Journal of Machine Learning Research 17 (2016) 1-23 Submitted 5/15; Revised 3/16; Published 4/16

Patient Risk Stratification with Time-Varying Parameters: A Multitask Learning Approach

Jenna Wiens WIENSJ@UMICH.EDU
Computer Science & Engineering
University of Michigan
Ann Arbor, MI

John Guttag GUTTAG@CSAIL.MIT.EDU
Department of EECS
Massachusetts Institute of Technology
Cambridge, MA

Eric Horvitz HORVITZ@MICROSOFT.COM
Microsoft Research
Redmond, WA

Editor: Benjamin M. Marlin, C. David Page, and Suchi Saria

Abstract

The proliferation of electronic health records (EHRs) frames opportunities for using machine learning to build models that help healthcare providers improve patient outcomes. However, building useful risk stratification models presents many technical challenges including the large number of factors (both intrinsic and extrinsic) influencing a patient's risk of an adverse outcome and the inherent evolution of that risk over time. We address these challenges in the context of learning a risk stratification model for predicting which patients are at risk of acquiring a *Clostridium difficile* infection (CDI). We take a novel data-centric approach, leveraging the contents of EHRs from nearly 50,000 hospital admissions. We show how, by adapting techniques from multitask learning, we can learn models for patient risk stratification with unprecedented classification performance. Our model, based on thousands of variables, both time-varying and time-invariant, changes over the course of a patient admission. Applied to a held out set of approximately 25,000 patient admissions, we achieve an area under the receiver operating characteristic curve of 0.81 (95% CI 0.78-0.84). The model has been integrated into the health record system at a large hospital in the US, and can be used to produce daily risk estimates for each inpatient. While more complex than traditional risk stratification methods, the widespread development and use of such data-driven models could ultimately enable cost-effective, targeted prevention strategies that lead to better patient outcomes.

Keywords: risk stratification, time-varying coefficients, multitask learning, *Clostridium difficile*, healthcare-associated infections

1. Introduction

Over recent years, there has been enormous growth in 1) our capacity to gather clinically relevant data and 2) the availability of such data sets. The collection of these data, in particular electronic health records (EHRs), holds out the promise of using machine learning to build models that can be harnessed to improve patient outcomes. Transforming patient data into actionable knowledge presents a barrage of pragmatic and technical challenges. But if we are successful in addressing these challenges, the knowledge embedded in these data has the potential to revolutionize clinical medicine.

One way in which these data can be leveraged is in the development of accurate data-driven models for predicting potentially avoidable adverse outcomes and using such predictions to guide interventions aimed at reducing the probability of these outcomes. The hypothesis is that we can extract from the data generalizable information that can help accurately identify a patient's *future* pathological states. If pathologies are predicted far enough in advance, then it may be possible for healthcare workers to intervene. Such targeted interventions could, in turn, lead to better patient outcomes.

In recent years, there has been a significant amount of research effort devoted to using clinical data to predict patient outcomes (Shoeb and Guttag, 2010; Syed and Rubinfeld, 2010; Syed et al., 2011; Chia et al., 2012; Saria et al., 2010; Saeed et al., 2011; Kleinberg and Hripcsak, 2011; Aboukhalil et al., 2008; Kansagara et al., 2011). We focus on the specific task of predicting which patients in a hospital will acquire an infection with *Clostridium difficile* (*C. difficile*), a largely preventable adverse outcome (Yokoe et al., 2008). *C. difficile* is a type of bacteria that takes over a patient's gut when normal flora get wiped out (often from receipt of antimicrobials). *C. difficile* infection (CDI) can lead to severe diarrhea and intestinal diseases (e.g., colitis), or even death. The infection is often treated with specific antimicrobials: oral vancomycin and metronidazole (and less frequently, fidaxomicin). However, it is estimated that approximately 20% of cases relapse within 60 days (Pépin et al., 2005). The incidence of CDI in the US is estimated at 200,000 cases per year (Dubberke et al., 2009); this is on par with the number of new cases of invasive breast cancer discovered each year in the US (DeSantis et al., 2014).

Infection with *C. difficile* is a type of healthcare-associated infection (HAI). HAIs are a serious problem in healthcare facilities across the world. It is estimated that, on any given day, HAIs affect approximately 1 in every 25 inpatients in US acute care hospitals (Magill et al., 2014). In addition to *C. difficile*, other common HAIs include ventilator-associated pneumonia, surgical site infection, and infections with methicillin-resistant *Staphylococcus aureus* (MRSA) and vancomycin-resistant *Enterococcus* (VRE). Though many risk factors are well-known (e.g., healthcare-associated exposure, age, underlying disease, etc.), HAIs continue to be a significant problem throughout the world (Klevens et al., 2007). In recent years there have been numerous articles citing our inability to prevent HAIs (Miller et al., 2011; Umscheid et al., 2011; Sievert et al., 2013). We hypothesize that one of the reasons HAIs remain so stubbornly prevalent is because we lack an effective clinical tool for accurately measuring patient risk. In this work, we chose to focus on infections with *C. difficile*, one of the most prevalent HAIs (Miller et al., 2011).

We take a data-centric approach to the problem of developing a model to predict a patient's daily risk of acquiring an infection with *C. difficile.* We leverage the contents of EHRs from over 50,000 patient admissions from a single hospital. These clinical data contain information regarding medications, procedures, in-hospital locations, healthcare staff, lab results, measurements of vitals, patient demographics, patient history and admission details.

We seek a mapping from this information describing a patient to an estimate of the patient's probability of acquiring an infection.

Automated patient risk stratification, based on the contents of the patient's EHR, can serve several purposes. Firstly, risk-stratification models can help clinicians match high-risk patients with the appropriate interventions, monitoring policies, or therapies. In the absence of effective risk stratification, widespread implementation of known interventions (e.g., isolating patients or performing specialized analyses of antibiotic regimens) is prohibitively expensive. Secondly, data-driven models can help generate hypotheses regarding potential risk factors, in turn improving our understanding of the disease. For example, the model could identify "hot-spots" within a hospital that could benefit from additional environmental cleanings. The construction of a predictive model can also help to frame new scientific hypotheses through the identification of discriminatory observations. Such insights can lead to the pursuit and confirmation of causal relationships. Thirdly, such models could aid in designing more efficient clinical trials by identifying a study population at higher risk for disease, increasing the fraction of patients expected to test positive in the trial. This could significantly reduce the cost of a clinical trial without compromising the statistical power of the study.

Learning accurate risk-stratification models from EHR data presents a number of technical challenges. Two main issues we focus on in our work include the high dimensionality of the problem and the complex temporal dependencies among the variables. There can be thousands of variables representing each day of a patient admission and it is likely that many of these variables affect a patient's risk of CDI. Moreover, many of these variables change over time. These time-varying data suggest that as a patient spends time in the hospital, his/her actual risk of CDI will vary. Furthermore, how these variables affect risk is likely to change over time. Recent efforts on building models for identifying patients at high risk of acquiring a CDI have ignored these issues. Prior work on risk-stratification models for CDI has centered on the consideration of a small number of risk factors selected by clinical experts and time-invariant parameters.

Our hospital-specific approach to patient risk stratification for CDI, produces *daily* estimates of patient risk. The novel aspects of our work and our main contributions are outlined as follows:

- We move beyond known risk factors to leverage the entire structured contents of the EHR. Our model is based on thousands of extracted binary variables, many of which are hospital-specific (e.g., the locations of patient rooms within the hospital).
- We include both time-varying and time-invariant variables and we explicitly consider the evolution of patient risk during admission, when estimating current risk.
- We develop a novel multitask learning approach to modeling the time-varying effects of risk factors, based on the domain adaptation techniques presented in Daumé III (2007).
- We propose an evaluation scheme that is representative of how the model will be applied in practice. In contrast to previous work, we do not evaluate how our model performs at a single point in time, but rather how the model performs when applied to each day of a patient's admission.

When tested on a holdout set consisting of 24,399 patient admissions from a single year, our proposed model achieved an area under the receiver operating characteristic curve (AUROC) of 0.81 (95%CI 0.78-0.84) and consistently outperformed a baseline model with time-invariant coefficients, for patients with longer risk periods. We have shown that our algorithm can be integrated into the health record system of a hospital and can be used to automatically calculate daily probabilities of risk of CDI for every adult inpatient. These estimates could in turn be used for the selective targeting of high-risk patients with specific interventions that could lead to changes in clinical practice and ultimately a reduction in the incidence of CDI and HAIs. Beyond HAIs, we believe the multitask approach for learning the time-dependent structure of risk factors is a promising methodology for building predicting models for other adverse outcomes.

2. Background and Related Work

In previous work, risk-stratification models for CDI considered no more than a dozen risk factors identified by experts and many of these risk factors pertained to time-invariant features (i.e., observations that do not change over the course of the hospitalization) and researchers ignored changes in patient risk over time (Tanner et al., 2009; Dubberke et al., 2011; Garey et al., 2008; Krapohl, 2011). In our work, we have shown how leveraging the entire structured contents of the EHR leads to significantly better predictions compared to a model based solely on a set of known risk factors easily extracted from the EHR (Wiens et al., 2014). Moreover, we have incorporated both time-varying (e.g., current medications) and time-invariant (e.g., gender) risk factors into the model. Dubberke et al. (2011) also consider a risk prediction model based on both variables collected at the time of admission and throughout the admission. However, they ignore any trend in patient risk. In contrast, we have investigated different methods for incorporating the evolution of patient risk into the current risk estimate, transforming the problem into a time-series classification task (Wiens et al., 2012a). These extensions lead to significant improvements in patient risk stratification.

While our previous work has touched on time-varying variables, to date, risk stratification models for CDI have only considered time-invariant model parameters. That is, although the patient changes over time, and so does the estimate of risk, the models used to compute patient risk do not. This approach does not allow the relative importance of risk factors to change over time as the patient spends more time the hospital. Models to date have not explicitly considered the time-dependence of such factors as a patient's susceptibility and exposure over time. We argue that, in addition to changes in patient state and hospital conditions, the relative importance of risk factors may change during an admission. For example, the important of evidence drawn from a patient's history may diminish as the patient spends more time in the hospital. We propose a methodology that can capture and represent such rich temporal dynamics of the relevance of risk factors in real-world healthcare settings.

Models with time-varying parameters have been studied in other contexts like survival analysis (Fan and Zhang, 2008). Over the years, standard approaches to survival analysis, like Cox proportional hazards, have been extended to include time-dependent parameters (Hastie and Tibshirani, 1993). Extensions typically involve the addition of interaction terms between features and time-varying functions (Gray, 1992; Murphy and Sen, 1991; Zucker and Karr, 1990; Tian et al., 2005; Sun et al., 2009). In many cases, the user must specify these functions. Researchers have developed non-parametric extensions, but these methods can be computationally inefficient for large, high-dimensional data sets. In practice, researchers often end up partitioning time into intervals and analyze each time period with a simple model. Our proposed approach is similar in the sense that we break the problem up into multiple tasks. However, instead of learning the models independently, we propose learning the models jointly using a multitask learning (MTL) framework.

MTL is a popular branch of machine learning that leverages the intrinsic relatedness among different tasks (Caruana, 1997). It has been studied extensively in many different applications, including healthcare (Caruana, 1996). As an example, Zhou et al. (2014) employ multitask learning in a patient risk stratification context for handling missing features values. In other work, (Zhou et al., 2011) employed an MTL framework in their work on Alzheimer's disease progression, centering on the prediction of the cognitive functioning of patients at different times in the future. They consider each time point as a different regression task and learn the tasks jointly. The authors employ a temporal group LASSO regularization framework, which ensures that only a small subset of the variables are chosen while penalizing large deviations of predictions at neighboring time points. Similarly, we treat each time point of prediction as a different task. However, instead of predicting multiple points into the future based only on the covariates at baseline, we predict risk each day based on time-varying covariates. In our application, a patient's risk of CDI is affected by several external risk factors that can change over the course of the hospitalization e.g., exposure to disease. We incorporate these changes at each time point, since ignoring them is likely to lead to inaccurate predictions.

3. Study Population

We considered all adult inpatient admissions to a large private hospital in the US over a two year period. The statistical analysis of retrospective medical records was approved by the Institutional Review Board of the Office of Research Integrity of the hospital network's

	Study Population (n=49,006)
Age, median (IQR)	60 (46-73)
Female gender, %	55.25
Hospital Service (%)	
medicine	47.85
cardiology	11.98
surgery	9.60
obstetrics	8.12
psychiatry	5.4
LOS (days), median (IQR)	5.4 (3.8-4.4)
CDI (%)	
current visit	1.02
1-year history	1.11
any history	1.54

Table 1: Demographics of our study population.

research institute. We considered patients admitted on or after 2011-04-12 and discharged on or before 2013-04-12 (n=73,454). We exclude admissions in which the patient was discharged or tested positive for *C. difficile* before the end of the third day (n=24,389) and admissions for which the patient had a positive test result for *C. difficile* within the 14 days preceding the current admission (n=59). This resulted in 49,006 unique admissions. These criteria exclude many predictable low-risk patients with shorter stays, and focus on those patients who we believe acquire the infection during the current hospital admission (as opposed to those who are already infected at the time of admission). Our final study population is described in Table 1.

CDI cases are typically defined as healthcare-associated if they occur within 48 hours of the time of admission. In our work, we exclude patients who test positive before the end of the third calendar day of admission. By defining the cutoff as the end of the third calendar day, we ensure that the minimum cutoff of 48 hours is achieved, while allowing for simultaneous predictions for every patient (at the end of the day). A uniform time of prediction makes sense from a clinical perspective, since it streamlines the risk-stratification process. Most recently, the Centers for Disease Control and Prevention (CDC) updated their definition of HAIs to include positive test results that occur on the third calendar day of admission (CDC, 2015). While we do not consider these cases here, this work could easily be extended to do so.

4. Methods

In this section, we begin with the problem setup and define notation that will be used throughout the paper. We then describe the feature extraction and learning algorithms used to train the risk prediction model.

4.1 Problem Setup & Notation

The task at hand is to learn a model to accurately predict an inpatient's risk of acquiring CDI during the current hospitalization. Predictions are made daily; we consider time at the granularity of a day, and each day t of an admission is represented by a d dimensional binary feature vector: $\mathbf{x}_t \in \{0,1\}^d$. While we focus on a binary representation of the data, we incorporate both continuous and discrete variables as discussed in Section 4.2. Since each admission in our study population consists of multiple days, the i^{th} patient admission is represented by a series of feature vectors: $(\mathbf{x}_1^{(i)}, \mathbf{x}_2^{(i)}, ..., \mathbf{x}_{m_i}^{(i)})$, where m_i varies across patient admissions since the length of a visit varies (note: $m_i \geq 3$ in our study population). We use boldface notation to denote vectors.

In addition to a series of feature vectors, each patient admission is also associated with a binary label $y \in \{+1, -1\}$. Each day of a visit in which the patient eventually tests positive is labeled $+1$ and -1 otherwise. Thus each patient admission $p^{(i)}$ consists of m_i (feature vector, label) pairs:

$$p^{(i)} = \{(\mathbf{x}_t^{(i)}, y_t^{(i)})\}_{t=1}^{m_i}$$

For patients who do not test positive, m_i is equal to one less than the length of the visit in calendar days. We do not consider the day of discharge in our analysis since by the time a patient is discharged the prediction is meaningless. For patients who eventually test positive, we consider a patient admission up to and including the day before the day of the positive test result. Finally, our data set is defined as:

$$\mathcal{D} = \{p^{(i)}\}_{i=1}^{n}$$

where n represents the number of unique patient admissions in the data set. Note that a patient may be represented multiple times in our data set if he/she has multiple admissions during the time period we consider.

Data	Description
Admission details	Admission details (e.g., date and time of admission, date and time of discharge, and type of visit) and other information pertaining to the admission such as the financial class code, the source of the admission, the hospital service, and the attending doctor are extracted for each patient admission.
Patient demographics	Information pertaining to patient demographics such as age at the time of admission, gender, race, marital status, and city of residence are extracted. Aside from age, all data in this table are categorical.
Laboratory results	Results pertaining to ordered laboratory tests are extracted. Each entry in the database table is associated with a patient admission, an observation identifier, an observation value, an observation time, a reference range (e.g., 120-200 for cholesterol) and an abnormal flag (e.g., H=high, L=low, C=critical, or empty=normal). We represent time-stamped laboratory results based on the observation identifier and the associated flag.
Diagnoses	Patient diagnoses are encoded using ICD-9 codes (NCHS, 2008). Patient visits can be associated with multiple ICD-9 codes; in our data the average visit (including outpatient visits) is associated with two distinct ICD-9 codes. ICD-9 codes, widely used for billing purposes, can get coded well after a patient is discharged (Iezzoni, 1990). For this reason, we do not use the codes associated with a patient's current visit in our model. Instead, we consider only the codes from a patient's most recent hospital admission.
Medications	Orders for medications are associated with an admission identifier, an 8-digit medication identifier, and a start/stop time. Each medication identifier is associated with a medication, a dosage and a form (e.g., in solution). Since the dosage and form are encoded in the 8-digit medication identifier, we represent patient medications using only this identifier.
Locations	For each hospital admission we have time-stamped location data. Location data refer to the patient's location within the hospital. Locations are collected at both the unit and the room level. Using these time-stamped data we can trace a patient's path through the hospital.
Vitals	Each entry in the vitals table corresponds to a visit, an observation identifier (e.g., "BPSYSTOLIC" for systolic blood pressure), an observation value, a reference range, an abnormal flag, and an observation timestamp. When extracting information about vitals for a patient we encode the observations the same way we encode laboratory results, i.e., as a concatenation of the observation identifier and an abnormal flag (e.g., "BPSYSTOLIC_H" for high systolic blood pressure).
Procedures	In the EHR, procedures are encoded using both Current Procedural Terminology (CPT) codes and ICD-9 procedure codes. Each row in the procedures table records a procedure, an admission identifier, and a procedure timestamp. Since both coding systems are used to describe procedures, in our analysis we consider both CPT and ICD-9 codes.

Table 2: Relevant information extracted from the EHR.

4.2 Feature and Label Extraction

As Paxton et al. (2013) state, there are many challenges that come with working with EHR data in research. Addressing these challenges requires careful consideration of the data and the intended application. Moreover, electronic health information systems will continue to change and therefore it is important that researchers take this into consideration when developing models based on EHR data. In our work these challenges motivated a simple, flexible, data-driven approach to extracting and representing EHR data.

4.2.1 Data Extraction

We represent each day of a patient's admission with a single feature vector, $\mathbf{x}_t$ for $t = 1...m_i$, composed of both *time-invariant* features collected at the time of admission and *time-varying* features collected over the course of each day. The time-invariant features aim to capture the baseline state of each patient while the time-varying features capture changes in patient state during the hospital admission. In the EHR, data are stored across different tables in several databases. We describe the relevant variables and how they are stored and extracted from the EHR in Table 2.

Each patient admission (i.e., encounter) is represented by a unique identifier, in addition each patient is associated with a unique identifier. These unique identifiers allow us to retrieve information across hospital databases for each admission, and across time for multiple admissions pertaining to the same patient. For each patient admission in our study population, we extract knowledge pertaining to the EHR data described above. We augment the data pertaining to the current admission with data extracted from previous admissions including diagnoses and medications.

4.2.2 Feature Engineering

The majority of the extracted data pertain to categorical features, e.g., medications or in-hospital locations. Vitals and laboratory results are also represented using categorical variables as described in Table 2. This eliminates the need to define our own cutoffs for discretization, since the cutoffs are encoded directly in the database using "reference ranges." Data pertaining to diagnoses were coded as ICD-9 codes, a hierarchical classification system with 13,000 unique codes. For our application, we do not expect that this level of granularity is informative. Diagnostic codes are used largely for billing purposes and are not timestamped, therefore their utility is limited. Given these limitations we focus on only diagnostic codes associated with the previous visit, and consider only the highest level of the codes.

We also consider a small number of continuous and discrete variables (e.g., age and statistics related to previous hospitalizations). We map all of these data to binary variables resulting in a high-dimensional feature space. Doing so allows us to later capture some of the nonlinear relationships that may be present in the data without using a nonlinear classifier. We discretize all continuous variables (except for age) using cutoffs based on quintiles from the training data.

From the laboratory data described in Table 2 and the in-hospital location data we were able to extract information regarding patient exposure to the disease throughout the hospitalization. *Colonization pressure* aims to measure the number of patients in a specific unit of the hospital colonized or infected with a particular disease. We define colonization pressure as in Wiens et al. (2014) and measure patient exposure over time based on the patient's location during the hospital admission. We measure exposure at both the hospital-wide and unit-wide level.

We focus on capturing events at the temporal resolution of a day. Thus, we do not consider the order of events within a single day. For example, we know which medications were ordered each day but not the temporal ordering of when the medications were taken. We handle cases where the same variable, e.g., blood pressure, is observed multiple times during a day by simply including all relevant values that were observed when building the daily feature vector. For our particular task this resolution suffices, though it may not be optimal.

4.2.3 Ground Truth

We label each example in our data set as either positive or negative depending on the laboratory data pertaining to positive stool tests for toxigenic *C. difficile*, as obtained during the hospital admission. Patient admissions with a positive test result are labeled positive in our data and negative otherwise. These laboratory results are time-stamped and thus we also noted the calendar day in which the patient tested positive. In the data set, patients are only tested if they exhibit symptoms. Thus, we expect the laboratory results to be highly sensitive and specific. However, since not all patients are tested every day for the disease there is a small possibility that some patients may have acquired an infection and yet were never tested.

4.3 Learning to Predict Daily Risk

We produce daily estimates of patient risk using a two-stage process as in Wiens et al. (2012a). However, in contrast to our previous work, our model in the first stage is not static but varies over the course of the hospitalization, incorporating time-varying model parameters. The first stage produces initial estimates of daily risk and the second stage incorporates risk estimates on previous days in order to capture the variation in risk over time. We explain both stages in detail below.

4.3.1 Baseline Approach

Given the set of labeled feature vectors $\mathcal{D} = \{(\mathbf{x}_t^{(i)}, y_t^{(i)})_{t=1}^{m_i}\}_{i=1}^{n}$ representing each day of a patient admission we can learn a classifier $\theta \in \mathbb{R}^d$, linear in $\mathbf{x}$, using $L2$-regularized logistic regression:

$$\min_{\theta} \frac{1}{2}\theta^T\theta + C\sum_{i=1}^{n}\sum_{t=1}^{m_i} log(1 + e^{-y_t^{(i)}\theta^T\mathbf{x}_t^{(i)}}) \tag{1}$$

Once we have learned θ we can produce an initial estimate of the i^{th} patient's risk on day t, $\hat{y}_t^{(i)} \in [0, 1]$:

$$\hat{y}_t^{(i)} = \frac{1}{1 + e^{(-\theta^T\mathbf{x}_t^{(i)})}} \tag{2}$$

This formulation simply pools all training examples together and learns a single model, θ ignoring the index t. As a patient spends additional time in the hospital, we expect the factors contributing to patient risk to change. Therefore, we extend the learning framework to produce a model that changes as the patient spends more time in the hospital.

4.3.2 Proposed Approach

We divide the problem into T learning tasks, splitting $\mathcal{D}$ into separate tasks: $\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3, ..., \mathcal{D}_T$. Where the task $\mathcal{D}_j$ is the task associated with the τ_j time period for $j = 1, 2, 3, ..., T$. The data are split into different tasks according to t, the index of the day of the visit, such that each task contains roughly an equal number of examples.

We could learn a separate model for each task independently of the others, but in doing so we would be limiting ourselves to using only $\frac{1}{T}$ of the original training data available. Moreover, the tasks themselves are related in time and thus are not independent. While the parameters may vary from one day to another, we do not expect large fluctuations in time. Therefore, we use a multitask learning framework to leverage the inherent relatedness among the different tasks and take advantage of the entire corpus of the training data.

We extend the $L2$-regularized logistic regression framework presented above to incorporate multiple tasks. We specifically chose $L2$ regularization over other regularization frameworks (e.g., L1) since many factors contributing to the risk of CDI are not well understood and effective interventions and preventative measures for reducing patient risk are still being studied. Thus we are more interested in capturing/identifying perhaps novel risk factors, than selecting a sparse subset. Moreover, given the collinearity present in the data, there's a risk that $L1$ regularization could "select-out" such known risk factors. We want to be careful that we do not exclude known risk factors in the final model. Even if other factors that are highly correlated with known risk factors remain in the model, clinicians are unlikely to use or trust a model that does not consider known risk factors. We employ domain adaptation techniques from Daumé III (2007), remapping each feature vector $\mathbf{x}_t^{(i)}$ to a feature vector $\in \{0,1\}^{d(T+1)}$ using the mapping function $\Phi(\mathbf{x}_t^{(i)})$:

$$\Phi(\mathbf{x}_t^{(i)}) = \quad [\mathbf{x}_t^{(i)}, \langle \mathbf{0} \rangle^{j-1}, \mathbf{x}_t^{(i)}, \langle \mathbf{0} \rangle^{T-j}] \quad \forall\, t \in \tau_j \quad \text{for } j = 1, 2, 3, ..., T$$

The new feature vectors consist of two copies of the original feature vector, padded with zeros $\mathbf{0} = [0_1, 0_2, 0_3, ..., 0_d]$. Here the notation $\langle \mathbf{0} \rangle^k$ represents k concatenated copies of the zero vector $\mathbf{0}$.

We then learn the regression parameters $\theta \in \mathbb{R}^{d(T+1)}$ using Equation (1) simply replacing $\mathbf{x}_t^{(i)}$ with $\Phi(\mathbf{x}_t^{(i)})$ in the objective function. One can decompose θ into $T+1$ vectors each in $\mathbb{R}^d$, the original dimensionality of the problem, $\theta = [\theta_0, \theta_1, \theta_2, ..., \theta_T]$. Here, θ_0 corresponds to a vector of shared feature weights since it is based on data from all days, where as θ_1 is based on only data from $t \in \tau_1$ and so on.

By substituting $\Phi(\mathbf{x}_t^{(i)})$ for $\mathbf{x}_t^{(i)}$ in Equation (2), the risk of patient i on day $t \in \tau_j$ becomes proportional to $(\theta_0 + \theta_j)^T \mathbf{x}_t^{(i)}$. Writing the function this way shows how learning the models jointly results in T different models all with a shared component θ_0.

$$\theta'_j = \quad \theta_0 + \theta_j \quad \text{for } j = 1, 2, 3, ..., T$$

In general, we can estimate the risk of a new patient day $\mathbf{x}_t^{(i)}$ using Equation 2, but replacing θ with θ'_j, without having to remap the feature vector.

4.3.3 Temporal Smoothing

Applying the model described above to a patient's data results in a single estimate of risk for each day $\hat{y}_t^{(i)}$ of a patient's visit. This estimate is based on both the baseline state of the patient (as captured by the time-invariant features in $\mathbf{x}_t^{(i)}$) and the measured time-varying variables. In previous work, we showed that this *snapshot* approach to measuring patient risk, ignores important information contained in the evolution of patient risk. Thus, we incorporate risk estimates from previous days using a cumulative moving average. Given the initial risk estimates, the predicted risk for patient i on day t is calculated as $risk_t^{(i)} = \frac{\hat{y}_1 + \ldots + \hat{y}_t}{t}$. This biases new estimates toward the estimates from previous days; while large fluctuations in patient risk in close temporal proximity are possible, they are unlikely. In earlier work, we considered a formulation based on a weighted average in which days closer to the current day receive more weight. However, we found these methods did not yield better estimates than a simple cumulative average (Wiens et al., 2012b). The approach considered here is equivalent to the *RP+Average* approach described in Wiens et al. (2012a), the simplest of the investigated approaches that significantly outperformed the snapshot approach.

4.4 Risk Stratification Model Evaluation - Daily Predictions

In the clinical literature, risk stratification models are often evaluated at a single point in time e.g., two days before an index event or at the time of admission (Tanner et al., 2009; Dubberke et al., 2011). Evaluating a model's ability to risk stratify patients at the time of admission is fine. However, a patient's risk is likely to change over the course of the hospitalization. Evaluating a model at a specific point in time, e.g., n days before an index event, brings to the fore two additional issues. First, such an evaluation requires you to define an index event for negative patients. For risk stratification for CDI the index event is often defined as the first positive test result for patients who test positive and the day of discharge otherwise. However, the task of distinguishing between patients about to test positive for CDI and patients about to be discharged from the hospital is relatively easy since the patients tend to look quite different in the feature space. More importantly, the ability to distinguish such patients is of little clinical utility. Second, evaluating a model at a specific point in time does not yield an accurate representation of how the model will perform in a clinical setting. While such an evaluation scheme is useful for comparing classifier performance, in practice, clinicians have no way of knowing when a patient is n days from an index event.

In a clinical setting, we expect to apply the risk stratification model to each day of a patient's visit. This results in multiple predictions for each patient admission, one corresponding to each day of the admission. We consider a model evaluation scheme that takes all of these predictions into consideration, yet still yields a single measure of performance.

One could imagine a validation scheme in which the performance of a classifier is evaluated for each day independently. While complete, this evaluation still lacks relevance from a clinical perspective since it is not clear how to interpret the utility of a classifier that correctly classifies a patient m days out of a total of n days. Here, we consider an evaluation scheme that is driven by the use case of the model. We assume that once a patient is identified as high-risk s/he will receive some form of an intervention (e.g., relocated to a private room or special stewardship of the patient's antibiotic protocol) and that this intervention will last for a certain period of time determined by the physicians treating the patient (e.g., 10 days or for the remainder of the visit). Thus, while the predicted risk remains low, we continue making daily predictions, each day deciding whether or not to intervene. However, once the predicted risk exceeds some threshold, we classify the patient as high-risk and discontinue making predictions, since the question of whether or not to intervene becomes irrelevant once physicians have intervened.

Thus, while the model's daily predictions are allowed to fluctuate (from low to high and high to low), when evaluating the model we choose a single decision threshold. We apply this decision threshold to each day of a patient's visit, up to the day before a positive test result is observed or the day before the day of discharge. If the patient's daily estimated risk ever exceeds the decision threshold they are classified as high risk, otherwise they are classified as low risk. This mimics the primary way we expect the model to be used in practice. By taking the maximum prediction for each patient and sweeping over different values of the decision threshold, we can evaluate the model in terms of the area under the receiver operating characteristic curve (AUROC). The AUROC alone is not enough to quantify the performance of the model; since there is high class-imbalance, we also consider the area under the precision recall curve (AUPR). We estimate 95% confidence intervals for these performance measures by applying a bootstrap method. Finally, in addition to these performance metrics, we also evaluate how *far in advance* we correctly identify positive cases. This is an important measure of performance that is often overlooked, but has crucial implication regarding the utility of the model in real-world clinical practice. The earlier we can identify a patient as high-risk, the earlier we can intervene, with the goal of reducing the likelihood of an adverse outcome for a patient and also reducing the spread of the pathogen.

5. Experiments and Results

Employing the methods described in the previous section, we learned and validated a risk stratification model for identifying inpatients at high-risk of acquiring an infection with *C. difficile* throughout their hospital admissions using data from our study population.

5.1 Learning the Risk Model

Our feature extraction yielded close to 10,000 binary variables for each patient day. To reduce the dimensionality, we filter out features that do not occur in at least 1% of the training set. This resulted in a lower dimensional feature vector, where each day is represented by a vector of 905 binary variables. The remaining variables are presented in Table 3. As shown in Table 3 the majority of the features pertain to laboratory results and medications, both time-varying variables.

We split the data into a training set and a holdout set based on time, training on data from the first year, and validating our model on data from the second year. The training data consisted of patient admissions from 2011-04-12 to 2012-04-11, totaling 190,675 visit days pertaining to 24,607 unique visits. Within the training data, 258 admissions had a positive test for *C. difficile* resulting in 2,608 training days with a positive label. To mitigate the influence of patients already showing symptoms, we removed patient days corresponding

Category	Feature Name	#Binary Features	Original Format
Patient History	Previous Medications	160	Categorical
	Previous Diagnoses	16	Categorical
	Number of Hospital Visits (90 days)	3	Continuous
	Avg. LOS of Hospital Visits (90 days)	6	Continuous
	Total LOS of Hospital Visits (90 days)	6	Continuous
	History of C. diff	1	Categorical
	1yr History of C. diff	1	Categorical
Patient Demographics	Age	5	Continuous
	Marital Status	5	Categorical
	Race	4	Categorical
	Gender	1	Categorical
	Financial Class	8	Categorical
	City of Residence	13	Categorical
Admission Details	Admission Month	12	Categorical
	Admission Year	3	Categorical
	Admission Type	4	Categorical
	Hospital Service	12	Categorical
	Admission Source	7	Categorical
	Attending Doctor	14	Categorical
	Expected Surgery	1	Categorical
Daily Admission Details	Laboratory Results	267	Categorical
	Medications	274	Categorical
	Vitals	24	Categorical
	Locations Units/Rooms	38	Categorical
	Procedures	4	Categorical
	Unit Exposure	6	Continuous
	Hospital Exposure	5	Continuous
	Day of Admission	5	Continuous

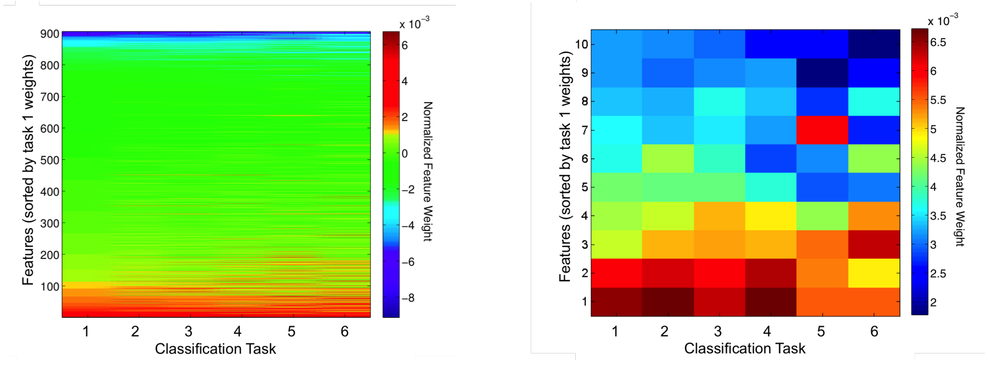
Table 3: High-level description of final features included in the model. The final feature vector consisted of 905 binary features belonging to four different categories.

to the day of and the day before the positive test result. In addition, when training the classifier, we randomly subsampled the data such that no patient contributed more than three days of the data to the training set. This is similar to reweighting samples such that each patient contributes equally to the overall classifier. If we had not done this, some patients would have been represented up to 10 times more often than other patients. Given the small number of positive examples, patients with longer visits could have significantly biased the classifier. In turn, this could have resulted in overfitting to patients with longer visits.

We selected the number of tasks T, and the corresponding temporal intervals τ_j for $j = 1, ..., T$ based on the number of training examples available for each interval. For our data, this resulted in six distinct tasks, corresponding to six distinct time periods: $\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3, \mathcal{D}_4, \mathcal{D}_5, \mathcal{D}_6$. We learned a model for each time period using $L2$-regularized logistic regression and the multitask learning framework described in Section 4.3.2. To select the hyperparameter C in (1), we performed repeated five-fold cross validation on the training data, choosing a setting that maximized the AUROC. The model parameters, i.e., θ, were solved for using LIBLINEAR (Fan et al., 2008). This resulted in six different models, $\theta'_t \in \mathcal{R}^{905}$:

$$\theta'_t = \begin{cases} \theta_0 + \theta_1 & : t \in [1] \\ \theta_0 + \theta_2 & : t \in [2] \\ \theta_0 + \theta_3 & : t \in [3] \\ \theta_0 + \theta_4 & : t \in [4, 5] \\ \theta_0 + \theta_5 & : t \in [6, 9] \\ \theta_0 + \theta_6 & : t \in [10, \infty) \end{cases}$$

θ'_t is allowed to deviate from the shared model θ_0 based on the data collected during each time period. Thus, the relative importance of risk factors is allowed to vary over time. Figure 1 (a) shows the extent to which the weights vary across time. The columns correspond to the different time periods (i.e., tasks) and the rows correspond to the different features. The features are sorted in descending order according to their weight on the first day. The color of each cell is related to the weight of the corresponding feature for the specified task, normalized by the sum of the absolute value of all weights for that task. All cells corresponding to features that have high positive weight are red and those with high negative weight are dark blue. If the relative importance of the weights remains constant over time, each column would appear identical to the first column (i.e., solid horizontal bands of color). However, as Figure 1 (a) shows, the relative importance of features changes. Still, as we expected, we see great deal of continuity in the feature ranking across time periods as all of the models share a common component.



(a) All features

(b) The top ten features

Figure 1: The changing relative importance of features over time. For each time period τ_j for $j = 1...6$ (i.e., task), the features are ranked according to the feature weight for the task associated with τ_1. The color represents the normalized feature weight for each task.

Figure 1 (b) shows that, even among the top ten features (from θ'_1), there are changes in the relative importance of features over time. In Figure 1 (b) the first two features become less important over time, while the third feature becomes more important. Table 4 lists the five features with the greatest positive weight in θ'_1 through θ'_6. Note that initially the most important feature is the patient's one year history of CDI. As a patient spends more time in the hospital, this feature loses importance relative to the location of the patient in the hospital.

Ranking	1	2	3	4	5
θ'_1	1 Yr. hist. of CDI	Hist. of CDI	Daily Units:XX	Temp.:High	Prev. Meds:Sevelamer
θ'_2	1 Yr. hist. of CDI	Hist. of CDI	Daily Units:XX	Temp.:High	Service:MED
θ'_3	1 Yr. hist. of CDI	Hist. of CDI	Daily Units:XX	Temp.:High	Prev. Meds:Sevelamer
θ'_4	1 Yr. hist. of CDI	Hist. of CDI	Daily Units:XX	Temp.:High	Mean Platelet Vol.:normal
θ'_5	Meds: Pantoprazole	1 Yr. hist. of CDI	Daily Units:XX	Hist. of CDI	Mean Platelet Vol.:normal
θ'_6	Daily Units:XX	1 Yr. Hist. of CDI	Temp.:High	Hist. of CDI	Service:MED

Table 4: We show features with greatest weight for tasks 1 through 6, using color to highlight certain trends.

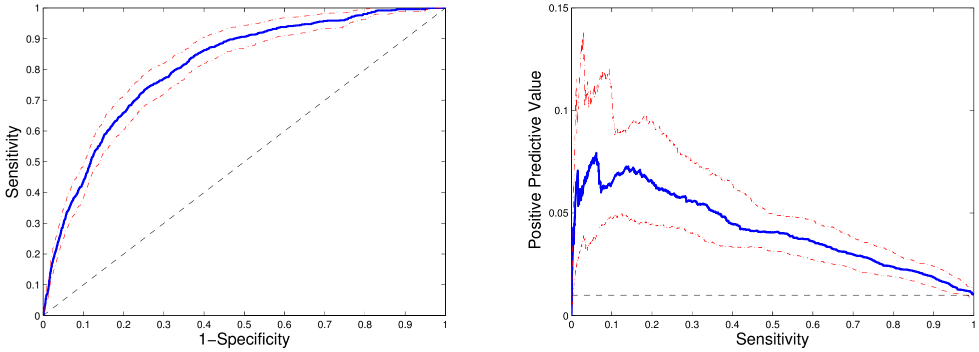
In Table 5, we note the 25 features with the greatest weight according to the shared model i.e., θ_0. We are not surprised that patient history of CDI appears at the top of this list. The medications that appear in Table 5 include drugs administered to patients receiving kidney dialysis, drugs for the treatment of high blood pressure and heart disease, and proton pump inhibitors. When interpreting these weights, it is important to keep in mind that many of the features in our model are highly correlated. These features may be directly or indirectly linked with an increased risk of CDI. Further analysis could generate hypotheses about causal relationships that could be tested in a randomized controlled trial.

Rank	Feature Index	Feature Name	Feature Description	Shared Weight
1	905	OneYear_History	positive test for toxigenic C. diff in past year	0.2472
2	904	All_History	positive test for toxigenic C. diff ever in the past	0.2314
3	124	daily_units:XX	medicine patient care unit	0.2084
4	427	daily_vitals:temporal_h	temperature oral high	0.1885
5	867	daily_meds:63604250	pantoprazole 40mg Inj	0.1508
6	44	v_hospital_service:MED	medicine hospital service	0.1466
7	605	prev_meds:63713135	sevelamer 800 mg Tab	0.1418
8	674	daily_meds:63715254	vitamin B comp w/C, FA Tab	0.1321
9	234	daily_labs:wbc_h	white blood cell count high	0.1320
10	475	prev_meds:63616924	vancomycin 1 gm/250 mL NaCl 0.9%	0.1177
11	269	daily_labs:mpv_	mean platelet volume measured	0.1175
12	433	daily_vitals:bgnas_	oxygen flow rate (nasal cannula)	0.1158
13	418	daily_labs:bun_h	blood urea nitrogen high	0.1094
14	74	attendingdoctornumber:xxxx	attending (anonymized)	0.1088
15	615	prev_meds:63708390	lisinopril 10 mg Tab	0.1081
16	590	prev_meds:63715676	zolpidem 5 mg Tab	0.1077
17	434	daily_vitals:tempax_	temperature axillary	0.1075
18	292	daily_labs:k_l	Potassium Lvl low	0.1059
19	587	prev_meds:63715254	vitamin B comp w/C, FA Tab	0.1038
20	591	prev_meds:63744403	pharmacy comment	0.1037
21	441	daily_vitals:bgpet_l	end tidal CO2 low	0.1036
22	506	prev_meds:63608947	metroNIDAZOLE 500 mg/100 mL 0.9% NaCl	0.1024
23	719	daily_meds:63713259	simvastatin 40 mg Tab	0.1007
24	267	daily_labs:phos_l	Phosphorus Lvl low	0.1003
25	688	daily_meds:63707897	K phos-Na phos Oral Pwdr	0.0990

Table 5: Features with greatest "shared" weight. Hospital unit and staff identifiers have been removed.

5.2 Evaluating the Model

In previous work, risk stratification models have been evaluated at a single point in time. Here, we employ a more realistic evaluation methodology (as described above). We believe this evaluation provides a more accurate representation of how the model will perform in a clinical setting.



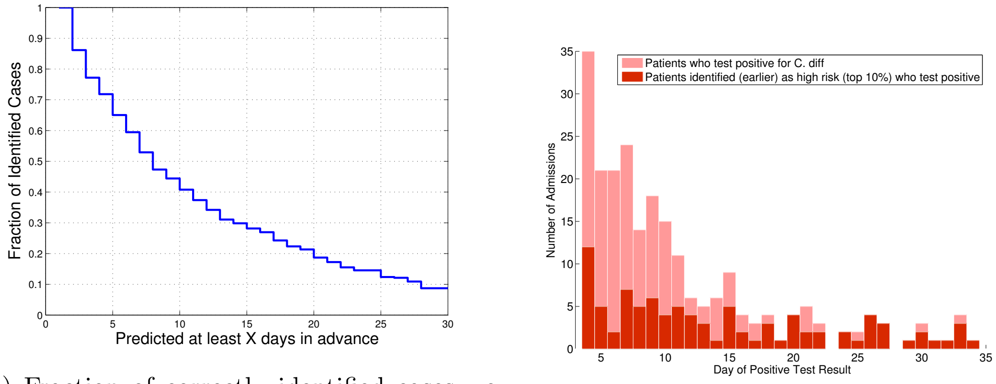
(a) Receiver Operating Characteristic Curve

(b) Precision Recall Curve

Figure 2: We plot two performance curves generated by applying the risk stratification method described in the previous section to the set of held-out patient admissions. We achieve an AUROC of 0.81 (95%CI 0.78-0.84) and a AUPR of 0.04 (95%CI 0.03-0.05). The 95% CI are represented by the red dashed lines. The performance of a random classifier is given by the black dashed lines. (A classifier no better than random achieves an AUROC=0.5 and AUPR=0.01)

We applied the model to the set of patient admissions held out for validation, which consisted of patient admissions from 2012-04-12 to 2013-04-12 and was composed of 24,399 admissions of which 242 had a positive test result for *C. difficile*. When validating the model, we did not subsample the test data as we did with the training data. Each patient admission in the held out set has at least three daily predictions of risk, since we considered only patients who were still present in the hospital at the end of the third day. However, we do not start applying the decision threshold until the end of the third day (given our exclusion criteria). Also, we do not evaluate the model on the final day of the admission (the day in which the patient was discharged) as such a prediction would be meaningless for guiding in-hospital interventions. Recall that the task is to predict a patient's probability of acquiring CDI during the current hospitalization and that predictions are made at the end of the day.

Figure 2 shows the ROC curve, and the precision recall curve for the held out patient admissions. Applied to the validation data, our model results in an AUROC of 0.81 (95%CI 0.78-0.84) and an area under the precision recall curve (AUPR) of 0.04 (95%CI 0.03-0.05). Although the AUPR appears low, the performance is significantly better than a baseline classifier (Precision=0.01), since the incidence of infection is approximately 1% in the study population.



(a) Fraction of correctly identified cases we can predict at least x days in advance of the patient testing positive for CDI.

(b) Distribution of time to positive test from admission (in days)

Figure 3: Classification performance resulting from a classifier based on the 90^{th} percentile.

To evaluate the ability of our model to distinguish high-risk patients from low-risk patients in the held out set set of patient admissions, we selected a decision threshold based on the 90^{th} percentile. We chose this cutoff to limit the number false positives, since CDI cases are relatively rare. Given this decision threshold, we correctly identify 103 patients out of 242 as high risk, and achieve a sensitivity of 0.43, a positive predictive value of 0.04, an F-score 0.08, and an odds ratio of 6.67 (confusion matrix TP=103 TN=21,820 FN=139 FP=2337). Lowering the decision threshold will increase the sensitivity and increase the number of false positives. Ultimately, the choice of decision threshold depends on the expected costs and benefits of the intervention that one intends to apply to high-risk patients.

Figure 3(a) illustrates how far in advance we can predict positive test results. We note that in approximately half of the cases correctly identified, we identify cases at least 7 days in advance of their testing positive for CDI. Figure 3(b) illustrates *when* patients are testing positive. While we identify more patients who test positive earlier, the *fraction* of patients we correctly identify increases as the length of stay increases.

Finally, in Figure 4, we illustrate the performance of our model (the *Multi-Task Joint* approach) compared to that of a model learned by simply pooling all the data (the *Single-Task* approach), in terms of the AUROC. In addition, we consider a third approach, *Multi-Task Independent*, that also learns multiple models, one for each day, but where the optimization is performed independently for each model. Applied to all patients in the held out set of patient admissions (risk period>3 days), the proposed approach, *Multi-Task Joint*, and the *Single-Task* classifier perform almost identically, with both achieving an AUROC of 0.81. In contrast, the *Multi-Task Independent* performs worse, achieving an AUROC of 0.80. We believe this reduction is due to the inability of the method to leverage all of the data as in the *Multi-Task Joint* approach. When we divide the patients into subsets based on their risk period the difference between the *Single-Task* and *Multi-Task* approaches be-

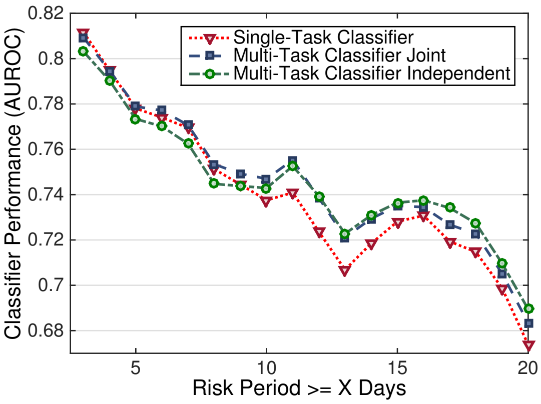


Figure 4: We compare a single-task classifier, where the model is time-invariant, to a multitask classifier where the model varies over time. The difference between the two classifiers is apparent for those patients who test positive later during the admission. These are the patients we are more interested in.

comes apparent. For patients with a longer risk period, the *Multi-Task* approach results in an improvement in performance over the *Single-Task* approach. While the difference is small, it is consistent. This difference is relevant, since the potential to intervene in a timely manner is greater for patients who test positive later in a visit. Therefore, the ability to identify such cases accurately is of considerable practical importance.

6. Discussion and Conclusion

In summary, we presented a novel data-driven patient risk stratification model for CDI that utilizes the entire structured contents of the EHR. The main contributions of our work are a novel approach to learning time-varying parameters and an evaluation scheme that is representative of how the model will be applied in practice.

The model provides estimates of patient risk daily based on time-varying parameters. We propose a multitask learning framework to efficiently learn the time-varying model. By learning the models jointly, we leverage the inherent relatedness among the different tasks. We observed changes in the ranking of features over time, suggesting that the proposed method could be used to further investigate the temporal effects of risk factors over the course of a hospital admission. These findings may shed new light on the relationship between risk factors and time, and in turn improve our understanding of the disease.

Furthermore, we demonstrated a consistent improvement in the classification performance using our multitask approach over a single-task approach, particularly among patients with longer risk periods. While consistent, the difference was not significant; this may be due to the fact that the number of cases with longer risk periods is small. We have approximately 5,000 visits in the heldout set with a risk period greater than 10 days, and only a small fraction of those patients end up testing positive. Additionally, in our formulation, we consider the same decision threshold every day. However, use of a variable decision threshold could lead to better results. Furthermore, our results showed a clear overall improvement when learning the models jointly versus independently. However, on patients with very long risk periods e.g., greater than 15 days, the independent classifiers appear to perform slightly better. This may be an indication that the averaging or smoothing effect of the joint optimization approach is perhaps too strong. This may be addressed by only penalizing differences between successive models. In future work, such models could be improved by considering additional temporal smoothness constraints.

We proposed a new evaluation scheme motivated by a clinical use case in which the model is used daily to evaluate patients in the hospital and a decision is made about whether or not to intervene. Prior to this work, such models have only been evaluated at a single point in time during a hospital visit (e.g., at the time of admission or n days before the index event). We argue that evaluating the model over the entire course of the admission is a more accurate approximation of how the model would be used and would perform in practice. We also evaluated our final model in terms of how many days in advance we could predict high-risk cases. In a clinical setting, how far in advance one can predict an infection is as relevant as traditional performance metrics like sensitivity and specificity. This approach to evaluation is independent of both the disease and our method of building a classifier. With the increasing interest in applying machine learning to clinical problems, ensuring that evaluation criteria are clinically relevant will be of great importance.

In addition, we were careful to split our data temporally, learning on data from one year and evaluating the model on data from the next year. Over time, hospital populations, physical layouts, clinical protocols, and staff can change. Furthermore, EHR systems can change both in terms of what is collected and the precise meanings of variables. Thus, when evaluating predictive models in medicine, it is important to ensure that all examples in the training set precede all examples in the set held out for validation. Failure to do this can produce misleading results since future changes may be captured by the training set. To this end, we expect the predictive performance of our model will deteriorate over time, if such changes are not readily incorporated. An important future direction of study is how such models transfer across time, and best practices for building models that incorporate not only changes but also take into consideration possible interventions.

The models learned in this work are based on hospital-specific data, and thus may not generalize to hospitals with very different patient populations or clinical protocols. However, we expect the *methods* to generalize beyond this specific hospital. Our data-driven approach to feature engineering can be applied in straightforward manner to the structured contents of any hospital database. Such a data-driven approach can readily incorporate hospital-specific features resulting in more accurate predictive models. Here, we considered only the structured contents of the electronic health record. Future models, however, could incorporate features extracted from clinical notes or even genomic data pertaining to the microbiome of patients.

We focused on developing predictive models for CDI. However, our contributions extend to building models for other types of healthcare-associated infections and other patient populations. Once incorporated into the hospital workflow, risk stratification models, like the one presented here have the potential to reduce the incidence of adverse patient outcomes. Widespread development and use of such data-driven models promise to enable cost-effective, targeted prevention strategies that will ultimately improve patient care.

References

A Aboukhalil, L Nielsen, M Saeed, R Mark, and G Clifford. Reducing false alarm rates for critical arrhythmias using the arterial blood pressure waveform. *Journal of Biomedical Informatics*, 41(3):442–451, 2008.

R Caruana. Algorithms and applications for multitask learning. In *International Conference on Machine Learning (ICML)*, pages 87–95, 1996.

R Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

CDC. Identifying healthcare-associated infections (HAI) for NHSN surveillance, January 2015.

CC Chia, I Rubinfeld, B Scirica, S McMillan, H Gurm, and Z Syed. Looking beyond historical patient outcomes to improve clinical models. *Science Translational Medicine*, 4(131):131ra49–131ra49, 2012.

H Daumé III. Frustratingly easy domain adaptation. In *Association for Computational Linguistics (ACL)*, volume 1785, page 1787, 2007.

C DeSantis, J Ma, and A Bryan, Land Jemal. Breast cancer statistics, 2013. *CA: a Cancer Journal for Clinicians*, 64(1):52–62, 2014.

E Dubberke, A Wertheimer, et al. Review of current literature on the economic burden of *Clostridium difficile* infection. *Infection Control and Hospital Epidemiology*, 30(1):57–66, 2009.

E Dubberke, Y Yan, K Reske, A Butler, J Doherty, V Pham, and V Fraser. Development and validation of a *Clostridium difficile* infection risk prediction model. *Infection Control and Hospital Epidemiology*, 32(4):360–366, 2011.

J Fan and W Zhang. Statistical methods with varying coefficient models. *Statistics and Its Interface*, 1(1):179, 2008.

RE Fan, KW Chang, CJ Hsieh, XR Wang, and CJ Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

K Garey, T Dao-Tran, Z Jiang, M Price, L Gentry, and H DuPont. A clinical risk index for *Clostridium difficile* infection in hospitalized patients receiving broad-spectrum antibiotics. *Journal of Hospital Infection*, 70(2):142–147, 2008.

RJ Gray. Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *Journal of the American Statistical Association*, 87(420): 942–951, 1992.

T Hastie and R Tibshirani. Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 757–796, 1993.

LI Iezzoni. Using administrative diagnostic data to assess the quality of hospital care: pitfalls and potential of ICD-9-CM. *International Journal of Technology Assessment in Health Care*, 6(02):272–281, 1990.

D Kansagara, H Englander, A Salanitro, D Kagen, C Theobald, M Freeman, and S Kripalani. Risk prediction models for hospital readmission: a systematic review. *Journal of the American Medical Association*, 306(15):1688–1698, 2011.

S Kleinberg and G Hripcsak. A review of causal inference for biomedical informatics. *Journal of Biomedical Informatics*, 44(6):1102–1112, 2011.

RM Klevens, J Edwards, C Richards, T Horan, R Gaynes, D Pollock, and DM Cardo. Estimating health care-associated infections and deaths in us hospitals, 2002. *Public Health Reports*, 122(2):160, 2007.

G Krapohl. Preventing health care-associated infection: Development of a clinical prediction rule for *Clostridium difficile* infection. PhD Thesis, 2011.

SS Magill, JR Edwards, W Bamberg, ZG Beldavs, G Dumyati, MA Kainer, R Lynfield, M Maloney, L McAllister-Hollod, J Nadle, et al. Multistate point-prevalence survey of health care–associated infections. *New England Journal of Medicine*, 370(13):1198–1208, 2014.

B Miller, L Chen, Dl Sexton, and D Anderson. Comparison of the burdens of hospital-onset, healthcare facility–associated *Clostridium difficile* infection and of healthcare-associated infection due to methicillin-resistant staphylococcus aureus in community hospitals. *Infection Control and Hospital Epidemiology*, 32(4):387–390, 2011.

SA Murphy and PK Sen. Time-dependent coefficients in a Cox-type regression model. *Stochastic Processes and their Applications*, 39(1):153–180, 1991.

NCHS. International classification of diseases, 9th revision, clinical modification. National Center for Health Statistics. http://icd9cm.chrisendres.com, 2008.

C Paxton, A Niculescu-Mizil, and S Saria. Developing predictive models using electronic medical records: Challenges and pitfalls. In *Americal Medical Informatics Association (AMIA) Symposium*, 2013.

J Pépin, ME Alary, L Valiquette, E Raiche, J Ruel, K Fulop, D Godin, and C Bourassa. Increasing risk of relapse after treatment of *Clostridium difficile* colitis in Quebec, Canada. *Clinical Infectious Diseases*, 40(11):1591–1597, 2005.

M Saeed, M Villarroel, A Reisner, G Clifford, LW Lehman, G Moody, Ts Heldt, T Kyaw, B Moody, and RG Mark. Multiparameter intelligent monitoring in intensive care ii (MIMIC-II): a public-access intensive care unit database. *Critical Care Medicine*, 39(5): 952, 2011.

S Saria, A Rajani, J Gould, D Koller, and A Penn. Integration of early physiological responses predicts later illness severity in preterm infants. *Science Translational Medicine*, 2(48):48ra65–48ra65, 2010.

A Shoeb and JV Guttag. Application of machine learning to epileptic seizure detection. In *International Conference on Machine Learning (ICML)*, pages 975–982, 2010.

D Sievert, P Ricks, J Edwards, A Schneider, J Patel, A Srinivasan, A Kallen, B Limbago, and S Fridkin. 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, 2009–2010. *Infection Control and Hospital Epidemiology*, 34(1):1–14, 2013.

Y Sun, R Sundaram, and Y Zhao. Empirical likelihood inference for the Cox model with time-dependent coefficients via local partial likelihood. *Scandinavian Journal of Statistics*, 36(3):444–462, 2009.

Z Syed and I Rubinfeld. Unsupervised risk stratification in clinical datasets: Identifying patients at risk of rare outcomes. In *International Conference on Machine Learning (ICML)*, pages 1023–1030, 2010.

Z Syed, C Stultz, B Scirica, and J Guttag. Computationally generated cardiac biomarkers for risk stratification after acute coronary syndrome. *Science Translational Medicine*, 3 (102):102ra95–102ra95, 2011.

J Tanner, D Khan, D Anthony, and J Paton. Waterlow score to predict patients at risk of developing *Clostridium difficile*-associated disease. *Journal of Hospital Infection*, 71(3): 239–244, 2009.

L Tian, D Zucker, and LJ Wei. On the Cox model with time-varying regression coefficients. *Journal of the American Statistical Association*, 100(469):172–183, 2005.

C Umscheid, Al Rajender, W Kendal, P Brennan, et al. Estimating the proportion of healthcare-associated infections that are reasonably preventable and the related mortality and costs. *Infection Control and Hospital Epidemiology*, 32(2):101–114, 2011.

J Wiens, J Guttag, and E Horvitz. Patient risk stratification for hospital-associated c. diff as a time-series classification task. In *Neural Information Processing Systems (NIPS)*, 2012a.

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

J Wiens, WN Campbell, ES Franklin, JV Guttag, and E Horvitz. Learning data-driven patient risk stratification models for clostridium difficile. In *Open Forum Infectious Diseases*, volume 1, page ofu045. Oxford University Press, 2014.

DS Yokoe, LA Mermel, DJ Anderson, KM Arias, H Burstin, DP Calfee, SE Coffin, ER Dubberke, V Fraser, DN Gerding, et al. Executive summary: a compendium of strategies to prevent healthcare-associated infections in acute care hospitals. *Infection Control*, 29 (S1):S12–S21, 2008.

J Zhou, L Yuan, J Liu, and J Ye. A multi-task learning formulation for predicting disease progression. In *Conference on Knowledge Discovery and Data Mining (SigKDD)*, pages 814–822. ACM, 2011.

Jiayu Zhou, Fei Wang, Jianying Hu, and Jieping Ye. From micro to macro: data driven phenotyping by densification of longitudinal electronic medical records. In *Conference on Knowledge Discovery and Data Mining (KDD)*, pages 135–144. ACM, 2014.

D Zucker and AF Karr. Nonparametric survival analysis with time-dependent covariate effects: a penalized partial likelihood approach. *Annals of Statistics*, pages 329–353, 1990.