A Data-driven Approach to Identifying Asymptomatic *C. diff* Cases

Hankyu Jang
hankyu-jang@uiowa.edu
Dept of Computer Science
The University of Iowa

Philip M. Polgreen
philip-polgreen@uiowa.edu
Dept of Internal Medicine
The University of Iowa

Alberto M. Segre
alberto-segre@uiowa.edu
Dept of Computer Science
The University of Iowa

Daniel K. Sewell
daniel-sewell@uiowa.edu
Dept of Biostatistics
The University of Iowa

Sriram V. Pemmaraju
sriram-pemmaraju@uiowa.edu
Dept of Computer Science
The University of Iowa

*For the CDC MInD-Healthcare Group

ABSTRACT

Asymptomatic carriers of an infection make it more challenging to understand the characteristics of that infection (e.g., parameters such as R_0) and to design, implement, and evaluate interventions. Asymptomatic carriers are usually not tested, which also means we do not have "ground truth" labels for these cases in our data. In this paper, we propose a 2-stage classification model for inferring asymptomatic carriers of *Clostridioides difficile* (*C. diff*) infections (CDI), a common healthcare-associated infection that causes almost half a million illnesses in the US each year. Guided by hypotheses derived from literature on risk factors for C. diff carriers, we design a Stage 1 model for detecting *asymptomatic* C. diff carriers that is trained on *symptomatic* CDI cases. We evaluate the performance of this Stage 1 model by designing a Stage 2 model to predict CDI incidence that uses among its inputs exposure to asymptomatic C. diff carriers inferred by our Stage 1 model. Results from this evaluation lead to two findings. First, our results show that the best performing Stage 1 model depends on all of the standard risk factors for CDI except for high-risk antibiotics. This is an intriguing finding that highlights an important difference between the risk profile of CDI patients and C. diff carriers. Second, we show that adding exposure to asymptomatic cases as an input to the Stage 2 CDI classification model leads to better performance. This result implies that asymptomatic C. diff carriers do in fact contribute to CDI spread, confirming an important conjecture from the CDI literature.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
epiDAMIK 2020, Aug 24, 2020, San Diego, CA
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-xxxx-XXXX-X…$15.00
https://doi.org/xx.xxxx/xxxxxxxxx.xxxxxxx

CCS CONCEPTS

• **Theory of computation → Semi-supervised learning**; • **Applied computing → Health informatics**.

KEYWORDS

Asymptomatic carrier detection, Clostridioides difficile, Colonization pressure, High-risk antibiotics, Spatio-temporal clustering

ACM Reference Format:
Hankyu Jang, Philip M. Polgreen, Alberto M. Segre, Daniel K. Sewell, and Sriram V. Pemmaraju. 2020. A Data-driven Approach to Identifying Asymptomatic *C. diff* Cases. In *epiDAMIK 2020: 3rd epiDAMIK ACM SIGKDD International Workshop on Epidemiology meets Data Mining and Knowledge Discovery*. ACM, New York, NY, USA, 8 pages. https://doi.org/xx.xxxx/xxxxxxxxx.xxxxxxx

1 INTRODUCTION

For many infections, *asymptomatic* cases present a major obstacle to understanding precisely how the infection is spread, and they make implementing effective interventions that much more challenging. Indeed, asymptomatic cases are widely believed to play a substantial role in the spread of COVID-19 [3, 21] and asymptomatic transmission of SARS-CoV-2 has been called the "Achilles' heel" of control strategies for COVID-19 [13].

The focus of this paper is on inferring asymptomatic cases of a common *healthcare-associated infection* (HAI) known as *C. diff infection*, or CDI. An HAI is an infection that a patient acquires in a healthcare facility while being treated for another condition. At any given time, 1 in 25 patients in the US has an HAI [23]. CDI is caused by the bacterium *Clostridioides difficile*, and is characterized by diarrhea and inflammation of the colon: there are almost half a million cases of CDI in the US each year [12]. CDI, and HAIs in general, pose a major challenge to healthcare systems worldwide, especially because some of these infections are becoming resistant to antibiotics, the primary treatment used to address these infections.

There is evidence that a substantial fraction of patients admitted to a healthcare facility are *asymptomatic* C. diff carriers [18, 19]. One particular study [19] found that up to 10% of patients admitted

to a tertiary hospital in Minnesota during March-April 2009 were in fact asymptomatic C. diff carriers. Yet the role of asymptomatic cases in the spread of CDI within healthcare facilities is largely unexplored [1], though there is accumulating indirect evidence that this role is substantial. For example, another study [10] found that 45% of CDI cases originated from sources other than symptomatic cases, suggesting a significant role for asymptomatic persons; a still more recent study [29], found that only 17% of CDI cases in a hospital ward had direct contact with other symptomatic patients, also suggesting that the pathogen had been acquired from other, presumably asymptomatic, sources.

Understanding the role of asymptomatic C. diff carriers is a critical element in designing effective interventions. Our paper presents a data-driven approach to identifying and understanding the role of asymptomatic C. diff carriers on the diffusion of CDI in a healthcare setting.

1.1 Results and Approach

Guided by literature on risk factors for being an asymptomatic C. diff carrier [16], we evaluate multiple data-driven models for inferring if a patient is an asymptomatic C. diff carrier. Our evaluation is based on retrospective data, for the time period 2007-2011, from the University of Iowa Hospitals and Clinics (UIHC), containing about 154K patient visits and associated demographic fields and rich spatio-temporal information on procedures, antibiotics, comorbidities, within-hospital transfers, etc. It is known that risk factors for (symptomatic) CDI include age, length of hospital stay, recent prior hospital admission, use of certain antibiotics considered high-risk for CDI, use of proton pump inhibitors, and severity of other comorbidities [9]. However, much less is known about the risk factors for asymptomatic C. diff carriage. Our first finding is that a predictive model for inferring asymptomatic C. diff carriage that uses all the (above mentioned) features that are risk factors for symptomatic CDI, *except for high-risk antibiotics* has good performance, relative to other models we consider. Specifically, excluding antibiotics as a risk factor seems to lead to a model with better performance than the model obtained by including antibiotics as a risk factor. This is an intriguing data-driven finding that is consistent with [16], where antibiotic use is not listed as a risk factor for asymptomatic C. diff carriage. However, as mentioned earlier, there is a lot unknown about risk factors for asymptomatic C. diff carriage and in other literature (e.g., [8]) the Cephalosporin class of antibiotics were found to be a risk factor for asymptomatic C. diff carriage.

