Summer 2015

Computational applications to hospital epidemiology

Mauricio Nivaldo Andres Monsalve
University of Iowa

COMPUTATIONAL APPLICATIONS TO HOSPITAL EPIDEMIOLOGY

by

Mauricio N. Monsalve

A thesis submitted in partial fulfillment
of the requirements for the Doctor of Philosophy
degree in Computer Science in the
Graduate College of
The University of Iowa

August 2015

Thesis Supervisor: Professor Sriram Pemmaraju

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

PH.D. THESIS

This is to certify that the Ph.D. thesis of

Mauricio N. Monsalve

has been approved by the Examining Committee for the
thesis requirement for the Doctor of Philosophy degree
in Computer Science at the August 2015 graduation.

Thesis Committee:

Sriram Pemmaraju, Thesis Supervisor

Alberto Segre

Philip Polgreen

Ted Herman

Padmini Srinivasan

The same experiment which at first glance seemed to show one thing, when more carefully examined, assures us of the contrary.

Galileo Galilei

Discourses and Mathematical Demonstrations Relating to Two New Sciences (1638)

ABSTRACT

Healthcare associated infections are a considerable burden to the health care system. The affected patients have their prognosis worsened and demand more resources from hospitals. Furthermore, the bacteria causing these infections are becoming increasingly resistant to antibiotics while also becoming more deadly and contagious. Contributing with knowledge for stopping these infections is, therefore, important.

This thesis reports on two projects centered on data collected at the University of Iowa Hospital and Clinics. The first project consisted in analyzing data collected by sensors that reported the location and hand washing behavior of health care workers. After extracting meaning from these radio signals, I studied two socially and epidemiologically relevant tasks: the inference of contact networks, which can be used to study the spread of infections in the hospital, and the study of associations between social pressure and hand washing, learning that effectively workers in proximity to others wash their hands more, but also that not all workers are as influential.

In the second project, I developed a data mining method for analyzing medical records aimed at tackling the problems of class imbalance and high dimensionality, and applied it to predicting Clostridium Difficile infection. The learnt models performed better than the state of the art and even improved prediction as the onset of symptoms approached. The main contribution, however, was in the information discovered: certain events in certain orders increased the risk of developing the infection, suggesting that reversing these orders could improve prognosis.

PUBLIC ABSTRACT

Every day, patients get admitted to hospitals for medical attention, but sometimes they get something else: a bacterial infection. As a result, the affected patients become less healthy and require more resources from the hospital. Furthermore, the bacteria causing these infections are becoming increasingly resistant to antibiotics while also becoming deadlier, more contagious, and increasingly harder to kill, making most prevention efforts fall short.

I analyzed data collected at the University of Iowa Hospital and Clinics to discover knowledge that can help health care workers understand and contain these infections. More precisely, I used computational methods for discovering information that would be difficult to elucidate otherwise.

I report two projects in this thesis. In the first, I analyzed data collected through a network of wireless sensors on movement and hygienic habits of workers in a hospital unit. I learned information for building "contact networks", which can help study the spread of infections, and found evidence that workers wash their hands more when coworkers are nearby.

In the second project, I developed a method for analyzing medical records for predicting whether patients will develop the infection caused by the *Clostridium Difficile* bacterium, while also learning which clinical events (for example, operations, prescriptions) put patients at risk. I discovered that some events increase risk if they occur in a specific order, and that high risk patients can be identified on admission.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

HAIs represent a significant burden to healthcare provision. In the US alone, about 2,000,000 patients acquire HAIs every year, resulting in nearly 100,000 deaths per year. Since several of these infections are contagious, researchers have been studying how epidemic outbreaks occur within healthcare facilities, and how to prevent or stop them. With the increasing availability of interaction data and means to analyze it, researchers have turned to contact network epidemiology. Contact network epidemiology, or simply *network epidemiology*, consists in the study of epidemics in populations represented as graphs or networks. It came as a novelty with respect to the previous methods (compartmental models, e.g., the SIR model) because it allowed for the integration of complex interactions, a better understanding of the progression of epidemics, and the possibility of evaluating the effectiveness of targeted vaccination [56, 78], which were impossible with the previous epidemiological models. Also, network epidemiology received great attention from the larger community of complex networks, as epidemic processes are analog to percolation and connected component formation, and vaccination is highly related to network resilience [88].

In this thesis, I reported on two projects centered on data collected at the University of Iowa Hospital and Clinics. The first project consisted in analyzing data collected by sensors that reported the location and hand washing behavior of health care workers, an effort taken by previous members of the compepi group [101, 115, 48]. I analyzed the recorded radio signals to determine location information of healthcare workers and hand washing, a task that involved a great amount of data preprocessing [82]. After making sense of this data, I studied two epidemiologically relevant tasks associated to human behavior. In the first, I studied the problem of link prediction for the inference of contact networks, with the aim of predicting when a working will come in contact (close proximity) to others based on whom they are usually in contact with [81]. This information can be used to build and simulate contact networks, which can be used to study the spread of infections in the hospital, a body of research known as *contact network epidemiology*. This work also served to validate

previous research carried in the compepi group by Curtis et al [16].

