be independent from each other. Thus, they do not capture the sequential trends over time inherent in these time series data. For example, the presence or absence of recordings of a time-variant feature may be more informative than the actual values (e.g., *heart rate high alarm* is only measured when a patient has an alarmingly high heart rate). In some cases, the change in an observation (e.g., increase in temperature) may be more important than the actual observed values. To effectively model these trends, in CREST, we introduce feature computation functions, capturing the following temporal synopsis features:

– **Trend-based features** include statistics such as minimum, maximum, and average values. In addition to an equal weighted average, linear and quadratic weighted averages are computed, giving more weight to later days. The relative times of the first and last recordings and of minimum and maximum recordings are also extracted to signal when in a patient's hospital stay these notable events occur.

– **Fluctuation-based features** capture the change characteristic of each time-variant feature. Mean absolute differences, number of increasing and decreasing recordings and the ratio of change in direction are examples of trends we extract to capture these characteristics.
– **Sparsity-based features** model frequency of measurements and proportion of missing values. For example, "heart rate high alarm" is recorded only if a patients heart rate exceeds the normal threshold.

Figure 2 illustrates the time-variant feature blood pressure for a patient and examples of trends we extract from this time series data.

3 Modeling Infection Risk In CREST

3.1 Robust Supervised Feature Selection

In CREST, each extracted feature set is fed into a rigorous feature selection module to determine the features that are most relevant to CDI risk. We denote $S_{n\times s}$, $D_{n\times d}$, and $T_{n\times t}$ to be the time-invariant, time-variant, and temporal feature matrices with n instances and s, d, and t features respectively. For a compact representation, we use X to represent S, D, and T. The goal is to reduce $X_{n\times p}$ into a new feature matrix $X'_{n\times k}$ where $X'_{n\times k} \subset X_{n\times p}$. To achieve this, we combine chi-squared feature selection, a supervised method that tests how features depend on the label vector Y, with SVMs. Two issues must be addressed when using this method, namely, determining the optimal cardinality of features, and which features to use.

Percentile Selection. We first determine the cardinality of features for each feature set. Using 10-fold cross validation over training data, we select the top K percent of features for $K = (5, 10, 15, \ldots, 100)$ and record the average AUC value by percentile for each of the three feature sets. We then select the percentiles that perform the best.

Robustness Criterion. Next, we select as few features as possible while ensuring adequate predictive power. We empirically select which features to use by choosing a robustness criterion, γ, which we define as "the minimum number of folds in which features must appear to be considered *predictive*". Since we have 10 cross-validation folds, $\gamma \in [1\colon 10]$, where $\gamma = 1$ implies all features selected for *any* folds are included in the final feature set (union) and $\gamma = 10$ implies all features selected for *every* fold is included in the final feature set (intersection).

We apply these steps to feature matrices S, D, and T, resulting in reduced feature matrices S', D', and T'.

3.2 CREST Learning Methodology

We represent a patient's CDI risk as the probability that the patient gets infected with CDI. To compute this probability, we estimate a function $f(X') \rightarrow Y$ using the reduced feature matrix X' (representing S', D', or T') and the label vector Y, consisting of binary diagnosis outcomes. The function outputs a vector of

predicted probabilities, $\hat{Y}$. In a hospital setting, CREST extracts a feature matrix X' every day of a patient's hospital stay. CREST then employs the classification function on X' (see Figure 1 for this continuous process). This section describes the process of estimating the function f, shown in Figure 4.

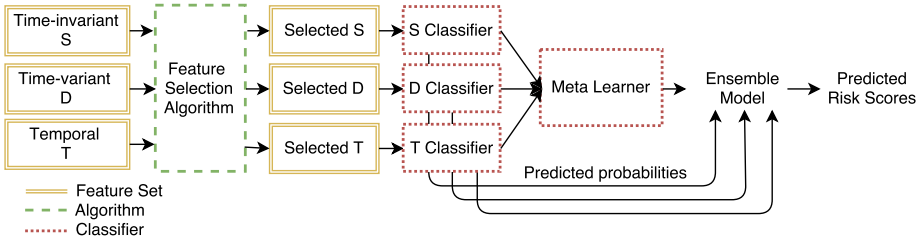


Fig. 4. Learning phase of the CREST Framework.

Type-specific Classification. We first train a set of type-specific classifiers built on each of the feature matrices. The task is to estimate $f(X') \rightarrow Y$ which minimizes $|Y - \hat{Y}|$. We use SVMs, Random Forests, and Logistic Regression to estimate f. Since imbalanced data is typical in this application domain, CREST uses a modified SVM objective function that includes two cost parameters for positive and negative classes. Thus, a higher misclassification cost is assigned to the minority class. Equation 1 shows the modified SVM objective function we used in CREST and Equation 2 shows how we choose the cost for positive and negative classes.

$$\begin{aligned} &\text{minimize } (\frac{1}{2} w \cdot w + C^{+} \sum_{i \in \mathcal{P}}^{l} \xi_i + C^{-} \sum_{i \in \mathcal{N}}^{l} \xi_i) \\ &\text{s.t. } \; y_i(w \cdot \Phi(x_i) + b) \geq 1 - \xi_i \; \xi_i \geq 0, \quad i = 1 \ldots l. \end{aligned} \tag{1}$$

$$C^{+} = C \; \frac{l}{|\mathcal{P}|}, \quad C^{-} = C \; \frac{l}{|\mathcal{N}|} \tag{2}$$

Where w is a vector of weights, $\mathcal{P}$ is the positive class, $\mathcal{N}$ is the negative class, l is the number of instances, C is the cost, ξ is a set of slack variables, x_i is i^{th} data instance, Φ is a kernel function, and b is the intercept.

A *static classifier*, trained on feature set S' extracted from admission time data, implies that only the information obtained upon a patient's admission is necessary to accurately predict their risk. This constitutes our baseline as it represents the current practice of measuring risk in hospitals and denotes risk on day 0. A *dynamic classifier*, trained on feature set D', constitutes a multiple-moments-in-time approach where the data from many different moments in a patient's stay are used as features. This approach allows us to quantify the relationship between the *physiological state* of the patient and their CDI risk. Finally, a *temporal classifier*, trained using feature set T', quantifies the relationship between a patient's *state-change* and their risk, complementing the time-variant

features.

Second-order Classification. Since the three type-specific classifiers capture different aspects of a patient's health and hospital stay, we combine them to produce a single continuous prediction based on comprehensive information. We hypothesize that this combination method, termed *second-order classification*, will provide more predictive power. To evaluate this hypothesis, we merge the predicted probability vectors from the type-specific classifiers into a new higher-order feature set $X_{meta} = (\hat{Y}_S, \hat{Y}_D, \hat{Y}_T)$. With this new feature matrix, our task becomes estimating a function $f(X_{meta}) \rightarrow Y$. Beyond naive methods such as model averaging to assign weights to the results produced by the type-specific classifiers, we also develop a stacking-based solution. We train meta learners fusing SVMs with RBF and linear kernels, Random Forests, and Logistic Regression on X_{meta} to learn an integrated ensemble classifier. Henceforth, final predictions are made by these new second-order classifier models.

4 Evaluation of CREST Framework

4.1 MIMIC-III ICU Dataset and Evaluation Settings

The MIMIC III Database [8], used to evaluate our CREST Framework, is a publicly available critical care database collected from the Beth Israel Deaconess Medical Center Intensive Care Unit (ICU) between 2001 and 2012. The database consists of information collected from ∼45,000 unique patients and their ∼58,000 admissions upon time of entry and throughout their stays. Each patient's record consists of laboratory tests, medical procedures, medications given, diagnoses, caregiver notes, etc.

Of the $58,000$ admissions in the MIMIC III Database, there are 1079 cases of CDI. Approximately half of these patients were diagnosed either before or within the first 4 days of their admission. To ensure that CDI cases in our evaluation dataset are contracted during the hospital stay, we exclude patients who test positive for CDI within their first 5 days of hospitalization based on the incubation period of CDI [19, 5]. For consistency between CDI and non-CDI patients, we also exclude non-CDI patients whose hospital stay is less than 5 days. As the vast majority of the MIMIC III Database consists of patients who do not contract CDI, we end up with an unbalanced dataset (116:1). To overcome this, we randomly subsample from the non-CDI patients to get a 2-to-1 proportion of non-CDI to CDI patients, leaving us with 1328 patient records.

Next, we define the feature extraction window for patients. For CDI patients, it starts on the day of admission and ends n days before the CDI diagnosis, $n \in \{1, \ldots, 5\}$. For non-CDI patients, there are a few alternatives for defining this window. Prior research has used the discharge day as the end of the risk period [17]. However, as the state of the patients can be expected to improve nearing their discharge, this may lead to deceptive results [14]. Instead, we use the halfway point of the non-CDI patient's stay as the end of the risk period or 5 days (minimum length of stay), whichever is greater.

We then split these patients into training and testing subsets with a 70%-30% ratio and maintain these subsets across all experiments. The training set is further split and 5-fold cross-validation is applied to perform hyper-parameter search. We use SVM with linear and RBF kernels, Random Forest and Logistic Regression. All algorithms were implemented using Scikit-Learn in Python.

4.2 Classification Results

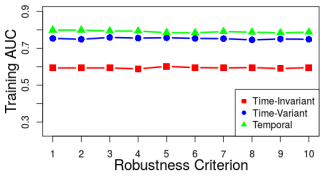


Fig. 5. Selection of robustness criterion

Using our feature selection module, we find the best cardinalities to be $K = 20$ for time-invariant, $K = 30$ for time-variant, $K = 90$ for temporal feature sets with robustness-criterion $\gamma = 10$ for all three feature sets. This choice of γ is motivated by an almost unchanging validation AUC over all potential γ values, as shown in Figure 5. This shows that mostly the same features are selected for each fold. By choosing $\gamma = 10$, we can be certain that only the features that are strongly related to the response variable are selected.

We first run a set of experiments with type-specific classifiers to determine the predictive power of each type of feature class. We then experiment with ensembles of the type-specific classifiers in two ways: (1) **Equal-weighted model averaging:** We calculate equal weighted averages of the probabilities produced by each type-specific classifier, (2) **Meta-learning:** We train second order meta learners using the outputs of the type-specific classifiers as the input of the meta learners. Table 1 shows the AUC, precision, recall and F-1 scores for each classification method.

Table 1. Classification results acquired on the test set.

		AUC	Precision	Recall	F-1
Static C.	SVM RBF	0.544	0.57	0.62	0.58
	SVM Linear	0.627	0.76	0.46	0.38
	Random F.	0.608	0.57	0.62	0.58
	Logistic R.	0.627	0.6	0.64	0.59
	Average	**0.602**	**0.63**	**0.59**	**0.53**
Dynamic C.	SVM RBF	0.779	0.73	0.73	0.71
	SVM Linear	0.756	0.71	0.72	0.69
	Random F.	0.818	0.75	0.76	0.75
	Logistic R.	0.758	0.72	0.73	0.71
	Average	**0.778**	**0.73**	**0.74**	**0.72**
Temporal C.	SVM RBF	0.815	0.76	0.77	0.76
	SVM Linear	0.817	0.76	0.72	0.72
	Random F.	0.832	0.77	0.77	0.77
	Logistic R.	0.809	0.75	0.76	0.75
	Average	**0.818**	**0.76**	**0.76**	**0.75**
Model Avg.		**0.817**	**0.76**	**0.71**	**0.65**
Meta Learn.	SVM RBF	0.838	0.76	0.76	0.75
	SVM Linear	0.833	0.76	0.73	0.74
	Random F.	0.815	0.74	0.75	0.74
	Logistic R.	0.831	0.76	0.77	0.76
	Average	**0.829**	**0.76**	**0.75**	**0.75**

Static classifiers constitute our baseline approach. The mean AUC of all static classifiers is 0.60, implying that a risk score can be assigned to a patient at the time of admission. Dynamic classifiers, which use time-variant features, achieve a much higher

AUC compared to the static classifiers. This shows that the physiological state of a patient is correlated with the CDI outcome. Among the type-specific classifiers, the temporal classifiers consistently attain the highest AUC. This highlights that patient-state changes are strongly predictive of CDI risk. To the best of our knowledge, ours is the first effort that uses this information to predict CDI risk for patients. Between our two ensemble methods, meta-learners further improve the prediction success over any of the type-specific classifiers, showing that considering all features together is beneficial. The highest AUC is achieved by meta-learners when an SVM with an RBF kernel is used. Figure 6 presents the ROC curves for type-specific classifiers and the meta learners, which show an increasing trend in diagnosis accuracy.

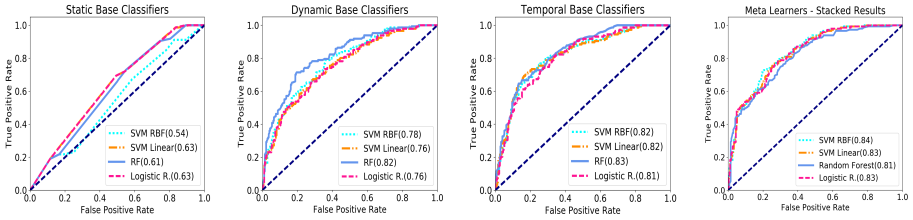


Fig. 6. ROC curves for Static, Dynamic, Temporal, and Meta Classifiers

4.3 Early Prediction of CDI

The earlier an accurate prediction can be made, the higher the likelihood that actions can be taken to prevent contraction of CDI. We evaluate the power of our model for early prediction using the best CREST meta learner (Section 4.2). Unlike the experiments above, we now train models using the data 1 to 5 days prior to CDI diagnosis. As presented in Figure 7, results indicate that early warnings can maintain high AUC values. In comparison with the baseline methods where the mean AUC is 0.60, CREST improves the accuracy of predicting high-risk CDI patients to 0.82 one day prior to diagnosis and to 0.76 five days prior to diagnosis, an improvement of 0.22 and 0.16 over the baseline respectively.

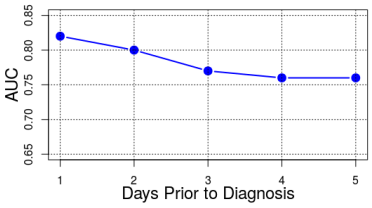


Fig. 7. AUC results of early prediction experiments

5 Related Work

Feature Extraction from Time Series. One strategy to deal with clinical time series in machine learning is to extract aggregated features. In healthcare, much work has gone into extracting features from signals such as ECG [9] or EEG [10, 11] using methods such as wavelet [12, 9] or Fourier [12] transformations. However, EHR time series have largely being ignored. Specially designing

feature extraction techniques for EHRs in our model, we demonstrate that prediction accuracy increases using these features over models that do not account for the temporal aspects of the data.

In-Hospital CDI Prediction. Recent work has begun to investigate prediction models for CDI. [14, 15] ignore temporal dependencies in the data and reduce this complex task to univariate time-series classification. [13] while combining time-variant and time-invariant data, neglect the trends in patient records. [4] uses ordered pairs of clinical events to make predictions, missing longer patterns in their data. In our work, we effectively apply multivariate time series classification while capturing temporal characteristics and long-term patterns in EHRs.

SVMs [3, 14, 15] and Logistic Regression [3, 4, 13, 17, 18] are popular tools for CDI risk prediction models. We apply a variety of models including SVM, Random Forest, Logistic Regression and ensembles of those to produce more comprehensive results.

6 Conclusion

CREST is the first system that stratifies a patient's infection risk on a continuous basis throughout their stay and is based on a novel feature extraction and combination method. CREST has been validated for CDI risk using the MIMIC Database. Our experimental results demonstrate that CREST can detect CDI cases with an AUC score of up to 0.84 one day before and 0.76 five days before the actual diagnosis. CDI is a highly contagious disease and early detection of CDI not only greatly improves the prognosis for individual patients by enabling timely precautions but also prevents the spread of the infection within the patient cohort. To our knowledge, this is the first work on multivariate time series classification to predict the risk of CDI. We also demonstrate that our extracted temporal synopsis features improve the AUC by 0.22 over the static classifiers and 0.04 over the dynamic classifiers.

Although we currently focus on the application of the CREST framework to solve the *Clostridium Difficile Infection* crisis, we plan to test our hypothesis in the future that our extracted CREST feature sets also have predictive power for tackling other hospital-acquired infections. Another important aspect of frameworks such as CREST is their applicability to other ICU environments. We are currently in discussion with Partners Health about the validation of CREST in the context of their healthcare systems.

Acknowledgments. The authors thank Dr. Richard T. Ellison, III, the head of Infection Control at UMass Memorial Medical Center, Worcester, MA, for his valuable comments that helped us understand the urgency of the CDI crisis. The authors also thank Dr. Alfred DeMaria, Medical Director for the Bureau of Infectious Diseases at Massachusetts Public Health Department for highlighting the effects of this crisis on the healthcare system across Massachusetts and beyond.

References

1. Centers for Disease Control and Prevention: https://www.cdc.gov/media/releases /2015/p0225-clostridium-difficile.html (2017)
2. Lessa, F.C., et al.: Burden of Clostridium difficile infection in the United States. New England Journal of Medicine, 372(9), 825-834 (2015)
3. Wu, J., et al.: Prediction modeling using EHR data: Challenges, strategies, and a comparison of machine learning approaches. Medical Care, 48(6), S106-S113 (2010)
4. Monsalve, M., et al.: Improving risk prediction of Clostridium Difficile Infection using temporal event-pairs. In: International Conference on Healthcare Informatics, IEEE, 140-149 (2015)
5. Cohen, S.H., et al.: Clinical practice guidelines for Clostridium difficile infection in adults: 2010 update by the Society for Healthcare Epidemiology of America (SHEA) and the Infectious Diseases Society of America (IDSA). Infection Control & Hospital Epidemiology, 31(05), 431-455 (2010)
6. Evans, C.T., Safdar, N.: Current trends in the epidemiology and outcomes of Clostridium difficile infection. Clinical Infectious Diseases, 60(suppl 2), S66-S71 (2015)
7. Centers for Disease Control and Prevention: Antibiotic resistance threats in the United States. https://www.cdc.gov/drugresistance/biggest threats.html (2017)
8. Johnson, A.E., et al.: MIMIC-III, a freely accessible critical care database. Scientific Data (3) (2016)
9. Sternickel, K.: Automatic pattern recognition in ECG time series. Computer Methods and Programs in Biomedicine, 68(2), 109-115 (2002)
10. Chaovalitwongse, W.A., Prokopyev, O.A. and Pardalos, P.M.: Electroencephalogram (EEG) time series classification: Applications in epilepsy. Annals of Operations Research, 148(1), 227-250 (2006)
11. Lemm, S., et al.: Spatio-spectral filters for improving the classification of single trial EEG. IEEE Transactions on Biomedical Engineering, 52(9), 1541-1548 (2005)
12. Zhang, H., et. al.: Feature extraction for time series classification using disc. wavelet coefficients. Advances in Neural Networks, ISNN-2006, 1394-1399 (2006)
13. Wiens, J., et al.: Learning data-driven patient risk stratification models for Clostridium difficile. Open Forum Infectious Diseases, 1(2), ofu045 (2014)
14. Wiens, J., et al.: Learning evolving patient risk processes for C. diff colonization. In: ICML Workshop on Machine Learning from Clinical Data (2012)
15. Wiens, J., Horvitz, E., Guttag, J.V.: Patient risk stratification for hospital-associated C. diff as a time-series classification task. In: Advances in Neural Information Processing Systems, 467-475 (2012)
16. Kuntz, J.L., et al.: Incidence of and risk factors for community-associated C. difficile infection: A nested case-control study. BMC infectious diseases, 11(1), 194 (2011)
17. Dubberke, E.R., et al.: Development and validation of a C. diff. infection risk prediction model. Infection Control & Hospital Epidemiology, 32(04), 360-366 (2011)
18. Tanner, J., et al.: Waterlow score to predict patients at risk of developing C. difficile-associated disease. Journal of Hospital Infection, 71(3), 239-244 (2009)
19. Dubberke, E.R., et al.: Hospital-associated Clostridium difficile infection: is it necessary to track community-onset disease?. Infection Control & Hospital Epidemiology, 30(04), 332-337 (2009)