The key difficulty in training and testing a predictive model for asymptomatic C. diff carriage is that we do not have any "ground truth" data, i.e., we have no labels identifying certain patients as being asymptomatic C. diff carriers. Our data – like most large-scale inpatient data from hospitals – only contain information on patients who tested positive for CDI, and these tests are invariably administered to patients who show symptoms. We overcome this missing label problem in two ways. First, we consider two alternative hypotheses on the relationship between CDI and asymptomatic C. diff carriage and use these hypotheses to generate a number of different prediction models for asymptomatic C. diff carriage. Second, we test out models indirectly by viewing these models for predicting asymptomatic C. diff carriage as constituting the first stage in a 2-stage model. We design the Stage 2 model for predicting *symptomatic* CDI cases. Inspired by the approach in [7, 9, 30], we use measures of exposure to asymptomatic C. diff carriers identified by the Stage 1 model as features in the Stage 2 model. Our second finding is that a model that includes exposure to asymptomatic C. diff carriers outperforms models that don't include this exposure. This finding simultaneously shows two things. First, it reveals the predictive power of our Stage 1 models and identifies Stage 1 models that outperform other models (e.g., the Stage 1 model that uses all CDI risk factors except for antibiotics). Second, it shows that exposure to asymptomatic C. diff carriers is a salient risk factor for CDI, something that has been conjectured widely in CDI literature [10, 29].

Additionally, we also investigate spatio-temporal clustering of the cases inferred to be C. diff carriers by our model. In prior work [27], we have shown that CDI cases at the UIHC exhibit spatio-temporal clustering. Using similar statistical tests, we show here that the observed CDI cases along with the inferred asymptomatic C. diff carriers also exhibit spatio-temporal clustering. This finding provides additional indirect evidence that in-hospital exposure to asymptomatic C. diff carriers may be playing a role in the spread of CDI in the hospital.

1.2 Other Related Work

Besides the papers cited earlier, there are two computational approaches to the problem of inferring asymptomatic cases, that are worth mentioning here. Makar et al. [20] define a generative probabilistic model for problem of inferring asymptomatic cases and their impact on other agents via exposure. Their main contribution is a computational method for solving for the parameters of the model. A different strand of research uses [28, 31, 32] the Steiner tree problem as a model for the problem where some nodes in a contact network are observably infected (i.e., symptomatic) and the infection status of other nodes is latent.

2 THE STAGE 1 MODEL: INFERRING ASYMPTOMATIC C. DIFF CARRIERS

2.1 The UIHC DataSet

The data used in this paper consist of anonymized electronic medical records (EMR) and admission-discharge-transfer records (ADT) for patient visits at the UIHC for the period 2007–2011. The 154,230 patient visits in the data are divided into two groups: (i) $visit_{CDI}$, visits during which patients tested positive for CDI and (ii) $visit_{CDIx}$, the rest of the visits. As in [11, 24, 30], we exclude short visits in both $visit_{CDI}$ and $visit_{CDIx}$, where patients are discharged within 48 hours of admission. The reason for excluding short visits from $visit_{CDI}$ is that such CDI cases are unlikely to be hospital-associated and the reason for then excluding short visits from $visit_{CDIx}$ is that otherwise the length of a visit field might end up being a prominent artificial signal of a non-CDI visit. For each visit in $visit_{CDIx}$, we generate one *instance* per day (*CDIx instances*) from the admission date to discharge date for that visit. Similarly, we generate daily instances (*CDI instances*) for each visit in $visit_{CDI}$, starting from the admission date, but only until three days before the CDI positive test date [22]. We exclude instances for the last three days

before a positive CDI test because there could be modifications to patient treatment during this period that could be in response to potential CDI. This process results in 8,946 CDI instances from 750 visits in $visit_{CDI}$ and 988,780 CDIx instances from 115,271 visits in $visit_{CDIx}$.

2.1.1 Individual risk factors for CDI. We include in each instance, 25 features extracted from the EMR and ADT data, which are considered risk factors for CDI in literature [7, 9]: length of stay of the visit until the date of the instance (*LOS*), *age*, *gender*, previous UIHC visit within 60 days (*PV*), the number of high-risk antibiotics prescribed (*ABX*s) and the number of gastric acid suppressors prescribed (*GAS*s) during the visit. Guided by literature on antibiotics that are considered high risk for CDI [25], we use the following five *ABX*s as features: (i) Amoxillin or Ampicillin (ABX 1), (ii) Clindamycin (ABX 2), (iii) Third generation Cephalosporin (ABX 3), (iv) Fourth genenration Cephalosporin (ABX 4), and (v) Fluoroquinolone (ABX 5). Similarly, guided by literature on risk factors for CDI [6], we use the following two *GAS*s as features: (i) H2-receptor antagonists (GAS 1), and (ii) proton pump inhibitors (GAS 2). We generate three features each for the seven medications (*ABX*s and *GAS*s): (i) prescription ($P_{medication}$), a binary feature, indicating if the medication was prescribed on the date of the instance, (ii) sum prescription count ($SP_{medication}$), number of days where the medication was prescribed to the patient, and (iii) mean prescription count ($MP_{medication} = \frac{SP_{medication}}{LOS}$) of the medication. We use ABX_x for $x \in \{1, 2, 3, 4, 5\}$ to denote the tuple $(P_{ABX_x}, SP_{ABX_x}, MP_{ABX_x})$ correspoding to ABX x. Similarly, we use GAS_x, for $x \in \{1, 2\}$ to denote the tuple $(P_{GAS_x}, SP_{GAS_x}, MP_{GAS_x})$.

2.1.2 Exposure risk factors for CDI. Colonization pressure is a measure of the proportion of patients infected or colonized with a specific pathogen in a specific physical area (e.g., a hospital ward or a geographic region) over a specified period of time [2]. Colonization pressure serves as a proxy measure for exposure, and the notion of colonization pressure has also been applied to CDI, albeit only those patients who have tested positive for CDI are included in the pressure calculation [7, 30]. Colonized patients who are asymptomatic are typically undetected and are usually excluded from pressure calculations. As has been done in other studies [7, 30], we compute this modified measure of colonization pressure, which we call *CDI pressure* and use it as an exposure risk factor for CDI.

We assume that CDI patients are infectious 3 days before the positive result and up to 14 days after the test date. For each visit in $visit_{CDI}$ and $visit_{CDIx}$, we keep track of the number of infectious CDI patients in the same room or unit, daily. From these counts, we generate the following four features:

- *Unit sum CDI pressure* (SCP_{unit}): cumulative daily number of infectious CDI patients in the same unit, from admission date up to the date of the instance
- *Room sum CDI pressure* (SCP_{room}): cumulative daily number of infectious CDI patients in the same room from admission date up to the date of the instance
- *Unit mean CDI pressure* (MCP_{unit}): $\frac{SCP_{unit}}{LOS}$
- *Room mean CDI pressure* (MCP_{room}): $\frac{SCP_{room}}{LOS}$

Table 1 summarizes basic statistics of these features for CDI visits and CDIx visits.

2.2 Training the Stage 1 Model

The goal of our Stage 1 model is to predict the likelihood of an individual being an asymptomatic C. diff carrier, as a function of certain hand-curated risk factors. As mentioned earlier, the fundamental obstacle to training this model is the fact that our data lacks "ground truth" labels. So the training of our Stage 1 model depends on hypotheses we make regarding how asymptomatic C. diff carriers relate to patients who have tested positive for CDI. The first hypothesis we consider is the following.

> **Hypothesis 1**: Asymptomatic C. diff carriers and CDI cases have similar risk profiles.

This hypothesis is not necessarily backed by studies in the literature; as mentioned earlier, the risk factors for asymptomatic C. diff carriage and the progression from C. diff carriage to CDI is not well understood. We propose this as a simple, reasonable hypothesis that allows us to train C. diff carriage prediction models that we can then evaluate. If we assume this hypothesis, we can train our Stage 1 model using CDI cases as instance labels. Then, patients who are assigned a high probability by a model trained in this manner, but are not CDI cases, are inferred to be asymptomatic C. diff carriers. Variants of this Stage 1 model can be obtained by using different subsets of features. More specifically, we partition the set of features into three groups: (i) *baseline* feature set B, consisting of *LOS*, *age*, *gender*, *PV*, GAS_1, and GAS_2, (ii) *colonization pressure* feature set CP, consisting of SCP_{unit}, SCP_{room}, MCP_{unit}, and MCP_{room}, and (iii) *ABX* feature set ABX, consisting of the 5 high-risk antibiotic feature tuples described earlier. For a subset

Table 1: Basic statistics of features. The values denote mean over each visit in $visit_{CDI}$ or $visit_{CDIx}$ and values in the bracket denote std. dev. For most of the features the values for $visit_{CDI}$ are much larger than the corresponding values for $visit_{CDIx}$ (e.g., LOS: 10.93 vs 7.58).

Feature	$visit_{CDI}$	$visit_{CDIx}$
LOS	10.93 (23.09)	7.58 (11.14)
age	53.5 (23.23)	44.23 (24.9)
$gender$	0.55 (0.5)	0.48 (0.5)
PV	0.35 (0.48)	0.19 (0.39)
SP_{GAS1}	1.71 (5.54)	0.92 (3.37)
SP_{GAS2}	5.81 (13.98)	2.98 (6.37)
MP_{GAS1}	0.17 (0.34)	0.11 (0.28)
MP_{GAS2}	0.42 (0.41)	0.33 (0.4)
SP_{ABX1}	0.46 (2.37)	0.48 (2.37)
SP_{ABX2}	0.1 (1)	0.05 (0.57)
SP_{ABX3}	0.39 (2.07)	0.21 (1.23)
SP_{ABX4}	1.2 (3.55)	0.24 (1.58)
SP_{ABX5}	1.58 (4.68)	0.73 (2.47)
MP_{ABX1}	0.04 (0.15)	0.05 (0.19)
MP_{ABX2}	0.01 (0.06)	0 (0.04)
MP_{ABX3}	0.04 (0.16)	0.03 (0.13)
MP_{ABX4}	0.1 (0.26)	0.02 (0.13)
MP_{ABX5}	0.11 (0.23)	0.08 (0.21)
SCP_{unit}	1.47 (3.18)	2.16 (4.84)
SCP_{room}	0.03 (0.23)	0.08 (0.75)
MCP_{unit}	0.23 (0.46)	0.26 (0.47)
MCP_{room}	0.01 (0.1)	0.01 (0.06)

$S \subseteq \{B, CP, ABX\}$, let D^S denote the dataset with every CDI and CDIx instance consisting of features from S. We train 4 different Stage 1 models using datasets $D^B, D^{B,CP}, D^{B,ABX}, D^{B,CP,ABX}$.

We train additional Stage 1 models on the basis of the following hypothesis.

> **Hypothesis 2**: The mechanism for acquiring (symptomatic) CDI consists of the patient first being an asymptomatic C. diff carrier and then being prescribed high-risk antibiotics.

Again, this hypothesis is not necessarily backed by medical studies, though mechanistic models for CDI (e.g., [33]) often attribute the transition from C. diff carriage to CDI to the use of additional high-risk antibiotics. This hypothesis has the following useful implication. Suppose A is the subset of patients who were prescribed high-risk antibiotics during their visit. Then, the subset $A_{CDI} \subseteq A$, consisting of patients who tested positive for CDI is exactly identical to the subset of A of patients who were asymptomatic C. diff carriers (prior to receiving antibiotics) and $A \setminus A_{CDI}$ is exactly the subset of A of patients who are not asymptomatic C. diff carriers. This motivates the restriction of our data set to just those daily instances where patients are prescribed to at least one ABX since admission. When a model is trained on this subset of data, the instances in $visit_{CDIx}$ that the model assigns the `True` label are inferred to be asymptomatic C. diff carriers. 5,483 CDI instances out of 359 visits from $visit_{CDI}$ and 374,821 CDIx instances out of 35,002 visits from $visit_{CDIx}$ result from this restriction. Using this restricted data set, we train 4 additional Stage 1 models using datasets $D^B_{ABX>0}, D^{B,CP}_{ABX>0}, D^{B,ABX}_{ABX>0}, D^{B,CP,ABX}_{ABX>0}$ that are obtained by considering different subsets of features.

2.2.1 Model training. Each dataset of instances mentioned in the previous section contains timestamped instances for the 5-year period 2007–2011. For each dataset, we build five prediction models, each model obtained by training on a 4-year subset, with one year excluded. Recall that the labels in our datasets correspond to a positive CDI test, whereas our goal for each model is to predict the likelihood of a patient being an asymptomatic C. diff carrier. For each dataset, a multi-layer perceptron model (MLP) is trained on the instances in 4 years (we use 20% of instances as a validation set, not used in training), and tested on the instances in the remaining year. We train a two-layer MLP, with a hidden layer size of 16, ReLU activation, and drop out of 0.5 using the Adam optimizer with a learning rate of 0.01 and maximum training for 200 epochs, but with an early stopping if the validation loss does not decrease for 3 consecutive epochs.

After the training and testing of the five models is completed, for each instance (a day during a patient visit), we have a probability that we interpret to be the likelihood of that patient being an asymptomatic C. diff carrier on that day. We now assign to each visit in $visit_{CDIx}$, the maximum probability of all the instances from the visit. We interpret this probability as the likelihood that the patient was a C. diff carrier during this visit. Our next step is to use these probabilities to mark a subset of the visits as being C. diff carrier visits. According to a survey [14] of studies on the prevalence of C. diff carriage, 0–17.5% of healthy adults were carriers of C. diff strains without clinical signs of CDI. Keeping this range in mind, we separately select the top 10%, top 5%, and top 3% of the visits in $visit_{CDIx}$ by probability and designate these sets of visits as $visit_{ACDI10\%}$, $visit_{ACDI5\%}$, and $visit_{ACDI3\%}$, respectively. Note that we have 8 different Stage 1 models, which means we have 8 different sets of $visit_{ACDI10\%}$, $visit_{ACDI5\%}$, and $visit_{ACDI3\%}$ as a result.

3 EVALUATING ASYMPTOMATIC C. DIFF CARRIER PREDICTIONS

The output of the Stage 1 Model is a subset of patient visits that are marked with the patient being an asymptomatic C. diff carrier during the visit. Note that the patients do not have a positive CDI test during these visits. As mentioned earlier, the key difficulty in evaluating this inference is that we do not have "ground truth" labels for asymptomatic C. diff carriers. We propose two indirect ways of validating and evaluating our Stage 1 model predictions.

(i) We design a 2-stage model for predicting symptomatic CDI cases that uses, in addition to standard risk factors of CDI, features that measure exposure to asymptomatic C. diff carriers (as predicted by our Stage 1 model). We investigate if this 2-stage model has improved performance due to inclusion of these additional exposure features. Furthermore, this framework also allows us to indirectly compare different Stage 1 models, by virtue of how well the 2-Stage model using that particular Stage 1 model performed.

(ii) We perform statistical tests to determine if the collection of CDI cases and asymptomatic C. diff carriers (as inferred by our Stage 1 model) exhibit spatio-temporal clustering. In our prior work [27], we observed statistically significant spatio-temporal interaction and clustering of CDI cases at the UIHC. Note that these were just the cases with a positive CDI test. We interpreted this finding as providing evidence of the within-hospital spread of CDI. A similar result for the collection of cases that additionally includes asymptomatic C. diff carriers will provide evidence that asymptomatic C. diff carriers also have a role to play in the within-hospital spread of CDI.

3.1 Training the Stage 2 Model

We now design a CDI prediction model that includes exposure to asymptomatic C. diff carriers (as predicted by our Stage 1 model) as features. We investigate the question of whether including these exposure features improves the CDI model prediction.

In Section 2.1.2 we defined 4 different measures, called CDI pressures, of exposure to CDI cases. In a similar manner, we define 4 measures of exposure to asymptomatic C. diff carriers. We start by assuming that any patient designated to be an asymptomatic C. diff carrier during a visit is infectious throughout the visit. This assumption leads to the following definition of *asymptomatic C. diff carrier pressures AP*, consisting of SAP_{unit}, SAP_{room}, MAP_{unit}, and MAP_{room}.

- Unit sum asymptomatic C. diff pressure (SAP_{unit}): cumulative daily exposure to asymptomatic C. diff carriers detected in the Stage 1 model in the same unit from admission date up to the date of the instance

- Room sum asymptomatic C. diff pressure (SAP_{room}): cumulative daily exposure to asymptomatic C. diff carriers in the same room from admission date up to the date of the instance
- Unit mean asymptomatic C. diff pressure (MAP_{unit}): $\frac{SAP_{unit}}{LOS}$
- Room mean asymptomatic C. diff pressure (MAP_{room}): $\frac{SAP_{room}}{LOS}$

Figure 1 shows the interaction between Stage 1 and Stage 2 models.

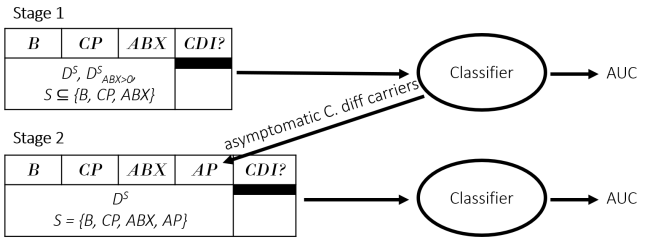

Figure 1: Diagram of the 2-stage model

In Section 2, we defined 8 different models for predicting asymptomatic C. diff carriers, 4 for each of the two hypotheses. From each of these 8 models, we get a different set of 4 exposure features, representing exposure to asymptomatic C. diff carriers. As a result, we evaluate 8 different Stage 2 models (Table 3) and for comparison we also evaluate one Stage 1 model (Table 2) without any feature corresponding to exposure to asymptomatic C. diff carriers.

3.2 Spatio-temporal Clustering of Symptomatic and Asymptomatic CDI Cases

In the previous work, we created a hospital graph of UIHC using room and spaces in a corridor as nodes (19K) and direct passage between node pairs in the 5-6m distance as edges (47K) [5]. We associate with each CDI case a timestamp (date of positive CDI test) and a location (room the patient was in at the time of positive CDI test). Two CDI cases are said to be in *spatio-temporal proximity* if the two cases occurred within 14 days of each other in rooms that are (roughly) within 30 m apart from each other, which is within 5 hop distance in the hospital graph [26]. This notion is conveniently described to be a *CDI case proximity graph* $G_{CDI} = (V_{CDI}, E_{CDI})$, where V_{CDI} is the set of CDI cases at the UIHC during the period 2007–2011 and E_{CDI} is the edges that connects pairs of CDI cases in spatio-temporal proximity. Note that CDI cases that tested positive for CDI within 48 hours of admission are not included in V_{CDI} because these cases are unlikely to be acquired during the hospital visit. We can generalize the notion of CDI case proximity graph in a natural way to include asymptomatic C. diff carriers. With each patient visit marked as an asymptomatic C. diff carrier case by our Stage 1 model, we associate a date, which is the date during the visit that was assigned the highest probability of being an asymptomatic C. diff carrier. Once a date is assigned to a visit, we can also associate a location to the visit, which is the room occupied by the patient on that date. For $x \in \{3, 5, 10\}$, let Let $G_{RCDIx\%} = (V_{RCDIx\%}, E_{RCDIx\%})$ denote the *revealed CDI case proximity graph*. Here $V_{RCDIx\%}$ is the union of the set of CDI cases and the set of asymptomatic C. diff cases output by our Stage 1 model when it was required to mark $x\%$ of visits in $visit_{CDIx}$ as asymptomatic C. diff carrier visits. Among the 8 sets of asymptomatic C. diff cases from 8 different Stage 1 models, we select the set of cases where adding the exposure features from these cases yields the *best* performance on the Stage 2 model. $E_{RCDIx\%}$ is the set of edges connecting pairs of nodes in $V_{RCDIx\%}$ that are in spatio-temporal proximity.

We compute a number of basic network statistics of $G_{RCDIx\%}, x \in \{3, 5, 10\}$ and compare these with corresponding statistics for G_{CDI} (Table 6). We then compute specific measures of network density and make a similar comparison (Table 7). Finally, we perform statistical tests on $G_{RCDIx\%}, x \in \{3, 5, 10\}$ (e.g., Knox test [17]) for testing if the union of the set of CDI cases and the set of asymptomatic C. diff cases exhibit spatio-temporal clustering. The results from these computations are described in Section 4.

4 RESULTS

4.1 Stage 1 Model

Table 2 summarizes the performance of the 8 Stage 1 Models, 4 models derived from each hypothesis (see Section 2). Recall that even though the purpose of these models is to predict asymptomatic C. diff carriers, they are trained on labeled data, where the labels indicate CDI. Table 2 shows how well these models are able to predict CDI. As an evaluation measure, we report AUC, the area under the receiver operating characteristic (ROC) curve, as the evaluation metric for our models since AUC is widely used as an evaluation metric for an imbalanced dataset. Note that our datasets are highly imbalanced: the imbalance ratio of datasets D^S and $D^S_{ABX>0}$, $S \subseteq \{B, CP, ABX\}$ is 111:1 and 68:1, respectively, which makes model training challenging. The AUCs reported in Table 2 are the test AUCs averaged over five years of training and testing on each dataset; this procedure is similar to k-fold cross-validation, but each fold corresponds to the instances in the same year. The 8 columns on the right of the table correspond to the 8 different models, as indicated by the column labels. The best performing Stage 1 Model is the one trained on $D^{B,ABX,CP}$, with a mean AUC of 0.719. This is not surprising because this model uses features of all the standard risk factors for symptomatic CDI. The next best model is the one trained on $D^{B,CP}$, with a mean AUC of 0.704. This result shows that *ABX* helps the prediction of symptomatic CDI. Again this is not surprising because high-risk antibiotics play an important role in predictive models for CDI. Exposure to CDI patients consistently help the prediction, as revealed by pairwise comparisons of models trained on features that use CDI pressures vs. those that do not use CDI pressures, e.g. $D^{B,ABX,CP}$ and $D^{B,ABX}$. The overall AUCs from the models trained on $D^S_{ABX>0}, S \subseteq \{B, CP, ABX\}$ is smaller compared to those trained on $D^S, S \subseteq \{B, CP, ABX\}$, though this comparison may not be fair since $D^S_{ABX>0}$ has a smaller set of instances compared to D^S.

Table 2: AUC on Stage 1 models

	D^B	$D^{B,ABX}$	$D^{B,CP}$	$D^{B,ABX,CP}$	$D^B_{ABX>0}$	$D^{B,ABX}_{ABX>0}$	$D^{B,CP}_{ABX>0}$	$D^{B,ABX,CP}_{ABX>0}$
AUC	0.676	0.635	0.704	*0.719	0.594	0.584	0.672	0.648

[a] AUC with asterisk denote best performer for $D^S, S \subseteq \{B, CP, ABX\}$

4.2 Stage 2 Model

The results of Stage 2 models are shown in Table 3. Each AUC in the table corresponds to the mean test AUC averaged over five years of training and testing on each dataset. As denoted by the label at the top (in the first row), every Stage 2 model evaluated here is trained on $D^{B,ABX,CP,AP}$, i.e., the dataset consisting of all risk factors for symptomatic CDI (*B*, *ABX*, and *CP*) along with *asymptomatic pressures AP*. The 24 models shown in this table differ in how asymptomatic C. diff carriers are identified in Stage 1. The 8 column labels in the second row on the right of the table correspond to the 8 different models on which the $visit_{ACDI10\%}$, $visit_{ACDI5\%}$, and $visit_{ACDI3\%}$ (bottom 3 rows in the table) are detected, as indicated by the column labels. The most important takeaway from this table is that using $D^{B,CP}$ as the dataset during Stage 1 consistently leads to the best performance. In other words, a model that uses baseline features (*B*) and colonization pressure features (*CP*), but not high-risk antibiotic features (*ABX*) to identify asymptomatic C. diff carriers, seems to most accurately identify C. diff carriers. This intriguing finding that is consistent with [16], seems to indicate that antibiotics that are risk factors for CDI are not associated with asymptomatic C.diff carriage.

The three Stage 2 models corresponding to $D^{B,CP}$ (AUC: 0.733, 0.729, 0.727) outperform the best performing Stage 1 model ($D^{B,ABX,CP}$, AUC: 0.719), clearly indicating that exposure to asymptomatic C. diff carriers impacts the spread of CDI. Most of the remaining Stage 2 models perform even worse than the Stage 1 model using $D^{B,ABX,CP}$. In other words, using exposure to asymptomatic C. diff carriers is worse than not using such exposure features, if asymptomatic C. diff carriers are detected poorly. Table 5 shows the *AP* of $visit_{CDI}$ and $visit_{CDIx}$ that is computed from asymptomatic C. diff carriers which are detected in Stage 1 Model on $D^{B,CP}$.

As a sensitivity test of our Stage 2 models, we train models on additional datasets that contain as features, exposure to *randomly* selected visits in $visit_{CDIx}$, instead of *AP*. We randomly select 10% of the visits in $visit_{CDIx}$, and generate 4 exposure features from these visits (*RP*) in the same manner as the *AP* features were generated. We repeat this five times to generate five different sets of random exposure features (*RP*s), namely $Random_i, i \in \{1 \cdots 5\}$. The results are in Table 4. The mean AUCs on these Stage 2 models are all worse than the AUCs obtained just by using the Stage 1 model on $D^{B,ABX,CP}$. This result shows that adding pressure features from a random subset of visits does not improve the CDI prediction.

4.3 Spatio-temporal Clustering

Table 6 shows the network statistics of G_{CDI} and revealed CDI case proximity graphs $G_{RCDIx\%} = (V_{RCDIx\%}, E_{RCDIx\%})$. Here $V_{RCDIx\%}$ is the union of the set of CDI cases and the set of asymptomatic C. diff cases output by our best-performing Stage 1 model ($D^{B,CP}$), with the requirement that $x\%$ of the visits from $visit_{CDIx}$ are marked as asymptomatic C. diff carrier visits. The number of nodes and edges ($|V|$, $|E|$), average, max and std dev of degrees ($\langle k \rangle$, k_{max}, and std), the clustering coefficient (cc), the average size of connected components $avg(|E_{cpnt}|)$, and the number of nodes and edges of the giant component ($|V_{giant}|$, $|E_{giant}|$), all increase as we add more asymptomatic C. diff cases to the graph.

Figure 2 shows a connected component of $G_{RCDI10\%}$ that contains 7 CDI cases and 48 asymptomatic C. diff carriers over 3 months period (March 21 - July 6 2011). The CDI case (July 6) in the bottom of the graph is only connected to an asymptomatic C. diff carrier (July 1) who has connections to CDI cases (June 18, June 19). This asymptomatic carrier may be attributable to the CDI case that is not directly connected with other CDI cases.

We compared the G_{CDI} and $G_{RCDIx\%}, x \in \{3, 5, 10\}$ on the four different measures of density: (1) $\frac{|E|}{|E*|}$, number of edges / number of possible edges, (2) $\frac{|E|}{|V|}$, number of edges / number of nodes, (3) $\frac{|E_{giant}|}{|V|}$, the size of the giant component / number of nodes, and (4) $\frac{avg(|E_{cpnt}|)}{|V|}$, average size of connected components / number of nodes. All of the density measures were larger in the revealed CDI case proximity graphs compared to those in the CDI case proximity graph. Furthermore, all four density measures of $G_{RCDI10\%}$ were the largest, followed by $G_{RCDI5\%}$ and $G_{RCDI3\%}$, as shown in Table 7.

Table 3: AUC on Stage 2 models

	$D^{B,ABX,CP,AP}$							
AP	D^{B}	$D^{B,ABX}$	$D^{B,CP}$	$D^{B,ABX,CP}$	$D^{B}_{ABX>0}$	$D^{B,ABX}_{ABX>0}$	$D^{B,CP}_{ABX>0}$	$D^{B,ABX,CP}_{ABX>0}$
10%	0.712	0.687	*0.733	0.710	0.700	0.724	0.697	0.703
5%	0.701	0.690	*0.727	0.685	0.693	0.714	0.689	0.702
3%	0.689	0.698	*0.729	0.690	0.710	0.704	0.686	0.711

[a]AUC with asterisk denote best performer

Table 4: AUC on Stage 2 models (pressures are computed from random selection of 10% of the visits in $visit_{CDIx}$)

	$D^{B,ABX,CP,RP}$				
RP	Random1	Random2	Random3	Random4	Random5
AUC	0.703	0.709	0.684	*0.711	0.696

[a]AUC with asterisk denote best performer

Table 5: Statistics of *AP* computed from $visit_{ACDI10\%}$ detected in Stage 1 model trained on $D^{B,CP}$

Feature	$visit_{CDI}$	$visit_{CDIx}$
SAP_{unit}	35.74 (83.24)	34.33 (64.16)
SAP_{room}	2.15 (18.26)	2.91 (16.02)
MAP_{unit}	2.99 (2.9)	3.93 (4.02)
MAP_{room}	0.15 (0.34)	0.27 (0.55)

Table 6: Network statistics

	G_{CDI}	$G_{RCDI3\%}$	$G_{RCDI5\%}$	$G_{RCDI10\%}$
$\|V\|$	783	4241	6546	12310
$\|E\|$	120	4150	10630	37842
$\langle k \rangle$	0.307	1.957	3.248	6.148
k_{max}	4	18	31	47
std	0.581	2.095	3.145	5.195
cc	0.013	0.306	0.443	0.561
$avg(\|E_{cpnt}\|)$	0.179	2.262	5.502	21.141
$\|V_{giant}\|$	8	118	245	1239
$\|E_{giant}\|$	10	232	738	6393

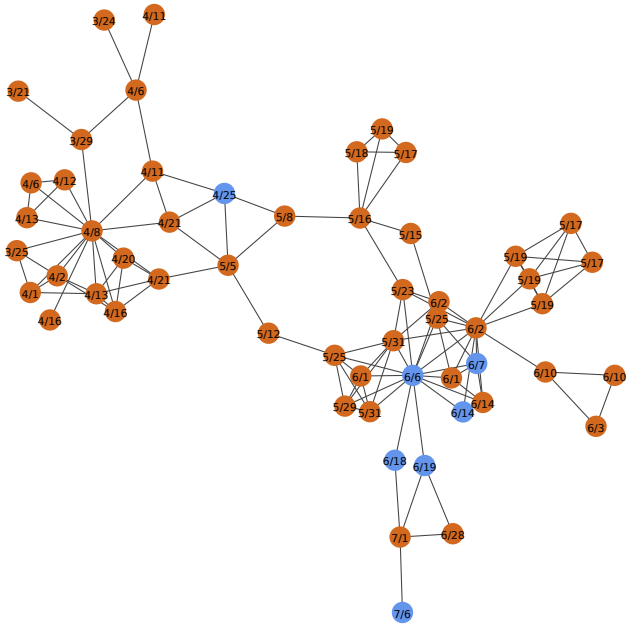

Figure 2: A connected component in $G_{RCDI10\%}$ composed of 7 CDI cases, in blue, and 48 asymptomatic C. diff carriers, in orange, over 3 months period (March 21 - July 6 2011). The CDI case (July 6) on the bottom-most of the graph is not connected to other CDI cases directly, but an asymptomatic C. diff carrier (July 1) connects them to CDI cases (June 18, June 19).

Additionally, we performed statistical tests on $G_{RCDIx\%}, x \in \{3, 5, 10\}$ to test if the union of the set of CDI cases and the set of asymptomatic C. diff cases exhibits spatio-temporal clustering. For each revealed CDI case proximity graph, we performed the Knox test by comparing the number of edges in $G_{RCDIx\%}$ with the distribution of the number of edges in the graphs that are obtained by permuting the timestamp of the cases in $G_{RCDIx\%}$ for random 100 permutations. Similarly, we test the statistical significance of the average size of the largest component ($avg(|E_{cpnt}|)$) and the size of the largest component ($|E_{giant}|$). In Table 8, the p-value of the Knox test and the average size of the largest component was 0 for all of the revealed CDI case proximity graphs that indicate spatio-temporal clustering of the cases. However, we observed that the size of the $|E_{giant}|$ in the permuted graphs is mostly larger than the revealed case proximity graphs of $G_{RCDI5\%}$ and $G_{RCDI10\%}$. Our conjecture regarding this last result is that the time interval of 5 years is not long enough to scatter the timestamps of cases far away from each other.

5 DISCUSSION AND FUTURE WORK

Our results point to several avenues for future work that involve gathering prospective clinical data. Our Stage 1 model for identifying patients who are likely to be asymptomatic C. diff carriers needs to be clinically tested. Designing low-cost clinical protocols for gathering these data and performing appropriate statistical tests is critical in order to have confidence in our results. One of our findings suggests that risk factors for asymptomatic C. diff carriage include most of the standard risk factors, with the exception of high-risk antibiotics. This finding needs to be made more precise and also tested by gathering prospective clinical data.

The datasets that are used in this paper are highly imbalanced with the imbalance ratio of 111:1 and 68:1 for $D^S_{ABX>0}$ and D^S, $S \subseteq \{B, CP, ABX, AP\}$, respectively, that makes the classification problem extremely difficult. To combat its extreme imbalance, we explored undersampling the majority instances in the training set during the training procedures of Stage 1 models; we gained some improvement in the training AUCs, but there was not much of a difference in the testing set AUCs, as we maintained the imbalance in the test set. We aim to explore oversampling strategies such as SMOTE [4] in our future work to improve the overall performance of our classifiers.

In this paper, we only consider the possibility of CDI cases being exposed to asymptomatic C. diff carriers. We do not consider more complicated chains of exposure involving sequences of asymptomatic C. diff carriers. Combining more complicated exposure chains with individual risk models is another avenue for future work. It seems possible to use formulations that involve the Steiner tree problem [28, 31, 32] for this purpose.

Another direction of the future work is using deep embedding approaches, such as Graph Convolutional Networks (GCN) [15] where we let the deep neural network to learn from individual risk factors in the EMR and their exposure to other patients that are captured in the ADT data.

Our asymptomatic C. diff carrier detection method can be applied in other infectious diseases where exposure plays an important role in disease diffusion. It is usually unknown if people we come in contact with are asymptomatic carriers of an infectious diseases. However, if data on an individual's risk factors to an infectious disease, contact information between these individuals, and a subset

Table 7: Network density

	G_{CDI}	$G_{RCDI3\%}$	$G_{RCDI5\%}$	$G_{RCDI10\%}$
$\frac{\|E\|}{\|E*\|}$	0.000392	0.000462	0.000496	0.000499
$\frac{\|E\|}{\|V\|}$	0.153257	0.978543	1.623892	3.074086
$\frac{\|E_{giant}\|}{\|V\|}$	0.012771	0.054704	0.112741	0.519334
$\frac{avg(\|E_{cpnt}\|)}{\|V\|}$	0.000229	0.000533	0.000841	0.001717

Table 8: Statistical test results on $G_{RCDIx\%}$ and the mean values of the statistics on the permuted graphs. Values in brackets denote std. dev.

		$G_{RCDI3\%}$	$G_{RCDI5\%}$	$G_{RCDI10\%}$
p-value	$\|E\|$, Knox test	0	0	0
	$avg(\|E_{cpnt}\|)$	0	0	0
	$\|E_{giant}\|$	0.37	0.99	0.77
statistics	$\|E\|$	3650 (58)	9213 (115)	33790 (223)
	$avg(\|E_{cpnt}\|)$	1.87 (0.04)	4.53 (0.09)	18.56 (0.28)
	$\|E_{giant}\|$	228 (75)	1091 (142)	6620 (325)

of individuals' infectious state is available, then our model would be able to detect the latent spreaders.

6 ACKNOWLEDGMENTS

This project is funded by CDC MInD-Healthcare via CDC cooperative agreement U01CK000531. The authors acknowledge feedback from other University of Iowa CompEpi group members.

REFERENCES

[1] Faisal Alasmari, Sondra M. Seiler, Tiffany Hink, Carey-Ann D. Burnham, and Erik R. Dubberke. 2014. Prevalence and Risk Factors for Asymptomatic Clostridium difficile Carriage. *Clinical Infectious Diseases* 59, 2 (04 2014), 216–222. https://doi.org/10.1093/cid/ciu258

[2] M. J. Bonten, S. Slaughter, A. W. Ambergen, M. K. Hayden, J. van Voorhis, C. Nathan, and R. A. Weinstein. 1998. The role of "colonization pressure" in the spread of Vancomycin-resistant Enterococci: an important infection control variable. *Arch. Intern. Med.* 158, 10 (May 1998), 1127–1132.

[3] Diana C Buitrago-Garcia, Dianne Egli-Gany, Michel J Counotte, Stefanie Hossmann, Hira Imeri, Aziz Mert Ipekci, Georgia Salanti, and Nicola Low. 2020. The role of asymptomatic SARS-CoV-2 infections: rapid living systematic review and meta-analysis. *medRxiv* (2020). https://doi.org/10.1101/2020.04.25.20079103

[4] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.

[5] D.E. Curtis, C.S. Hlady, G. Kanade, S.V. Pemmaraju, P.M. Polgreen, and A.M. Segre. 2013. Healthcare Worker Contact Networks and the Prevention of Hospital-Acquired Infections. *PLOS One* (December 2013). https://doi.org/doi:10.1371/journal.pone.0079906

[6] Sandra Dial, JAC Delaney, Alan N Barkun, and Samy Suissa. 2005. Use of gastric acid–suppressive agents and the risk of community-acquired Clostridium difficile–associated disease. *Jama* 294, 23 (2005), 2989–2995.

[7] E. R. Dubberke, K. A. Reske, M. A. Olsen, K. M. McMullen, J. L. Mayfield, L. C. McDonald, and V. J. Fraser. 2007. Evaluation of Clostridium difficile-Associated Disease Pressure as a Risk Factor for C difficile-Associated Disease. *Archives of Internal Medicine* 167, 10 (05 2007), 1092–1097. https://doi.org/10.1001/archinte.167.10.1092

[8] Erik R. Dubberke, Kimberly A. Reske, Sondra Seiler, Tiffany Hink, Jennie H. Kwon, and Carey-Ann D. Burnham. 2015. Risk Factors for Acquisition and Loss of Clostridium difficile Colonization in Hospitalized Patients. *Antimicrobial Agents and Chemotherapy* 59, 8 (2015), 4533–4543. https://doi.org/10.1128/AAC.00642-15

[9] Erik R Dubberke, Yan Yan, Kimberly A Reske, Anne M Butler, Joshua Doherty, Victor Pham, and Victoria J Fraser. 2011. Development and validation of a Clostridium difficile infection risk prediction model. *Infection Control & Hospital Epidemiology* 32, 4 (2011), 360–366.

[10] David W. Eyre, Madeleine L. Cule, Daniel J. Wilson, David Griffiths, Alison Vaughan, Lily O'Connor, Camilla L.C. Ip, Tanya Golubchik, Elizabeth M. Batty, John M. Finney, David H. Wyllie, Xavier Didelot, Paolo Piazza, Rory Bowden, Kate E. Dingle, Rosalind M. Harding, Derrick W. Crook, Mark H. Wilcox, Tim E.A. Peto, and A. Sarah Walker. 2013. Diverse Sources of C. difficile Infection Identified on Whole-Genome Sequencing. *New England Journal of Medicine* 369, 13 (2013), 1195–1205. https://doi.org/10.1056/NEJMoa1216064 PMID: 24066741.

[11] Centers for Disease Control and Prevention. Jan, 2020 (accessed June 11, 2020). *Identifying Healthcare-associated Infections (HAI) for NHSN Surveillance.* https://www.cdc.gov/nhsn/pdfs/pscmanual/2psc_identifyinghais_nhsncurrent.pdf

[12] Centers for Disease Control and Prevention. Reviewed Nov 13, 2019 (accessed June 10, 2020). *Clostridioides difficile Infection.* https://www.cdc.gov/hai/organisms/cdiff/cdiff_infect.html

[13] Monica Gandhi, Deborah S. Yokoe, and Diane V. Havlir. 2020. Asymptomatic Transmission, the Achilles' Heel of Current Strategies to Control Covid-19. *New England Journal of Medicine* 382, 22 (2020), 2158–2160. https://doi.org/10.1056/NEJMe2009758

[14] Schäffler H. and Breitrück A. 2018. Clostridium difficile - From Colonization to Infection. *Frontiers in Microbiology* 9, 646 (2018). https://doi.org/10.3389/fmicb.2018.00646

[15] Thomas N Kipf and Max Welling. 2016. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv preprint arXiv:1609.02907* (2016).

[16] Ling Yuan Kong, Nandini Dendukuri, Ian Schiller, Anne-Marie Bourgault, Paul Brassard, Louise Poirier, François Lamothe, Claire Béliveau, Sophie Michaud, Nathalie Turgeon, et al. 2015. Predictors of asymptomatic Clostridium difficile colonization on hospital admission. *American journal of infection control* 43, 3 (2015), 248–253.

[17] Martin Kulldorff and Ulf Hjalmars. 1999. The Knox Method and Other Tests for Space-Time Interaction. *Biometrics* 55, 2 (1999), 544–552. http://www.jstor.org/stable/2533804

[18] Lorraine Kyne, Michel Warny, Amir Qamar, and Ciarán P Kelly. 2000. Asymptomatic carriage of Clostridium difficile and serum levels of IgG antibody against toxin A. *New England Journal of Medicine* 342, 6 (2000), 390–397.

[19] Surbhi Leekha, Kimberly C Aronhalt, Lynne M Sloan, Robin Patel, and Robert Orenstein. 2013. Asymptomatic Clostridium difficile colonization in a tertiary care hospital: admission prevalence and risk factors. *American journal of infection control* 41, 5 (May 2013), 390–393. https://doi.org/10.1016/j.ajic.2012.09.023

[20] Maggie Makar, John V. Guttag, and Jenna Wiens. 2018. Learning the Probability of Activation in the Presence of Latent Spreaders. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018.* 134–141. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16980

[21] Kenji Mizumoto, Katsushi Kagaya, Alexander Zarebski, and Gerardo Chowell. 2020. Estimating the asymptomatic proportion of coronavirus disease 2019 (COVID-19) cases on board the Diamond Princess cruise ship, Yokohama, Japan, 2020. *Eurosurveillance* 25, 10, Article 2000180 (2020). https://doi.org/10.2807/1560-7917.ES.2020.25.10.2000180

[22] M.N. Monsalve, S.V. Pemmaraju, S. Johnson, and P.M. Polgreen. 2015. Improving Risk Prediction of Clostridium Difficile Infection Using Temporal Event-Pairs. In *2015 International Conference on Healthcare Informatics.* 140–149. https://doi.org/10.1109/ICHI.2015.24

[23] U.S. Department of Health and Human Services. Jan 15, 2020 (accessed June 10, 2020). *Health Care-Associated Infections.* https://health.gov/our-work/health-care-quality/health-care-associated-infections

[24] Jeeheh Oh, Maggie Makar, Christopher Fusco, Robert McCaffrey, Krishna Rao, Erin E Ryan, Laraine Washer, Lauren R West, Vincent B Young, John Guttag, et al. 2018. A generalizable, data-driven approach to predict daily risk of Clostridium difficile infection at two large academic health centers. *infection control & hospital epidemiology* 39, 4 (2018), 425–433.

[25] Robert C Owens Jr, Curtis J Donskey, Robert P Gaynes, Vivian G Loo, and Carlene A Muto. 2008. Antimicrobial-associated risk factors for Clostridium difficile infection. *Clinical Infectious Diseases* 46, Supplement_1 (2008), S19–S31.

[26] S. Pai, S.V. Pemmaraju, P.M. Polgreen, A.M. Segre, and D.K. Sewell. In press. Spatiotemporal Clustering of In-Hospital *Clostridioides difficile* Infection (CDI). *Infection Control and Hospital Epidemiology* (June In press).

[27] Shreyas Pai, Philip M. Polgreen, Alberto Maria Segre, Daniel K. Sewell, and Sriram V. Pemmaraju. 2020. Spatiotemporal clustering of in-hospital Clostridioides difficile infection. *Infection Control & Hospital Epidemiology* 41, 4 (2020), 418–424. https://doi.org/10.1017/ice.2019.350

[28] Polina Rozenshtein, Aristides Gionis, B. Aditya Prakash, and Jilles Vreeken. 2016. Reconstructing an Epidemic Over Time. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16).* Association for Computing Machinery, New York, NY, USA, 1835–1844. https://doi.org/10.1145/2939672.2939865

[29] García-Fernández S, Frentrup M, Steglich M, Gonzaga A, Cobo M, López-Fresnena N, Cobo J, Morosini MI, Cantón R, Del Campo R, and Nübel U. 2019. Whole-genome sequencing reveals nosocomial Clostridioides difficile transmission and a previously unsuspected epidemic scenario. *Sci Rep.* 9 (May 2019). Issue 1.

[30] N. Khan P.M. Polgeen A.M. Segre D.K. Sewell T. Riaz, A. Kharkar and S.V. Pemmaraju. 2020. Highly Local CDI Pressures As Risk Factors for CDI. To appear in Decennial 2020: 6th International Conference on Healthcare Associated Infections, Atlanta, GA.

[31] Han Xiao, Çigdem Aslay, and Aristides Gionis. 2018. Robust Cascade Reconstruction by Steiner Tree Sampling. In *IEEE International Conference on Data Mining, ICDM 2018, Singapore, November 17-20, 2018.* IEEE Computer Society, 637–646. https://doi.org/10.1109/ICDM.2018.00079

[32] Han Xiao, Polina Rozenshtein, Nikolaj Tatti, and Aristides Gionis. 2018. Reconstructing a cascade from temporal observations. In *Proceedings of the 2018 SIAM International Conference on Data Mining, SDM 2018, May 3-5, 2018, San Diego Marriott Mission Valley, San Diego, CA, USA*, Martin Ester and Dino Pedreschi (Eds.). SIAM, 666–674. https://doi.org/10.1137/1.9781611975321.75

[33] Laith Yakob, Thomas V. Riley, David L. Paterson, and Archie C.A. Clements. 2013. Clostridium difficile exposure as an insidious source of infection in healthcare settings: an epidemiological model. *BMC Infectious Diseases* 13, 376 (2013). https://doi.org/10.1186/1471-2334-13-376