Having contact network data is essential for proper network epidemiology. Random graph models, which often offer close-form solutions to epidemiological questions, only account for simplistic interactions. For more general networks, however, there are not simple, general solutions; as a fact, finding the probability that an individual gets infected during a SIR outbreak, and related questions, such as estimating the epidemic size, are NP-hard [107]. This enhances the importance of having real data to study custom situations. Historically, however, there has been little data of intra-hospital contact networks. Due to the unavailability of data, early studies used synthetic network models to study the spread of infection [79, 56, 78]. In spite that contact network epidemiology was proposed more than 30 years ago [45], it has been only in the recent years (mostly since 2011) that research groups started measuring detailed contact network data to study the transmission of infection [118, 50, 5, 74, 59, 4, 75, 120, 36, 19, 60], an effort joined by the compepi group [18, 17, 102, 81, 48, 82, 16, 83].

The other problem I addressed with this data was the study of associations between social pressure and hand washing. By doing so, I found that workers in proximity to others wash their hands more, but also that not all workers were as influential [83]. Hopefully, the results of this research will inform the design of new guidelines to increase hand hygiene adherence.

In the second project I cover in this thesis, I address the problem of using data collected by the hospital (electronic medical records) to measure the risk of developing Clostridium Difficile infection (CDI). CDI occurs when Clostridium Difficile, an antibiotic resistant bacterium, colonizes the intestines of a patient, leading to severe diarrhea and intestinal damage due to the toxins released by the bacterium. In the worst cases, CDI may turn into toxic megacolon (the colon enlarges dramatically due to inflaction), which may require surgery, and even death. For this project, I extended an existing database which contained clinical, architectural, and computer usage data, that was initiated by previous members of the compepi group [18, 17, 102, 16].

For the problem of predicting CDI, I developed a data mining method aimed at tackling the problems of class imbalance and high dimensionality, and applied it to predicting CDI. The developed method consisted in building an ensemble of logistic regression models that, far different from existing developments, perform feature selection at the individual classifier level as opposed to the ensemble level. The models resulting from this methodology performed better than the state of the art models. They also improved prediction as the onset of symptoms approached, i.e., produced dynamically changing risk curves of the development of CDI. The main contribution, however, was in the information discovered: certain events in certain orders increased the risk of developing the infection, suggesting that reversing these orders could improve prognosis.

The overall line connecting my work has been the application of computational methods to epidemiological problems (i.e., *computational epidemiology*). So far, the most related discipline is biostatistics, which is concerned on analyzing data medical and public health data. So, there is room for the use of computational methods for analyzing such data, especially if they involve data that does not follow the usual statistical assumptions, such as radio signals, that follow the usual representations, such as dynamic networks and temporarily organized events (and partial orders), or falls in the big size or high dimensionality domain. Thus, perhaps, there is need for more research in *computational epidemiology* for the improvement of the quality of healthcare provision.

I left out of my dissertation the early research on syphilis in the US [84] and results regarding to my previous research [80], because they are outside the theme of my current work.

CHAPTER 2
THE MICU SENSOR NETWORK DEPLOYMENT

2.1. Introduction

To measure both the interactions between healthcare workers and hand hygiene adherence (tightly related to in-hospital epidemics), the compepi group deployed a wireless sensor network in the 20-bed Medical Intensive Care Unit (MICU) of the UIHC from June 1 to June 10, 2011 [81, 82]. We continuously measured healthcare worker location information as well as their proximity with respect to each other, and recorded alcohol dispenser usage. The collected data tells a story of contact patterns within the MICU, interactions between healthcare workers and patients, *hand hygiene adherence*, and how the proximity to other workers affected this *adherence*. The experiment was a follow-up to a previous, similar deployment by the compepi group in the same unit during 2010 [48].

To keep in mind, by *hand hygiene adherence* we mean the likelihood healthcare workers washed their hands at the specific moments they should do so. Guidelines for hand hygiene state that workers must comply with these rules to prevent the spread of infection. The moments at which the worker must wash their hands are called *opportunities* in the literature. The opportunities we are concerned with are the moment before entering a patient room, and the moment right after exiting that room. These two opportunities, which we call *at entry* and *at exit*, are the opportunities mentioned in the guidelines of the CDC.

In this chapter, I recount my work related to processing the signals recorded in the 2011 deployment. This work has lead to several publications [81, 82, 83]. Still, some of the developments presented here have not taken part in publications.

2.1.1. The MICU sensor network deployment

(Adapted from [83].) We deployed a wireless sensor network to measure interactions between HCWs (e.g., close proximity contacts), their individual location (e.g., "inside patient room", "in hallway", "at nurses' station", etc.) and hand-hygiene activity (i.e., alcohol dispenser usage) in the Medical Intensive Care Unit (MICU) of the University of Iowa Hospital

and Clinics (UIHC) for 10 consecutive days. This sensor network consisted of small pager-sized wireless sensors or motes. These programmable, battery-powered devices consist of a small processor with flash memory and an IEEE 802.15.4-compliant wireless radio. We programmed the motes to broadcast a brief message every 7 to 12 seconds. When received by other motes within range, these messages encode the unique identifier of the sender mote, the received signal strength index (RSSI) associated with the message (a proxy for distance, as RSSI increases with proximity), and the time the message was received. These data were recorded in the flash memory of the receiving mote for later analysis. The motes' radios communicated through an unused portion of the Wi-Fi spectrum to avoid interfering with medical equipment.

Our wireless sensor network consisted of stationary sensors or beacons, and wearable sensors or badges. Beacons were placed inside all 20 patient bedrooms as well as outside rooms (e.g., in hallways, at nurses' stations) throughout the unit, as depicted in Fig. 2.1. This spatial grid of sensors served to locate workers in the unit from the collected data. Beacons also included instrumented alcohol dispensers that broadcast messages whenever their pumps were used. We only instrumented and considered alcohol dispensers located immediately outside every room, ignoring the dispensers sporadically located within patient rooms.

Badges were worn by HCWs and were collected from and distributed to workers at the beginning of each shift. HCWs were divided into three different job types:

1. *Doctors*: staff physicians, fellows and residents;

2. *Nurses*: MICU nurses, nurse assistants and nurse managers; and

3. *Critical-care support*: clerks, pharmacists and respiratory therapists.

Badges were assigned randomly to workers within each job type, ensuring that individual workers could not be identified, and workers could not be tracked across different shifts. The different types of sensors nodes used are depicted in Fig. 2.2.

Figure 2.1: Placement of all the sensors in the MICU, June 2011. Symbols: triangles represent pyramids, circles represent alcohol dispensers, and squares represent *door minders*. Darker symbols represent in-room sensors. Floor patterns: patient bedrooms are chess-tiled, worker spaces (nurse stations and physician workrooms) are diagonal-stripped, corridors are white, and unmeasured areas are dark shaded.

Since the deployment was part of a process-improvement project and no patient information was collected, it was ruled *non-human-subjects research* by the University of Iowa's Institutional Review Board.

2.1.2. Completed research

My work on processing and analyzing the sensor data consists in three well separated parts. First, I gave considerable attention to preprocessing the data, so any analysis done on it is meaningful. The data collected from the sensors suffered from strong variability, noticeable biases, and absence of tagged ground truth dataset.

The second part consisted in the problem of predicting, using room entry information, whether healthcare workers were in contact (close proximity). This work came to validate previous work done in the compepi group.

And the third part consisted in the study of the presence of *peer effects* with respect to hand hygiene in the unit, i.e., whether proximity to other healthcare workers had any effect on the proneness to wash hands. And we found evidence of such effects.

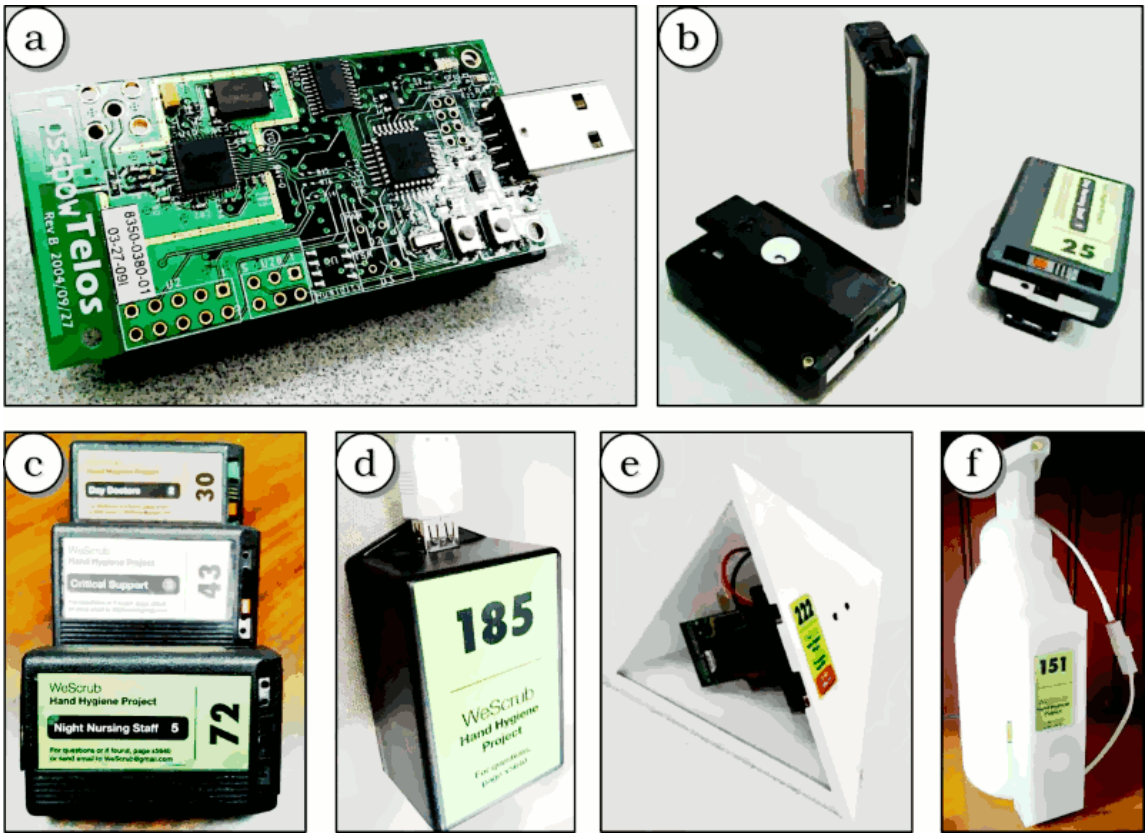


Figure 2.2: Sensors used in the deployment: (a) Crossbow's Telos-B mote, (b) badges in 3 different angles, (c) the color of the badge's label indicates job type, (d) listener mote, (e) pyramid (beacon), (f) soap/alcohol dispenser (beacon).

2.2. Similar experiences

Few works have used contact networks generated from sensor network data in healthcare settings. The SocioPatterns research group has done extensive work on this. They have collected network data from: 1 week in a pedriatic ward [50, 75, 36], 4 days in a geriatric unit [120], and 3 months of 50 rooms in two hospitals [74]. The other deployments were the ones performed by the compepi group in the MICU of the UIHC, first in 2010 [48] and later in 2011 [81, 82, 83].

There is little literature on experiences in processing data from such deployments. Kazandjieva et al reported difficulties associated to sensor malfunction due to human factors, and ordered the recorded data by time using reference orders instead of clock synchronization [55]. Cattuto et al deployed a network of sensors that could not transmit

further than 1-1.5 meters, which limited and simplified data processing [50, 120]. Friggery et al deployed a sensor network more similar to ours, but used a simplified visit detection algorithm [32, 74].

2.3. Processing the data

2.3.1. Main challenges raised by the data

One of the salient features of the RSSI values collected was their high variability. Using moving averages seemed unsuitable given that the sample rate was 8 seconds, time much larger than the one needed by a normal person to move enough to completely change the RSSI values.

Secondly, the RSSI values in reciprocal communication (from A to B and from B to A) seemed to suffer from consistent biases, as shown in Fig. 2.3. This ultimately meant that the RSSI values were in *different scales*, in other words, that the relation of RSSI and distance was different across different motes.

The final challenge, and the most important one, was the unavailability of experimental data on RSSI values at different distances that would allow a characterization of RSSI values, variability and all, changing with distance (especially at short range). Indeed, such data was missing, which also implied the need to calibrate the RSSI values using the main experiment's data. Still, we had data on RSSI values from the bedmotes for when the badges were inside a room.

I first address the problem of calibration, then move to the variability issue (designing a filter), and finally address the problems associated to RSSI and distance (detecting contacts, when a worker visited a room, etc.).

2.3.2. Calibration (normalization) algorithm

(Adapted from [82].) The RSSI calibration procedure starts from the simple model

$$x_{AB}^{A}(t) \approx s_A x_{AB}(t),$$

where x_{AB}^{A} are the RSSI read by A and sent from B, x_{AB} are the ideal RSSI values (unbiased

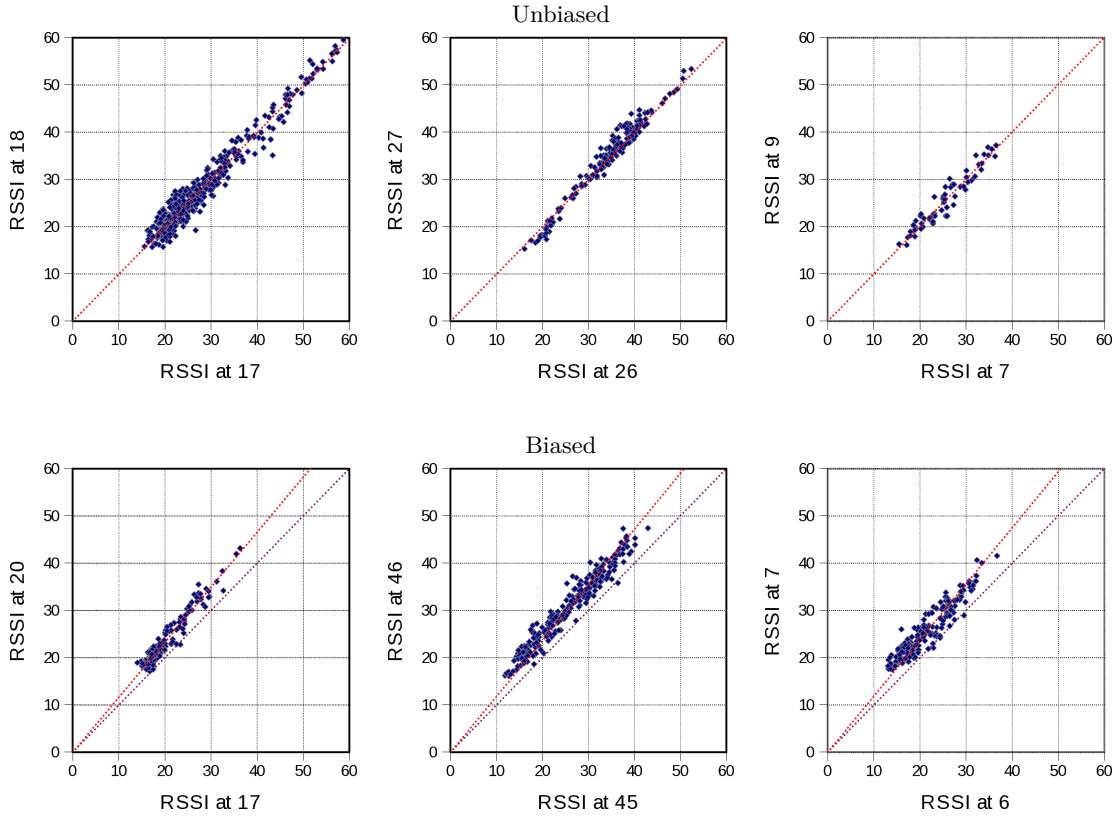


Figure 2.3: RSSI in reciprocal communication. The three plots a the top depict nearly symmetrical readings while the three plots at the bottom depict asymmetrical readings. The dotted red lines illustrate the bias (inclination) of the reciprocal readings and the dotted purple lines illustrate the identity line (in the biased plots).

and, thus, symmetric, i.e., $x_{AB}(t) = x_{BA}(t)$), and s_A is the scaling bias of badge A. This model states that the RSSI values read by a badge are biased by a multiplicative factor. The model is inspired by the data (see Fig. 2.3).

If the model is approximately correct, then we can estimate s_A/s_B by computing the ratio $\alpha_{AB} = \left\langle x^A_{AB}(t) \right\rangle / \left\langle x^B_{AB}(t) \right\rangle$ ($\langle\circ\rangle$ denotes *average*) and we can estimate s_A/s_C indirectly as

$$\frac{s_A}{s_C} = \frac{s_A}{s_B}\frac{s_B}{s_C} \Rightarrow \alpha_{AC} \approx \alpha_{AB}\alpha_{BC},$$

an estimation which can be improved by taking the average over all intermediate badges

$$\tilde{\alpha}_{AC} = \frac{1}{n}\sum_B \alpha_{AB}\alpha_{BC},$$

where n is the number of intermediate badges.

Estimating the ratios α_{AB} through $\tilde{\alpha}_{AB}$ allows us to estimate the ratios that we cannot compute directly from the data. Our calibration procedure requires knowing all the ratios α_{AB}. Thus, the quality of the estimation $\tilde{\alpha}_{AB}$ is relevant, and we assessed it through experimentation.

The following calibration algorithm is an immediate consequence of this empirical property. First, we compute all ratios $\alpha_{AB} = \left\langle x^A_{AB}(t) \right\rangle / \left\langle x^B_{AB}(t) \right\rangle$. Then, for all pairs A, B such that α_{AB} is missing, we estimate α_{AB} through $\tilde{\alpha}_{AB}$. Once every α_{AB} is computed, we proceed to compute constants $\alpha_A = \frac{1}{n} \sum_B \alpha_{AB}$, n being the number of badges. Then, replace the RSSI values $x^A_{AB}(t)$ by $x^A_{AB}(t)/\alpha_A$. All the RSSI will be in the same scale.

The rationale behind the above is very simple. If $\alpha_{AB} = s_A/s_B$, then $\alpha_A = \frac{1}{n} s_A \sum_B s_B^{-1}$. Thus, $\alpha_A = \alpha s_A$, $\alpha_B = \alpha s_B$, etc., for $\alpha = \frac{1}{n} \sum_A s_A^{-1}$. Then, we have that $x^A_{AB}(t)/s_A = x_{AB}(t)/\alpha$. Since all readings become proportional to the *ideal* readings through the same scaling factor α^{-1}, all the readings are in the same scale.

2.3.3. RSSI smoothing

(Adapted from [82].) The objective of applying filters to proximity-related RSSI values is two-fold: to produce continuous estimates of RSSI values and to reduce their variability.

Signal strength attenuates with both distance and obstacles such as the human body. We would like to reduce the effect of obstacles as much as possible. So far, we can only identify when communication was blocked when a very low RSSI value or missing reading was preceded and followed by large RSSI. For example, $RSSI = 40$ (-60 dBm) surrounding a missing reading represent walking 40 to 60 meters in 16 seconds, unless communication was blocked. In general, we would like to ignore quick reductions in signal strength. At the same time, high RSSI values are more reliable than the rest, because they can only occur in close proximity.

Our approach consists in estimating $x_{AB}(t)$ through a weighted moving average of RSSI observations. We consider observations within a time window of 60 seconds around time t, i.e. in the interval $[t - 30, t + 30]$. We prioritize observations according to their proximity

in time to t and according to their *relative magnitude* with respect to the surrounding observations.

Let x_τ be an observation (RSSI) taken at time τ and let $S(t) = \{x_\tau : t-30 \leq \tau \leq t+30\}$ be the set of observations associated to the time window $[t-30, t+30]$. Now, let us define the *temporal weights* $\omega_T(\Delta t)$ as:

$$\omega_T(\Delta t) = \frac{1}{a + \Delta t^2},$$

where a is a tuning constant. From this definition, we proceed to define the moving average filter $\mathcal{A}(t)$ as:

$$\mathcal{A}(t) = \frac{\sum_{x_\tau \in S(t)} x_\tau \omega_T(t-\tau)}{\sum_{x_\tau \in S(t)} \omega_T(t-\tau)}.$$

Filter $\mathcal{A}(t)$ defines the *local magnitude* of the set of observations $S(t)$.

The temporal weights $\omega_T(\Delta t)$ configures how much an observation weights when it is Δt time away from time t. We chose a so that an observation 15 seconds away from t weights 50% of an observation at time t; thus, we chose $a = 15^2$.

Having defined the local magnitude, we define the relative magnitude of x_τ as the difference between x_τ and $\mathcal{A}(t)$: $\Delta x_\tau(t) = x_\tau - \mathcal{A}(t)$. This definition depends on both τ and t, so an observation x_τ has a diferent relative magnitude according to the different time window in which it is used.

Now, let us give priority to observations according to their difference with respect to the local magnitude. Let us define the *magnitude weight* $\omega_M(x, t)$ as:

$$\omega_M(x, t) = b + \max(0, x - \mathcal{A}(t)).$$

An observation below the local magnitude $\mathcal{A}(t)$ will receive a minimum weight b. We have chosen $b = 4^2$ because filter $A(t)$ is very local, and any deviation from it appears significative.

We now introduce the full formula of our *priority* filter $\mathcal{P}(t)$. For this filter, we use both the temporal and magnitude weights, and replace observations below the local magnitude

$A(t)$ with the local magnitude. We define $\mathcal{P}(t)$ as:

$$\mathcal{P}(t) = \frac{\sum_{x_\tau \in S(t)} \max(x_\tau, \mathcal{A}(t))\omega_T(t-\tau)\omega_M(x_\tau, t)}{\sum_{x_\tau \in S(t)} \omega_T(t-\tau)\omega_M(x_\tau, t)}.$$

Note that, if constant b of ω_M is adjusted accordingly, this filter is commutative with increasing linear transformation applications on x_τ. Thus, our calibration methods work both before or after application of this filter.

This filter has a strong weakness: it fails to provide reasonable estimation in the presence of strong packet loss. This can be fixed, however, by introducing *false zero observations* to set $S(t)$. Using the modified $S'(t) = S(t) \cup \{\hat{x}_{t-30} = 10, \hat{x}_{t+30} = 10\}$ instead of $S(t)$ should suffice. This way, the false observations will have an effect when there are few readings in the interval $[t-30, t+30]$.

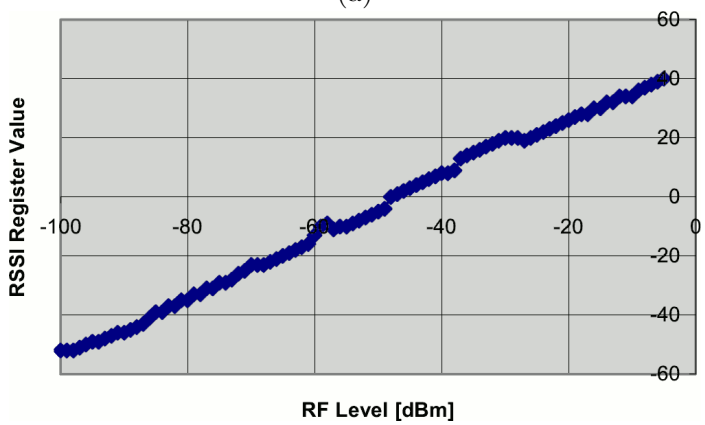
2.3.4. Detection of hand washing

(Adapted from [82].) Alcohol dispensers sent three signals in a row whenever someone pressed their pump. Also, it was normal for several healthcare workers to be around a dispenser when an activation occurred. All this led to redundant replication of dispenser signals in the data.

A simple solution to this problem consisted in matching received signals to dispensers, and considering all signals occurring in succession to come from the same activation. Given the slight differences in clock synchronization among the badges, a window of 10 seconds was chosen. The *activation event* was then assigned to the badge that read the highest RSSI.

Still, some activations were associated to very low RSSI values. These activation very are unlikely to have been caused by healthcare workers carrying badges. We had to choose a threshold R_{HMIN} below which we would ignore these signals. The fact that there were activations caused by people not participating in the experiment meant that we would have to deal with false activations that we could not get rid of. Fortunately, the distribution of detected dispenser events by R_{HMIN} suffered a distortion in the range $[36, 40]$ (it is a bell that grew a second, small mode), so our *a priori* choice is $R_{HMIN} = 38$.

(a)



(b)

PA_LEVEL	TXCTRL register	Output Power [dBm]	Current Consumption [mA]
31	0xA0FF	0	17.4
27	0xA0FB	-1	16.5
23	0xA0F7	-3	15.2
19	0xA0F3	-5	13.9
15	0xA0EF	-7	12.5
11	0xA0EB	-10	11.2
7	0xA0E7	-15	9.9
3	0xA0E3	-25	8.5

Figure 2.4: Key insets from the technical specifications of the CC2420 radio. Inset (a) shows that RSSIs are internally shifted from signal strength (dBm) using an additive constant that varies by device (but is around 45), and inset (b) shows the output power of the antenna for different configurations.

2.3.5. The RSSI-contact threshold

(This subsection contains unpublished work, but it informed [83].) I wanted to detect whether two healthcare workers were in close proximity to each other, but there was no reference to RSSI and distance in the experiments. So, I went through some published experiments to determine a threshold: $RSSI = 35$ roughly means a distance of 4 meters.

The Telos-B mote uses a CC2420 radio [100]. The CC2420's technical specifications state that the RSSI values received from the antenna are the signal strength plus an additive constant (Fig. 2.4-(a)) and that the output power is customizable (Fig. 2.4-(b)) [3].

The TinyOS code written for the motes sets the RSSI offset to -30 (signal strength was subtracted 30). It also states that the saved RSSI values were added 60 to make them

positive and that the power level was 15 (output of -7 dBm). Combined, the first two facts state that our RSSI can be transformed to dBm just by subtracting 100. The third fact states that, after transforming our RSSI values to dBm, we still need to add 7 dBm to them to make them comparable to the figures in literature (as they are normally reported with an output power of 0 dBm).

It has been reported that at a separation of 4 meters (13 feet) between the antennas, signal strengths range between -50 to -55 dBm [100] and between -47 to -61 dBm [9] for an output power of 0 dBm. The intervals are $[38, 43]$ and $[32, 46]$ in our scale. In light of this, a value of 35 dBm seems relaxed enough to pick most interactions at around 4 meters.

2.3.6. The RSSI-room threshold

(This is unpublished work.) Before the sensor network measured the interactions in the MICU, a researcher (Phil) walked around the unit carrying three badges, entering into each room with a bedmote for a few (2 to 5) minutes. We could then use this data to find the RSSI value at which we can determine when a worker was inside a room.

I used the following procedure to determine the threshold at which a worker is *definitely* inside a room. To ease the presentation, let us suppose that we are working only with one badge and one bedmote. My idea was to determine the interval of time the badge was inside the room by testing when $RSSI \geq R_{IN}$ (a threshold). If the duration of the interval is consistent with the *few minutes* reference, then the threshold R_{IN} determines when a badge is inside a room.

Given the variability of the readings, I am interested in the first and last time $RSSI \geq R_{IN}$ to define an interval. Let us call r_t to an RSSI value obtained from a signal received at time t. Then, let us define functions

$$left(x) = \min\{t : r_t \geq x\}$$

and

$$right(x) = \max\{t : r_t \geq x\}.$$

These functions characterize the interval obtained using thresholds. Let us define the *duration* function then:

$$duration(x) = \begin{cases} 0, & \max\{r_t\} < x \\ right(x) - left(x) + 8, & \sim \end{cases}$$

This function returns the estimated duration of the interval when using the $RSSI \geq R_{IN}$ method. The +8 term is added because the sample rate is 8 seconds; if a value greater or equal to, say, 60 was received only once, then $left(x) = right(x)$ would make $duration(x) = 0$.

To measure how likely is such a threshold to accept RSSI when the badge is inside a room, let us define the following functions:

$$histogram(x) = |\{r_t : r_t = x\}|,$$

which counts the frequency of RSSI values received, stratified by value, and

$$tail(x) = \sum_{y \geq x} histogram(x),$$

which counts how many RSSI were received above a value (*tail distribution*).

Fig. 2.5-(a) shows the *duration* function (*interval bound*) plotted against the *tail* and *histogram* curves, which were multiplied by 8 to make them comparable, for bedmote 185. For this plot, $R_{IN} = 74..52$ reproduce the time inside the room. Ranges were also frequent with other bedmotes, but they were slightly different. Values around 50 and 51 consistently discovered the proper duration. By comparing the curves, we note that $RSSI \geq R_{IN}$ only 1/4 to 1/5 of the time. This was also consistent with other bedmotes.

As stated previously, the RSSI from the bedmotes were comparable to each other. Fig. 2.5-(b) shows an experiment proving this; for each pair of bedmotes, I took all the readings received from both of them in intervals of 8 seconds (e.g., $(30, 45)$), left the smallest RSSI

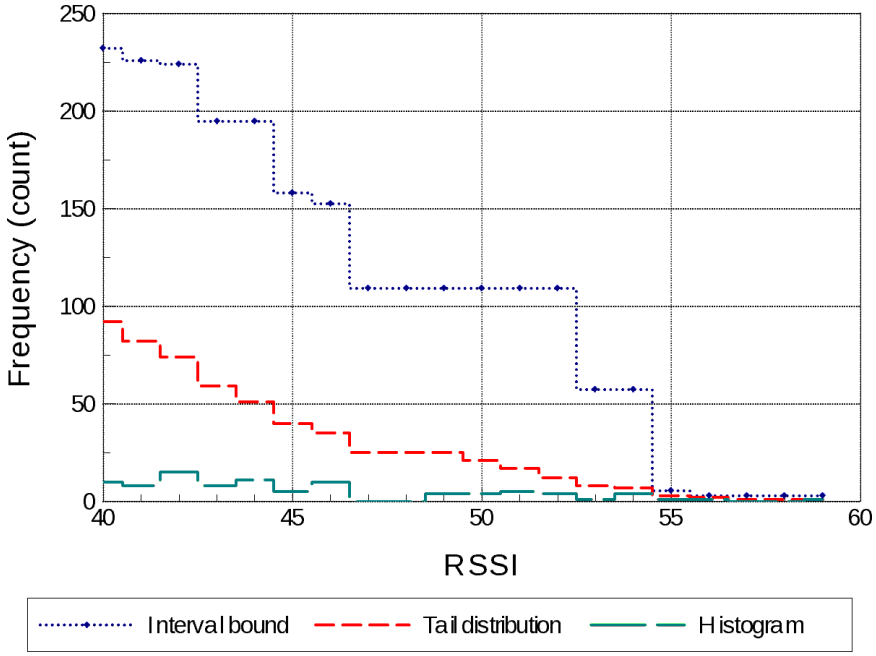


Figure 2.5: RSSI and room thresholds. The plot compares the interval duration with the number of RSSI received above the threshold (tail distribution and histogram curves).

(e.g., $(30, 45) \rightarrow 30$), and took the maximum of all those readings. For that pair of bedmotes, a value of 30 means that there was a place both bedmotes sent RSSI of 30 or more at the same time. Finally, I removed all the pairs of bedmotes with values smaller than 45. This left the graph shown in Fig. 2.5-(b), which shows the adjacent bedrooms in MICU, granted their bedmotes were working. This shows that their RSSI are comparable and that 45 is too small to guarantee that a badge is inside a room.

2.3.7. Detection of room visits

(Adapted from [82].) We can guarantee that a badge is inside a room when $RSSI \geq R_{IN}$. However, a badge can still be inside if this condition is not met, because communication can be easily blocked. Therefore, we needed a robust method for identifying a visit to a room.

The solution I proposed uses three RSSI thresholds: one to determine a visit with certainty (R_{IN}), another to determine when a worker was outside the room (R_{OUT}), and the last one to determine when a worker was probably entering the room (R_{DOOR}). There

is also a threshold for the minimum amount of time (t_{MIN}) for a *candidate visit* so that it can be considered as a real one.

The detection of a visit ocurred like this. Let us suppose that we are dealing with a single badge and a single room, first. Let us suppose the badge is classified as *out-of-room*. If $RSSI \geq R_{DOOR}$, then we badge *may* have entered the room (or be at the door), so we remember the time this condition was met and transition to *maybe-room* state. Next, if $RSSI < R_{DOOR}$, then the state becomes *out-of-room* again. However, if eventually $RSSI \geq R_{IN}$, then we say that the badge IS inside the room, so it enters the *in-room* state.

A visit ends when $RSSI < R_{OUT}$. But the time set for when the visit ends is analogous to the time the visit starts: it is the last time $RSSI \geq R_{DOOR}$ was met. