2015 International Conference on Healthcare Informatics

Improving risk prediction of Clostridium Difficile Infection using temporal event-pairs

Mauricio Monsalve*, Sriram Pemmaraju*, Sarah Johnson†, Philip M. Polgreen‡
*Department of Computer Science,
The University of Iowa, Iowa City, IA 52242
{mauricio-monsalve, sriram-pemmaraju}@uiowa.edu
†Global Health Sciences,
The Medicines Company, Parsippany, NJ 07054
sjjohn25@mchsi.com
‡Department of Internal Medicine,
The University of Iowa, Iowa City, IA 52242
philip-polgreen@uiowa.edu

***Abstract*—Clostridium Difficile Infection (CDI) is a contagious healthcare-associated infection that imposes a significant burden on the healthcare system. In 2011 alone, half a million patients suffered from CDI in the United States, 29,000 dying within 30 days of diagnosis. Determining which hospital patients are at risk for developing CDI is critical to helping healthcare workers take timely measures to prevent or detect and treat this infection. We improve the state of the art of CDI risk prediction by designing an ensemble logistic regression classifier that given partial patient visit histories, outputs the risk of patients acquiring CDI during their current hospital visit. The novelty of our approach lies in the representation of each patient visit as a collection of co-occurring and chronologically ordered pairs of events. This choice is motivated by our hypothesis that CDI risk is influenced not just by individual events (e.g., being prescribed a first generation cephalosporin antibiotic), but by the temporal ordering of individual events (e.g., antibiotic prescription followed by transfer to a certain hospital unit). While this choice explodes the number of features, we use a randomized greedy feature selection algorithm followed by BIC minimization to reduce the dimensionality of the feature space, while retaining the most relevant features. We apply our approach to a rich dataset from the University of Iowa Hospitals and Clinics (UIHC), curated from diverse sources, consisting of 200,000 visits (30,000 per year, 2006-2011) involving 125,000 unique patients, 2 million diagnoses, 8 million prescriptions, 400,000 room transfers spanning a hospital with 700 patient rooms and 200 units. Our approach to classification produces better risk predictions (AUC) than existing risk estimators for CDI, even when trained just on data available at patient admission. It also identifies novel risk factors for CDI that are combinations of co-occurring and chronologically ordered events.**

I. Introduction

Clostridium Difficile Infection (CDI) is a healthcare-associated infection that is a major cause of morbidity and is associated with significant healthcare costs. In 2011 alone, half a million patients suffered from CDI in the United States, and 29,000 died within 30 days of diagnosis [1]. Moreover, CDI is increasing: there were only an estimated 139,000 cases in 2000. Furthermore, mortality from CDI has increased at an even greater rate: as recently as 2000, there were only 3,000 CDI-related deaths [2].

Predicting which patients will develop CDI could help to confirm cases more quickly and perhaps prevent outbreaks by helping to determine when to isolate infected patients. Several researchers have developed models for predicting CDI cases using medical records. Dubberke et al introduced a model for predicting CDI for any patient admitted to the hospital [3]. Garey et al developed a CDI-prediction model, but only for patients receiving broad-spectrum antibiotics [4]. Wiens et al. have built models that compute evolving risk scores for patients [5, 6], and another model that only uses data that were available within 24 hours of a patient's admission to the hospital [7]. In addition, there are simplified risk scores that can be calculated using only 4-5 features [8].

Predicting outcomes from medical records is difficult for a number of reasons. Medical records consist a variety of data types [9], and medical records often contain inaccuracies, biases and censoring [10]. In addition, medical records evolve over time. Some researchers have mined frequent events or common patterns of clinical events using partial orders [11, 12], sequences [13], and temporal abstractions [14], but these approaches have shortcomings as well, including inflexible representations of time and poor interaction between the classifier and the patterns discovered [15, 16].

In this manuscript, we propose a methodology for predicting patient-level CDI by mining clinical events that occur during the hospital stay as well as information that is known at the time of admission. Our classifier consists of an ensemble of logistic regression models, as in [17], fitted with regularization, as in [18]. The novelty of our approach, however, lies in the description of the visit using co-ocurring and chronologically ordered pairs of events, a simplification of the partial order patterns used in [11, 12]. The contributions of our work to the literature are multiple. First, our method performs better at predicting CDI than alternative methods. Second, we were able to produce moving risk curves for CDI, meaning that the risk predicted by our classifiers increases if the patient is going to develop CDI in the following days. Third, by representing patient visits as pairs of events, our classifier returns human-interpretable data. Fourth, we showed how to

978-1-4673-9548-9/15 $31.00 © 2015 IEEE
DOI 10.1109/ICHI.2015.24

successfully make use of hierarchical relations among the features to produce more informative models. And fifth, we contribute with a methodology for building classifiers for the situation of class imbalance and high dimensionality.

II. Overview of CDI

CDI is a hospital-associated infection that causes diarrhea in affected patients. CDI is caused by Clostridium difficile (c. diff.), a bacterium that is commonly present in the intestines of healthy people [2, 19]. Antimicrobials disrupt the normal flora in the intestine allowing c. diff. to grow and produce toxins that result in CDI. Risk factors for CDI include antibiotic use, especially clindamycin, cephalosporins and quinolones [20, 21], advanced age [21], and underlying severity of illness [3, 20].

C. diff. and its spores can be spread from one patient to another via the hands of healthcare workers and by contact with the hospital environment. Indeed, c. diff. spores can survive in the healthcare environment for long periods of time and are resistant to most cleaning agents [22]. Long hospital stays are associated with CDI [21]. Also, staying a room previously occupied by a CDI patient, or staying on a unit with a high rate of CDI have also been associated with CDI [3, 23, 24]. However, recent work has suggested that symptomatic cases are often not genetically related, at least during non-outbreak periods, raising fundamental questions about the scientific knowledge with respect to c.diff, particularly about the role of spread within hospitals [25, 26].

III. The data

A. Structured medical records data

Our data consist of HIPAA-compliant anonymized medical records representing patient-visit data (i.e., *encounter* data), containing diagnoses, prescriptions, procedures, etc., associated with each admission from October 2006 to December 2011 at the University of Iowa Hospitals and Clinics (UIHC), collected from billing data, patient flow data, and architectural drawings. We collected data on 208,902 visits involving 126,265 distinct patients. Visits consist of two types of data: (i) general visit data, and (ii) clinical event data. *General visit data* include patient demographics, visit information, service information, attending healthcare workers, and diagnoses. Patient demographics include information such as age, gender, ethnicity, and the zip code where the patient resided at the time of admission. Visit information includes the admission and discharge dates, the type and source of admission, and the disease severity and mortality. Service information refers to the service providing most of the care (e.g., psychiatry, dermatology, etc.). Attending healthcare workers include the lead physician and assistants. Diagnoses contain the conditions present on admission, those that developed during care, and those of unclear onset (it is not always possible to accurately determine when a disease originated).

Clinical event data represent what occurred during the visit. We consider four types of event data: prescriptions, which associate a patient with a medication; procedures, which associate a patient with an *operation* and the associated physician; transfers, which associate a patient with a physical location; and positive CDI laboratory tests, which tell us when healthcare workers identified that the patient had CDI. Table I presents sample visit statistics.

Item	*Avg*	*Range*
Age (years)	43.27	0–105
Length of stay (days)	5.83	1–462
Room transfers	2.04	0–102
Diagnoses	7.41	0–40
present on admission†	4.13	0–35
acquired during visit†	0.64	0–19
Prescriptions	37.28	0–5513
unique medications	9.51	0–107
Procedures	2.78	0–26
unique procedures	2.73	0–18
Physicians	2.94	1–15

† Not all diagnoses can be classified in either category.

Table I: Statistics of patient visits.

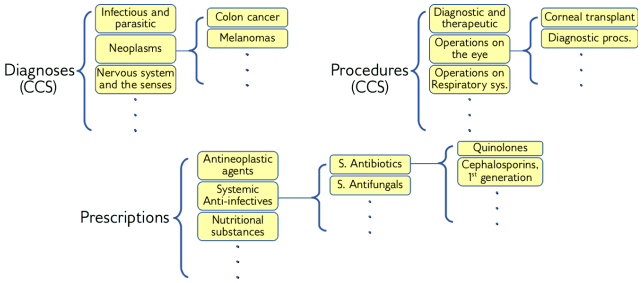


Figure 1: Partial view of the hierarchies associated with diagnoses, procedures, and prescriptions.

Diagnoses, procedures, and prescriptions are associated with hierarchies, as illustrated by Fig. 1. Diagnoses and procedures are categorized into ICD-9 and CCS codes. We prefer CCS codes [27], which group ICD-9 codes by similarity. CCS codes are grouped by chapters, providing a natural ontology. Prescriptions are also associated with hierarchies. Each prescription is associated with a medication, which in turn belongs in a three level-hierarchy comprised of: major class, minor class, and subminor class. We describe medications through the subminor class, because we deemed the *medication id* to be unnecessarily specific.

B. CDI at the hospital

Preliminary analysis of our data confirms general knowledge of types of patients at high risk for CDI and typical clinical management of CDI. Fig. 2 shows the rate of CDI in patients by age group, in bins of 5 years. CDI cases tend to be more frequent for the elderly, as consistent with the literature [2, 19]. Community-acquired CDI, i.e., present on admission to the hospital, totals 486 cases (26.25% of visits associated with CDI). The fact that CDI at admission corresponds roughly to $1/5$ of the cases, suggests that other cases of CDI may be caused by acquisition of c.diff outside the hospital as well. Table II shows the most common comorbidities, antibiotics prescribed and procedures performed on patients diagnosed with CDI. The conditions presented in Table IIa are known to co-occur with CDI [28]. But the presence of CDI introduces changes in their care: antibiotic prescription increases roughly 2.7 times after the development of CDI. Table IIb shows the most common antibiotics prescribed to patients who develop

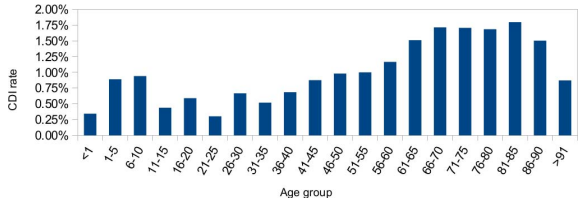


Figure 2: Proportion of admissions that result in CDI in the UIHC, stratified by age at time of admission.

Condition/diagnosis	*Frequency*
Acute kidney failure, unspecified†	245
Acidosis†	178
Unspecified septicemia	155
Acute respiratory failure	121
Pneumonia, organism unspecified	120
Unspecified pleural effusion	113
Septic shock	99
Congestive heart failure, unspecified	91
Unspecified protein-calorie malnutrition†	88
Hyposmolality and/or hyponatremia†	71

† Associated with diarrhea and intestinal failure, consistent with CDI [29].

(a) Ten most frequent diagnoses of CDI patients.

	Frequency	
Antibiotic name	*Before CDI*	*After CDI*
Metronidazole (systemic)	668	5879
Vancomycin	1120	4354
Piperacillin/tazobactam	777	1254
Ciprofloxacin (systemic)	641	1113
Cefepime	555	1018

(b) Five most frequent antibiotics prescribed to patients after CDI.

	Frequency	
Procedure name	*Before CDI*	*After CDI*
Injection of antibiotic†	228	134
Transfusion of packed cells	110	97
Computerized axial tomography of abdomen†	22	53
Parenteral infusion of nutritional substances†	42	45
Transfusion of platelets	48	44

† Procedures consistent with occurrence of CDI.

(c) Five most frequent procedures performed on patients after CDI.

Table II: Most frequent diagnoses, antibiotics, and procedures associated with patients that suffered CDI during their visit.

CDI, both before and after the diagnosis. Metronidazole and vancomycin are associated with the largest increases in prescription rates after the diagnosis of CDI. Three of the five procedures shown in Table IIc are consistent with treatment of CDI. Note that the frequency of antibiotic injections decreases after CDI develops, which is consistent with the switch to the *oral* route for the treatment of CDI.

Fig. 3 provides a simplified description of a real case of CDI in the hospital. A child is admitted to the hospital for a surgery intended to repair the aorta. Shortly upon admission, the child is sent to the surgery waiting room, given blood medications, anesthesia, and Cephalosporin antibiotics. The child undergoes surgery, and is sent to the pediatric ICU with a catheter and oxygen. The next day, the child starts a regimen of histamine$_2$ antagonists and diuretics, medications which can treat nausea and dehydration symptoms. Day 6 arrives, the laboratory confirms CDI, and the child is treated

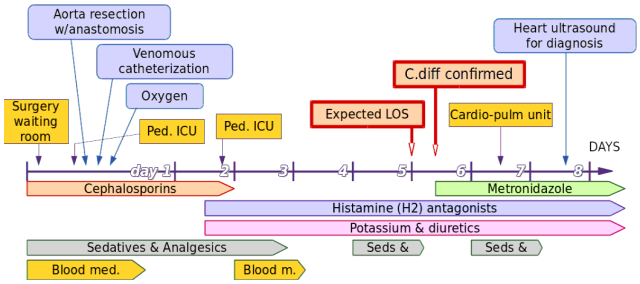


Figure 3: A case of CDI in the hospital. A child is admitted to the hospital for a scheduled cardiac surgery and develops CDI while in care.

with Metronidazole. This treatment appears to be successful as the child leaves the pediatric ICU the following day and is discharged one day later (day 8).

IV. Feature Engineering

A. Overview of the classification approach

Our approach to classification consists in converting the original data, which mostly consists of temporal, event data, into a static equivalent that can be described in tabular format, as shown in Fig. 4a, as has been previously reported [5–7, 11–14, 30–33]. Our instances consist of **days in a visit**, not of whole visits. After describing the visits in tabular format, we pass these data to the classifier. Our classifier, described in Sec. V, consists of an ensemble of logistic regression models. The model produced by the classifier consists of the collection of individual logistic regression models. The risk estimate of CDI is, then, computed from the probabilities generated by the logistic regression models.

B. Bare events and pairs of events

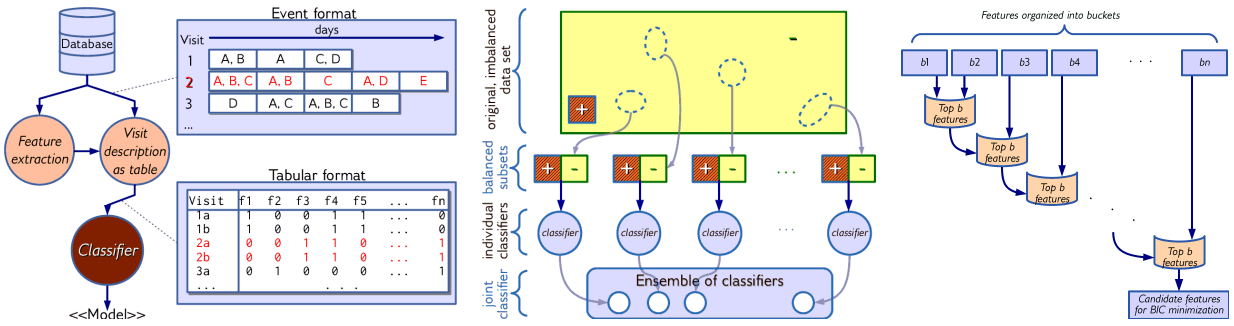
We now formally define the notion of events, visits, ordered pairs of events, and translate these into features.

Definition 1. We define a *visit* V as a set of *events* corresponding to a patient's stay in the hospital. Each *event* $(t,e) \in V$ is comprised of a time t and an *event action* e.

Example 1. The first visit shown in Fig. 4a is described as $V_{ex} = \{(1,A),(1,B),(2,A),(3,C),(3,D)\}$.

An event (t,e) states that event action e occurred at time t. An event action represents the action associated with the event. For example, *injection of antibiotics* (procedure) and *transfer to ICU* (transfer) are event actions. Time is represented in days, where $t = 1$ means the first day of that patient's visit. Thus $(1, injection\ of\ antibiotic)$ is an event indicating that a patient received an injection of antibiotics on the first day of his or her visit.

We view a visit as entirely consisting of events. Admission data can be converted to events if we assign them to time $t = 0$ and convert them to event actions such as $@Age = 10$, $@Diag = 135$ (diagnosis is CCS code 135), $@Severity = Major$, etc. We use the @ symbol for admission data.



(a) From temporal data to its static equivalent. (b) The ensemble approach to class imbalance. (c) First pass of feature selection.

Figure 4: Overview of our approach: (a) feature engineering, (b) the ensemble approach to class imbalance, and (c) first pass of feature selection.

Definition 2. We define the *partial visit* of visit V at time t as $V(t) = \{(t', e) \in V : t' \leq t\}$, i.e., of events until day t.

Example 2. Partial visits from V_{ex} include $V_{ex}(0) = \emptyset$, $V_{ex}(1) = \{(1,A),(1,B)\}$ and $V_{ex}(2) = \{(1,A),(1,B),(2,A)\}$. Note that, $V_{ex}(t) = V_{ex}$ for any $t \geq 3$.

We need the notion of a partial visit because we want to be able to predict the risk of acquiring CDI at any point during a patient's visit.

Definition 3. We define the *bare events description* of [partial] visit V as $BE(V) = \{e : \exists t, (e,t) \in V\}$, i.e., as the set of the event actions in the events of V.

Example 3. For the whole visit V_{ex}, $BE(V_{ex}) = \{A,B,C,D\}$. For partial visit $V_{ex}(2)$, $BE(V_{ex}(2)) = \{A,B\}$.

The bare events description, or just *bare events*, represents the basic representation of visits in previous research [3, 4, 7]. Feature *received laxatives* in Dubberke et al is an example of this [3]. Our proposal, however, consists in combining event actions in temporal order for representing visits.

Definition 4. We define the *ordered pairs of events description* of [partial] visit V as $PE(V) = \{(e_1,e_2) : \exists t_1,t_2,\ t_1 \leq t_2 \wedge (e_1,t_1) \in V \wedge (e_2,t_2) \in V \wedge e_1 \neq e_2\}$.

Example 4. For the whole visit V_{ex}, $PE(V_{ex}) = \{(A,B),(A,C),(A,D),(B,A),(B,C),(B,D),(C,D),(D,C)\}$. For partial visit $V_{ex}(2)$, $PE(V_{ex}(2)) = \{(A,B),(B,A)\}$.

The ordered pairs of events description, or just *pairs of events*, couples event actions of events that succeed each other temporally or occur during the same day. Note that a pair of events is a simpler version of a partial order [11, 12]. Also note that this interpretation changes with admission data; in pair (e_1,e_2), if e_1 and e_2 represent admission data, then the pair represents an AND operation over admission data. For example, pair $(@Age = 60, @Diag = 135)$ means that the patient is $60-69$ years of age AND was admitted with intestinal infection ($@Diag = 135$). If only e_1 represents admission data, then the pair represents event action e_2 occurring in a visit with admission data including e_1. For example, pair $(@Age = 60, To = OR)$ means that a patient of age $60-69$ was transferred to the operating room (OR).

We now proceed to describe the role of hierarchies in our methodology.

Definition 5. The relation $\prec$ stands for the hierarchy relation ($e \prec e'$ means e is more specific than e') and $\prec^\star$ is the transitive closure of $\prec$.

Example 5. As Fig. 1 shows, we have that *Cephalosporins* $\prec$ *antibiotics* but *Cephalosporins* $\prec^\star$ *anti-infectives* as well as *Cephalosporins* $\prec^\star$ *antibiotics*.

To allow our classifier to make use of these hierarchies, we let the classifier decide on the level of granularity it needs to describe the data. For example, if all antibiotics increased the risk of CDI equally, the classifier could then assign the risk to the antibiotics category rather than to each individual antibiotic. The idea is to let the classifier decide on a small number of features (via feature selection) and make use of the aggregating power embedded in the hierarchies. Hence, we introduce *redundant* event actions in the visits, as we describe below.

Remark 1. Let event action e belong in category e', i.e., $e \prec^\star e'$. Then, e' is an event action and, for every pair $(t,e) \in V$, also $(t,e') \in V$, for all visits V.

Definition 6. We define the *hierarchically aware pairs of events description* of visit V as $PE_H(V) = PE(V) - \{(e,e') : e \prec^\star e' \vee e' \prec^\star e\}$.

Example 6. Let us suppose that $C \prec D$. Then, we have that $PE_H(V_{ex}) = \{(A,B),(A,C),(A,D),(B,A),(B,C),(B,D)\}$.

The definition of PE_H removes redundant information from PE. If event actions e and e' are related through a hierarchy, then we are not interested in knowing that e' generalizes e; this is not visit specific information, and therefore it does not help in classification.

For each visit (or partial visit), the application of functions like BE and PE_H defines a sparse description of the instance. For a data set, such descriptions induce the more standard tabular or matricial description of the data, necessary for classification with linear models.

Definition 7. Let $\mathcal{D}$ be a set of descriptions (either bare events or ordered pairs of events). We say that vector $\mathcal{H} = (h_1,...,h_m)$ is a *header* of $\mathcal{D}$ if:

1) The dimension of $\mathcal{H}$ is $m = |\cup_{E\in D} E|$.
2) Each component $h_i \in \cup_{E\in D} E$, for $i \in \{1,...,m\}$.
3) The components are not repeated, i.e., $h_i \neq h_j$ for all $i \neq j$, $i, j \in \{1,...,m\}$.

The *header* of a set of [partial] visit descriptions introduces a strict total order on the features that are used to describe the visits. In practice, we arrange the features alphabetically. Having defined the header, we now introduce the tabular description of a set of [partial] visits.

Definition 8. For [partial] visit description D and header $\mathcal{H}$, the *vectorial description* $v^{\mathcal{H}}(D)$ is a binary $|\mathcal{H}|$-dimensional vector such that, for all $i \in \{1,...,|\mathcal{H}|\}$,

$$V_i^{\mathcal{H}}(D) = \begin{cases} 1, & \text{if } h_i \in D, \\ 0, & \text{otherwise.} \end{cases}$$

Definition 9. Let F be a function that takes a [partial] visit and returns a bare events or ordered pairs of events description, i.e., $F \in \{BE, PE_H, BE \cup PE_H\}$, and let $\mathcal{T}$ be a set of tuples such that for every $(V,c) \in \mathcal{T}$, V is a [partial] visit and c is a class label. Then, the *tabular description of $\mathcal{T}$ induced by F* is defined as the pair $(\mathcal{H}, E)$, such that:

1) $\mathcal{H}$ is a header for the set $\{F(V) : (V,c) \in D\}$.
2) E is a set of tuples $\{(v^{\mathcal{H}}(V), c) : (V,c) \in D\}$.

Using $F \in \{BE, PE_H, BE \cup PE_H\}$ induce different tabular descriptions of a data set of visits. Classification is done on these tabular descriptions; for a tabular description $(\mathcal{H}, E)$, we construct the data matrix for the linear problem by using vectors $e \in E$ as its rows. In practice, however, we take advantage of the sparse description of visits by using sparse matrices in our code.

C. Additional features

In addition to the admission information and clinical events that naturally describe a visit, we hand-crafted a small number of additional features that are known be risk factors for CDI. We introduced features describing whether the patient was readmitted once or twice in the last 60 and 90 days, whether the patient had CDI within one year of admission, the diagnoses from the previous admission (if any), and the CDI testing method in place. We also computed the CDI *Colonization Pressure* [3, 5–7], which measures how many patients who had CDI stayed in the same unit as the patient. We use the daily version of the colonization pressure, describing it as event actions *Pressure=LOW*, *Pressure=MODERATE*, and *Pressure=HIGH,* and their generalization, *Pressure*. A pressure of zero means no event is introduced.

D. Dimensionality growth

The introduction of pairs of events drastically increases the feature space from around 3,000 bare events to around 300,000 pairs of events. Such high dimensionality threatens the purpose of yielding a human-interpretable prediction model, which would benefit from few but relevant features. Moreover, such high dimensionality threatens classification, especially considering that the minority class consists of just 950 visits (out of 200,000), and also because of the increased computational complexity of the classification algorithm, because of the enlargement of the data set. For these reasons, we split the training set into smaller chunks, feed them to an ensemble classifier, and perform extensive feature selection. The methodology is presented in the next section.

V. Classification

A. Addressing class imbalance through ensembles

Our approach to classification consists of building an ensemble of logistic regression classifiers to estimate risk. Using ensembles allows us to reduce training on a large data set to training on several smaller data sets, as well as address class imbalance by constructing subsets of the data that are balanced, as done by Lim et al [17]. As shown in Fig. 4b, we train the classifiers on subsets of the data that contain the whole minority class and a random subset of the majority class, so that both classes are balanced. This results in a collection of classifiers that are agnostic to class imbalance, a property inherited by the ensemble. We do not need to oversample the minority class, introduce perturbations, etc., as is often done in other research [34].

Formally, if n is the number of visits in the minority (CDI) class ($n = 950$), we pick n random visits from the majority class. Then, for each visit, we pick R days sampled at random ($R = 3$) and represent them using the methodology described in Section IV-B, producing training sets of nR instances for each class. Each training set will be used by one logistic regression classifier. For the CDI class, we do not sample days later than 3 days before CDI was detected. (The guidelines recommend testing patients that have had diarrhea or other symptoms of CDI for at least 3 days.)

B. Feature selection

Our objective is to reduce the number of features to a reasonably low number, to facilitate interpretability of the model. Some researchers address high dimensionality through ensembles, by randomly selecting features in the classifiers [35–40]. Instead, we perform feature selection inside each classifier. This produces an ensemble classifier that considers substantially fewer features, aimed to facilitate interpretability. Dimensionality reduction approaches, such as Johnson-Lindenstrauss and PCA, do not necessarily result in using few features (one dimension might span too many features) and non experts in dimensionality reduction often interpret the produced dimensions incorrectly [41].

We filter features in two stages. In the first stage, we pass over the feature space, quickly discarding the least relevant features. We do so by using Algorithm 1, which processes a tabular description of the data (Def. 9) and returns the m "most likely" predictive features, with $m = 1000$[1].

In the second stage, we use L_1-regularization to further reduce the number of features. The L_1-regularized cost function for fitting logistic regression is

$$L(\alpha, \beta; \lambda) = \lambda|\beta|_1 + \sum_{(x,y)\in S} \ln\left(1 + \exp(-y(\alpha + \beta^T x))\right), \quad (1)$$

[1]Passing too much data to a classifier can worsen its runtime and memory consumption. A value of m that is too large can make classification infeasible. Smaller values of m can increase the overall speed of classification at the expense of model quality (compensated with a larger ensemble). A value m that is too small can lead to ignoring potentially relevant features.

Algorithm 1 Greedy randomized embedded feature filter

Input: m: bucket size, $\mathcal{H}$: features, $\mathcal{S}$: rows of the data set

1: Randomly partition $\mathcal{H}$ into sets $B_1, ..., B_k$, so that $|B_i| = m$ for every i such that $1 \leq i \leq k-1$.
2: Let $C = B_1$.
3: **for** i = 2 to k **do**
4: Fit logistic regression model to $\mathcal{S}$ projected on $C \cup B_i$
5: For $h \in C \cup B_i$, define $s(h)$ as the number of classification errors introduced when β_h is set to 0 (β_h is the coefficient of the logistic regression model for feature h).
6: Update C to be the m features in $C \cup B_i$ with the highest $s(h)$.
7: **end for**
8: **return** C

where α and β describe the logistic regression model, λ is the penalty on $|\beta|_1$, S is the set of instances of the tabular description of the data (Def. 9), and in $(x,y) \in S$, x is the instance vector and $y \in \{-1,+1\}$ is the class. (When $\lambda = 0$, we have traditional logistic regression.) Ideally, λ is chosen through cross validation. Since this can be expensive, we follow Fan and Tal [18], choosing λ by minimizing the Bayesian Information Criterion (BIC) of L, which applied to our problem is

$$BIC = -2L + (1 + |\beta|_0) \ln |S|. \quad (2)$$

Using L_1-regularization for feature selection has theoretical backing. The $L_0 - L_1$ *equivalence* [42] result states that, in sparse data, L_1-regularization can effectively approximate the ideal L_0-regularization, a notorious NP-hard problem, in polynomial time. However, using BIC instead of cross validation should reduce the quality of the approximation slightly.

We use L_1-regularization to approximate L_0-regularization, with the intention of making each classifier in the ensemble as small as possible. Since we create an ensemble, the number of selected features is bound to increase in any case, and, therefore, we could consider using L_2-regularization or LASSO over L_1-regularization (see the work of Vidaurre et al [43] for L_1 and other regularization schemes). However, using less strict regularization may increase the number of features without improving prediction quality (Table VI supports this point).

C. Estimation

We predict CDI by converting each visit into a feature description (Def. 9), and supplying this description to each logistic regression model in the ensemble. Then, the ensemble counts the number of times CDI is predicted; if above 50%, the ensemble predicts CDI. This approach gives the same results as averaging the probabilities produced by the models.

VI. Experiments

A. Experiments outline

Our first experiment compares our methodology to the state of the art: the work of Wiens et al [7]. The experimental setting lies between their setting and ours, in that we predict using 1-2 days worth of data only, but we perform 10-fold cross validation rather than using one year to predict the next. The second experiment further compares bare events and pairs of events, and shows that the classifiers generate evolving risk curves. The third experiment compares using only information known at admission time versus using only clinical events. Using only information known at admission time produces fairly predictive results, while using only clinical events produces less predictive results.

	PEC	BEC	SAC
Pairs of events (PE_H)	√		
Bare events (BE)	√	√	√
Ensemble of logistic regression	√	√	
Feature selection	√	√	
Compensation for class imbalance	√	√	

Table III: Characteristics of the three classifiers.

The general settings for our classifiers consist of ensembles of 30 logistic models each, sampling each visit into three partial visits (randomly), and filtering 1000 features in stage 1 of feature selection. For each classifier, we report its area under the curve (AUC), which we use as the main parameter for comparison. We also report the sensitivity (true positive rate) and specificity (true negative rate) for completeness. Additionally, we report the number of active features (the ones included in the classifier, i.e., with $\beta_i \neq 0$) and the inactive features (with $\beta_i = 0$). A feature is inactive if discarded through feature selection or deemed irrelevant during regression.

B. Improvement over baseline

In the first experiment, we compare three classifiers: the pairs of events classifier (PEC), the bare events classifier (BEC), and the state of the art classifier (SAC). PEC consists of the methodology presented in this paper, with features produced by both BE and PE_H. BEC is identical to PEC, except that features are only described through BE. SAC is an adaptation of the work of Wiens et al [7]. Table III summarizes the characteristics of PEC, BEC, and SAC. We compare the classifiers using 10-fold cross validation and data limited to 1 or 2 days after admission, for fair comparison against Wiens et al. Note that SAC does not fully follow their research. They used L_2-regularized logistic regression to predict cases of CDI using data known at admission time (e.g., demographics, initial diagnoses) as well as clinical events (e.g., procedures, prescriptions) and laboratory values (e.g., blood pressure, temperature) until 24 hours after admission. We cannot use such data, because we do not have laboratory values and our discrete notion of time does not permit us to cut visits exactly 24 hours after admission. We compensate for the later by considering visits up to day 1 or 2. Wiens et al also used data from one year to predict the next, which overcomes the problem of changes in the testing of CDI. Since we introduce a feature indicating the CDI testing method being used at the time of admission, we do not consider it necessary to train on one year to predict the next, thus training SAC identically to PEC and BEC.

Table IV summarizes general statistics of the different classifiers in the task of predicting whether patients will develop CDI using data known at 1 or 2 days after admission. The best performing classifier was PEC, followed closely by BEC. A more detailed visual description of the performance of the classifiers in this task is shown in the ROC curves of Fig. 6a. Note that the AUC of SAC is similar to the one shown in Wiens et al [7]. The low sensitivity and high specificity

Classifier	AUC	Sensitivity	Specificity	Active fs	Inactive fs
SAC	80.57%	17.19%	99.32%	1999.0	713.2
BEC	83.94%	76.32%	76.06%	461.7	2250.5
PEC	**85.19%**	**78.04%**	**75.86%**	**3263.4**	**150740.8**

Table IV: Performance predicting CDI cases using data known at 1 or 2 days after admission. For each classifier, we report its AUC, sensitivity, specificity, and number of active and inactive features. The values are averaged over the 10-fold cross validation tests.

Classifier	AUC	Sensitivity	Specificity	Active fs	Inactive fs
Any day					
BEC	85.26%	78.04%	76.44%	539.8	2201.2
PEC	**86.61%**	**82.21%**	**74.82%**	**3729.7**	**262222.0**
Later days					
BEC	85.21%	77.12%	76.76%	576.4	2135.8
PEC	86.53%	82.25%	74.26%	4132.1	269594.2

Table V: Performance predicting CDI at any day of a patient's visit. For each classifier, we report its AUC, sensitivity, specificity, and number of active and inactive features. The values are averaged over the 10-fold cross validation tests.

of SAC come from the fact that class imbalance was not addressed. With respect to the number of active features, BEC considered much fewer features than SAC, which is consistent with the use of feature selection. Furthermore, BEC performed better than SAC. On the other hand, PEC considered more features than BEC, but taken from a much larger pool of more than 150,000 features.

C. Up-to-date risk estimation

In this experiment, we compare PEC and BEC for the task of predicting whether the patient will develop CDI during a visit. We consider two alternative training sets: *any day* and *later days*. The *any day* training set consists of patient visits cut off at days sampled uniformly at random. The *later days* training set cuts patients' visits with linearly increasing probability, making later days more likely to be sampled than earlier days. The idea behind *later days* is that interesting [pairs of] events might occur later during the visit. For both training sets, the CDI class can be sampled until 3 days before diagnosis. The non-CDI class can be sampled until the very end of the visit under *any day*, but only up to until 2 weeks under *later days*, to reduce the effect of extremely long visits (months, years) on classification. (As Table I shows, most visits last only a few days.)

Table V shows the performance of PEC and BEC when trained under the *any day* and *later days* data sets. The BEC classifiers lag slightly behind the PEC classifiers under both training sets, but their difference in AUC is small. Fig. 6b shows that the PEC classifiers perform nearly identically, while the BEC classifiers lag closely behind. The effect of the training set appears irrelevant. Note that the AUCs of PEC and BEC in this experiment are similar to the previous one (Table IV). This hints that data on admission and early events may be good predictors of the outcome of the patient.

A side-result of the classifiers trained is that they are more likely to predict CDI when the onset of symptoms approaches.

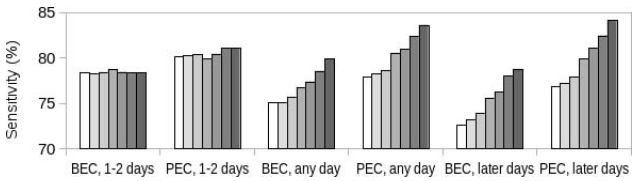


Figure 5: BEC and PEC classifiers as the time to CDI approaches. The bars represent the sensitivity of the classifiers from 7 days to the day before the onset of symptoms.

	AUC		Active features	
Classifier	With	Without	With	Without
1-2 days				
BEC	85.07%	84.10%	461.7	1567.2
PEC	86.20%	83.97%	3263.4	5143.6
Any day				
BEC	85.26%	84.25%	539.8	1629.6
PEC	86.61%	84.49%	3729.7	5191.8
Later days				
BEC	85.21%	84.10%	576.4	1630.8
PEC	86.53%	84.34%	4132.1	5264.8

Table VI: Impact of L_1-regularization with BIC minimization on the classifiers. For each classifier, the AUC on the *any day*, *later days* and *1-2 days* testing sets are presented, as well as the number of features, for the cases *with* and *without* regularization. The values are averaged over the 10 fold cross validation tests.

Fig. 5 shows the sensitivity of the PEC and BEC classifiers as the onset of CDI approaches. The classifiers were trained on *any day* and *later days*, as well as in *1-2 days*, which represents the setting from the previous experiment. As expected, training on admission and early events only (BEC and PEC on *1-2 days*) does not produce risk curves. Training on *any day* and *later days* produces risk curves, without much difference between them; PEC outperforms BEC for this task.

Table VI shows that BIC minimization contributed slightly to improve out-of-sample performance while noticeably reducing the number of active features in the ensembles. The use of BIC minimization signified a reduction in at least 19.6% of the features. The average in-sample accuracy of each regression model in the BEC classifiers is 83.2%. For PEC regression models, accuracy is 93% on average. This suggests that the ensembles help compensate for underfit and overfit regression models.

D. Admission data versus clinical events

As using only data available prior to day 2 seem to suffice for predicting whether the patient will develop CDI during the visit, we considered the question of prediction accuracy using either admission data or clinical events data. In this experiment, we compare PEC and BEC when trained on either admission data only or clinical events only. Table VII shows the performance of BEC and PEC classifiers trained using either admission data or clinical events while Fig. 6c compares their ROC. Classifiers using admission data clearly outperform classifiers using clinical events, which confirms that information available at admission time can indeed be

Classifier	AUC	Sensitivity	Specificity	Active fs	Inactive fs
Admission data					
BEC	82.02%	72.63%	77.07%	73.6	1696.3
PEC	83.13%	68.00%	80.54%	280.0	58331.4
Clinical events data					
BEC	77.58%	64.34%	75.43%	272.2	485.0
PEC	78.83%	69.62%	72.11%	3180.0	117536.3

Table VII: Performance predicting CDI using either admission data or clinical events data. For each classifier, we report its AUC, sensitivity, specificity, and number of active and inactive features. The values are averaged over the 10-fold cross validation tests.

used to predict whether a patient will develop CDI during the visit.

E. Features selected in PEC

In the *PEC, any day* classifier, the *most influential* features are dominated by bare events and admission data. To us, the influence of a feature is its absolute log-odds ratio, i.e., $|\beta_i|$ for feature i. Table VIII shows the 20 most influential features in the classifier. Features of the form `[x < y]` represent pairs of events. Most of the features in Table VIII involve bare events and/or admission data, which explains the previous results, showing that using admission data alone could lead to good prediction. If we extend the analysis to the 100 most influential features, we see a similar picture. 36 features correspond to bare events, with 33 being about admission data, while 64 features correspond to pairs of events, with 62 involving admission data. Moreover, 59 pairs of events are strictly about admission data, i.e., they do not involve clinical events. The 5 features that do not involve admission data are: `[To=PORR < To=OR]`, which states that visiting the PORR (post-OR) before the OR (operating room) reduces the risk of CDI; `Proc=231`, which states that undergoing *another therapeutic or diagnostic* procedure increases risk; `[To=OR < Proc=Diag/Therap]`, which states that visiting the OR to undergo *any* diagnostic or therapeutic procedure increases risk; `Proc=223`, which states that enteral or parenteral nutrition increases risk; and `To=4JPW`, which states that visiting a specific unit (4JPW) decreases risk. To be noted, these features cannot just be interpreted in isolation; they co-occur with many events, because medical events are associated with the patient's condition. This explains the two most influential pairs that partially involve admission data: `[@AdmType=NEWBORN < Proc=Cardiovasc]` and `[@SvcCat=PEDIATRICS < Proc=Cardiovasc]`. Both state that either a newborn or a child that undergoes a cardiovascular procedure is more likely to develop CDI.

Extending to the top 1000, which make up for the dominating features of the classifier, we see that only 154 do not involve admission data, 134 are pairs of events. Of these pairs, prescriptions occur in 108 (54 are exclusively pairs of prescriptions), procedures occur in 72, and transfers are the least frequent, occurring in 12 pairs.

Diagnoses participate in 50 features of the top 100 and in 578 of the top 1000. The most influential diagnoses, several present in Table VIII, include *intestinal infection* (@Diag=135), *osteoarthritis* (203), *pancreatic disorders except diabetes* (152), *nutritional deficiencies* (52), *abdominal hernia* (143), and *other lower respiratory disease* (133).

Feature name	Log odds (β_i)
`@Diag=135`	5.2718
`@Severity=Major`	2.7682
`@Severity=Extreme`	2.4677
`@Severity=Moderate`	1.8935
`@AdmSrc=NEWBORN PREMATURE BIRTH`	1.7402
`[@Severity=Minor < @AdmType=ELECTIVE/ROUTINE]`	1.4477
`@SvcCat=INTERNAL MEDICINE`	1.1968
`@Diag=203`	-1.1257
`@AGE=20`	-1.1054
`[To=PORR < To=OR]`	-1.0547
`@PCR_period`	-0.9199
`[@PCR_period < @Diag=152]`	0.8856
`[@SvcCat=INTERNAL MEDICINE < @AdmType=ELECTIVE/ROUTINE]`	0.8567
`@SvcCat=FAMILY MEDICINE`	-0.8476
`[@SvcCat=PSYCHIATRY < @AdmType=EMERGENCY]`	-0.8273
`@AGE=30`	-0.7447
`[@AdmType=URGENT < @AGE=20]`	-0.7261
`@AdmSrc=UIHC CLINIC`	-0.7136
`@Diag=52`	0.6986
`@Readm_90D`	0.6706

Table VIII: Top 20 most influential features in PEC, any day.

F. Role of time

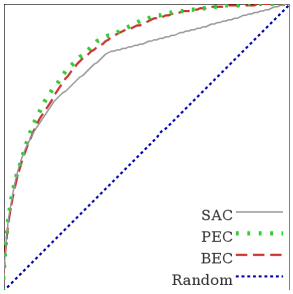
For the most part, temporal ordering played a small, but significant role in classification as evidenced by the fact that PEC performed consistently better than BEC.

Table IX shows the top pairs of events where order matters the most, i.e., those with the highest difference $\beta_{[x<y]} - \beta_{[y<x]}$. Observe that the log-odds in some of them even change signs. Some of these orders are consistent with the literature of risk factors of CDI; for example, that receiving antibiotics after some other event signaling exposure (e.g., repiratory intubation) increases the risk of CDI. In other cases, pairs of events can be seen as markers of the progression of the infection, as in the case when parenteral nutrition was needed (`[Proc=223 < Proc=231]`) and when nutritional agents were given before medication for vertigo-nausea (`[RxMaj=40 < RxSmin=562210]`).

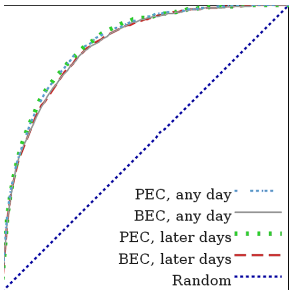
Antibiotics were present in several of the pairs shown in Table IX. Overall, pairs of events involving systemic antibiotics represent 6.16% of all the features, almost always participating in pairs of events rather than bare events. Most of the time, the whole category of antibiotics is mentioned. Otherwise, first and fourth generation Cephalosporins, Penicillins and Aminopenicillins are the antibiotics mentioned. Antibiotics seem to increase risk when the patient has received metabolic agents, and when receiving anticoagulants and anticonvulsants. The last two seem to suggest that such patients underwent severe dehydration and nausea, which are common symptoms of CDI. Systemic antifungals seem to also increase the risk of CDI.

G. C.difficile exposure

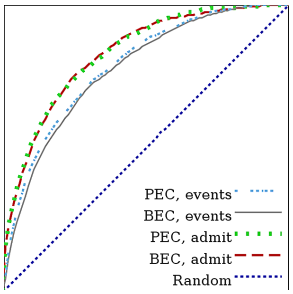
In much of the literature, it has been argued that c.diff is highly contagious. Hence, one might expect features in our classification to demonstrate this. For example, one might expect to see pairs of events of the type "high colonization pressure, then exposed to antibiotics" would increase the risk



(a) Prediction using 1 or 2 days of a visit.



(b) Prediction at any day of a visit.



(c) Prediction using admission data only (admit) or clinical events only (events).

Figure 6: ROC curves of the classifiers in the task of predicting CDI, using different aspects of the data. The ROC curves are averaged over the 10-fold cross validation tests. The *Random* curve stands for the uninformed classifier.

Pair of events	Readable version	$\beta_{[x<y]}$	$\beta_{[y<x]}$
`[To=OR < To=PORR]`	*transferred to OR, then transferred to PORR*	-0.1831	-1.0547
`[To=OR < Proc=Diag/Therap]`	*transferred to OR, then underwent miscellaneous diagnostic-therapeutic procedure*	0.3556	-0.0004
`[Proc=223 < Proc=231]`	*underwent 'enteral and parenteral nutrition', then underwent 'other therapeutic procedures'*	0.1982	-0.0019
`[Proc=231 < RxMin=812]`	*underwent 'other therapeutic procedures', then prescribed 'antibiotics systemic'*	-0.0140	-0.2062
`[Proc=Diag/Therap < RxSmin=81206]`	*underwent miscellaneous diagnostic-therapeutic procedure, then prescribed 'fourth generation cephalosporins'*	0.1781	0.0144
`[RxMaj=40 < RxSmin=562210]`	*prescribed 'nutrients/nutritional agents', then prescribed '5ht3 receptor antagonists'*	0.1362	0.0079
`[Proc=216 < RxSmin=81219]`	*underwent 'respiratory intubation and mechanical ventilation', then prescribed 'extended-spectrum penicillins'*	0.0042	-0.1182
`[To=6RCE < Proc=Diag/Therap]`	*transferred to 6RCE, then underwent miscellaneous diagnostic-therapeutic procedure*	-0.0055	-0.1120
`[RxSmin=280892 < RxSmin=81203]`	*prescribed 'misc analgesics systemic', then prescribed 'first generation cephalosporins'*	0.1132	0.0096
`[Proc=177 < RxSmin=280808]`	*underwent 'computerized axial tomography (ct) scan head', then prescribed 'opiate agonists'*	-0.0161	-0.1191

Table IX: Top 10 pairs of events where the order is relevant. For each pair of events, we include a human readable description of it as well as the log-odds of the original order $\beta_{[x<y]}$ and the converse order $\beta_{[y<x]}$.

of developing CDI. But this was not the case. In fact, the Pressure events were, for the most part, ignored by our classifiers. Furthermore, the good performance of the classifiers on only 1-2 days worth of visit data seems to downplay the role of exposure in the development of CDI. This may suggest that c.diff exposure plays a lesser role in CDI, as suggested in recent research [25, 26].

However, we need to emphasize the limitations of using our classifier's output to estimate the "importance" of features. Our classifier aims to produce a simple, minimally redundant explanation of risk, e.g., if two features have a similar explanatory power, the classifier will choose only one. Since clinical management of the patient is highly dependent on the patient's condition, we can easily explain many procedures, medications, and locations associated with a patient-visit just by knowing the patient's admission information. Moreover, many procedures are associated with particular locations in the hospital, because of the medical speciality associated. Thus, it is likely that pressure-related features (that have an important spatial component) were subsumed by other features that collectively provided a minimally redundant explanation.

VII. Conclusion

We addressed the problem of predicting CDI using temporal information from medical records. We described the temporal information (events) that occurred during a patient's visit as *ordered pairs of events* (pairs of events). A pair of events (x,y) or `[x < y]` states that event `x` took place the day before or the same day as event `y`. We crafted an ensemble of logistic regression models, where each classifier of the ensemble was trained on a balanced subset of the data and performed extensive feature selection, addressing the problems of class imbalance and high dimensionality, respectively. Our methodology slightly outperforms baseline classifiers in the task of predicting CDI. However, our most salient contribution is that of a classifier that produces interpretable information which could later inform medical decision making.

Our work is subject to several limitations. First, by describing visits as ordered pairs of events, we are missing the opportunity of learning what happens when orders are longer, e.g., with ordered triples of events. Second, we have not introduced a methodology for recommending the parameters of the classifier (ensemble size, visit resamples, bucket size in feature selection, etc.). Third, even though the models produced by our methodology are readable, the narrative they produce is simple but incomplete. Groups of co-occurring clinical events are likely to be ignored, except for one or two. This implies that meaningful associations can be hidden. For example, causal relation $a \rightarrow b$ could be described by pair (c,d) if c co-occurs with a and d co-occurs with b, even though c and d are causally unrelated. Furthermore, even though the ensembles reduce the number of features to use, the least predictive features are less likely to be consistently chosen among logistic regression models, hence increasing the total number of features chosen in the ensemble, which is

detrimental to classification. These limitations, especially the last one, are the subject of future work.

Despite our limitations, we describe how novel approaches to using existing medical records can help anticipate an important healthcare-associated infection. However, our approach may also have applications for other hospital-associated infections and adverse events, which together contribute to significant hospital-associated morbidity and mortality.

Acknowledgements: This work was funded in part by the University of Iowa's eHealth and eNovation Center.

REFERENCES

[1] F. Lessa, Y. Mu, W. Bamberg *et al.*, "Burden of clostridium difficile infection in the united states," *New England Journal of Medicine*, vol. 372, no. 9, pp. 825–834, 2015.

[2] Centers for Disease Control and Prevention (CDC), "Vital signs: preventing clostridium difficile infections," *MMWR Morbidity and Mortality Weekly Report*, vol. 61, no. 9, pp. 157–62, 2012.

[3] E. Dubberke, Y. Yan, K. Reske *et al.*, "Development and validation of a clostridium difficile infection risk prediction model," *Infection Control and Hospital Epidemiology*, vol. 32, no. 4, pp. 360–6, 2011.

[4] K. Garey, T. Dao-Tran, Z. Jiang *et al.*, "A clinical risk index for clostridium difficile infection in hospitalised patients receiving broad-spectrum antibiotics," *Journal of Hospital Infection*, vol. 70, no. 2, pp. 142–7, 2008.

[5] J. Wiens, J. Guttag, and E. Horvitz, "Learning evolving patient risk processes for c. diff colonization," *Machine Learning for Clinical Data Analysis. ICML 2012*.

[6] ——, "Patient risk stratification for hospital-associated c. diff as a time-series classification task," *NIPS 2012*.

[7] J. Wiens, W. Campbell, E. Franklin *et al.*, "Learning data-driven patient risk stratification models for clostridium difficile," *Open Forum Infectious Diseases Advance Access, June 2014*.

[8] S. Fujitani, W. George, and A. Murthy, "Comparison of clinical severity score indices for clostridium difficile infection," *Infection Control and Hospital Epidemiology*, vol. 32, no. 3, pp. 220–8, 2011.

[9] P. Jensen, L. Jensen, and S. Brunak, "Mining electronic health records: towards better research applications and clinical care," *Nature Genetics*, vol. 13, pp. 395–405, 2012.

[10] C. Paxton, A. Niculescu-Mizil, and S. Saria, "Developing predictive models using electronic medical records: Challenges and pitfalls," *AMIA 2013*.

[11] D. Patnaik, P. Butler, N. Ramakrishan *et al.*, "Experiences with mining temporal event sequences from electronic medical records: Initial successes and some challenges," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '11, 2011, pp. 360–368.

[12] D. Patnaik, N. Ramakrishnan, L. Parida *et al.*, "Mining significant partial order patterns in electronic medical records (poster)," *AMIA 2011*.

[13] N. Sundaravaradan, N. Ramakrishnan, and D. Hanauer, "Factorizing event sequences," *IEEE Computer*, vol. 45, no. 12, pp. 73–75, 2012.

[14] N. Lee, A. Laine, H. Hu *et al.*, "Mining electronic medical records to explore the linkage between healthcare resource utilization and disease severity in diabetic patients," in *2011 First IEEE International Conference on Healthcare Informatics, Imaging and Systems Biology (HISB)*, 2011.

[15] J. Li, A. Fu, H. He *et al.*, "Mining risk patterns in medical data," *KDD 2005*.

[16] R. Henriques, S. Pina, and C. Antunes, "Temporal mining of integrated healthcare data: Methods, revealings and implications," *2013*.

[17] N. Lim, H. Ahn, H. Moon, and J. Chen, "Classification of high-dimensional data with ensemble of logistic regression models," *Journal of Biopharmaceutical Statistics*, vol. 20, no. 1, pp. 160–71, 2010.

[18] Y. Fan and C. Tang, "Tuning parameter selection in high dimensional penalized likelihood," *Journal of the Royal Statistical Society, series B*, vol. 76, no. 3, pp. 531–552, 2012.

[19] AHRQ Effective Health Care Program, *Treating and Preventing C-diff Infections: A Review of the Research for Adults and Their Caregivers*.

[20] T. Henrich, D. Krakower, A. Bitton, and D. Yokoe, "Clinical risk factors for severe clostridium difficile-associated disease," *Emerging Infectious Diseases*, vol. 15, no. 3, pp. 415–22, 2009.

[21] J. Fashner, M. Garcia, L. Ribble, and K. Crowell, "Clinical inquiry: what risk factors contribute to c difficile diarrhea?" *Journal of Family Practice*, vol. 60, no. 9, pp. 545–7, 2011.

[22] R. Jump, M. Pultz, and C. Donskey, "Vegetative clostridium difficile survives in room air on moist surfaces and in gastric contents with reduced acidity: a potential mechanism to explain the association between proton pump inhibitors and c difficile-associated diarrhea?" *Antimicrobial Agents and Chemotherapy*, vol. 51, no. 8, pp. 2883–7, 2007.

[23] M. Hamel, D. Zoutman, and C. O'Callaghan, "Exposure to hospital roommates as a risk factor for health care-associated infection," *American Journal of Infection Control*, vol. 38, no. 3, pp. 173–81, 2010.

[24] M. Shaughnessy, R. Micielli, D. DePestel *et al.*, "Evaluation of hospital room assignment and acquisition of clostridium difficile infection," *Infection Control and Hospital Epidemiology*, vol. 32, no. 03, pp. 201–206, 2011.

[25] A. Walker, D. Eyre, D. Wyllie *et al.*, "Characterisation of clostridium difficile hospital ward-based transmission using extensive epidemiological data and molecular typing," *PLoS Medicine*, vol. 9, no. 2, p. e1001172, 2012.

[26] D. Eyre, M. Cule, D. Wilson *et al.*, "Diverse sources of c.difficile infection identified on whole-genome sequencing," *New England Journal of Medicine*, vol. 369, no. 13, pp. 1195–205, 2013.

[27] A. Elixhauser, C. Steiner, and L. Palmer, "Clinical classifications software (ccs), 2014," *U.S. Agency for Healthcare Research and Quality*.

[28] M. Siemann, M. Koch-Dörfler, and G. Rabenhorst, "Clostridium difficile-associated diseases: The clinical courses of 18 fatal cases," *Intensive Care Medicine*, vol. 26, no. 4, pp. 416–21, 2000.

[29] NHS Choices (UK), "Complications of clostridium difficile infection."

[30] I. Batal, H. Valizadegan, G. Cooper, and M. Hauskrecht, "A pattern mining approach for classifying multivariate temporal data," *IEEE BIBM 2011*.

[31] ——, "A temporal pattern mining approach for classifying electronic health record data," *ACM TIST 2012*.

[32] R. Moskovitch, C. Walsh, G. Hripcsak, and N. Tatonetti, "Prediction of biomedical events via time intervals mining," *BigCHat Workshop, KDD 2014*.

[33] R. Moskovitch and Y. Shahar, "Medical temporal-knowledge discovery via temporal abstraction," in *AMIA annual symposium proceedings*, vol. 2009, 2009, p. 452.

[34] H. He and E. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[35] A. Brahim and M. Limam, "Robust ensemble feature selection for high dimensional data sets," *HPCS 2013*.

[36] K. Hwang, I. Lee, J. Park *et al.*, "Reducing false-positive incidental findings with ensemble genotyping and logistic regression based variant filtering methods," *Human Mutation*, vol. 35, no. 8, pp. 936–944, 2014.

[37] J. Shankar, S. Szpakowski, N. Solis *et al.*, "A systematic evaluation of high-dimensional, ensemble-based regression for exploring large model spaces in microbiome analyses," *BMC Bioinformatics*, vol. 16, no. 31, 2015.

[38] S. Wang, X. Chen, J. Huang, and S. Feng, "Scalable subspace logistic regression models for high dimensional data," *APWeb 2012, LNCS 7235, pp. 685-694, 2012*.

[39] P. Yang, W. Liu, B. Zhou *et al.*, "Ensemble-based wrapper methods for feature selection and class imbalance learning," in *Advances in Knowledge Discovery and Data Mining*. Springer, 2013, pp. 544–555.

[40] R. Zakharov and P. Dupont, "Ensemble logistic regression for feature selection," in *Pattern Recognition in Bioinformatics*. Springer, 2011, pp. 133–144.

[41] J. M. Lewis, L. Van Der Maaten, and V. R. de Sa, "A behavioral investigation of dimensionality reduction," in *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, 2012, pp. 671–676.

[42] D. Lin, D. Foster, and L. Ungar, "A risk ratio comparison of l0 and l1 penalized regressions," *University of Pennsylvania, techical report, 2010*.

[43] D. Vidaurre, C. Bielza, and P. Larranaga, "A survey of l1 regression," *International Statistical Review*, vol. 81, no. 3, pp. 361–387, 2013.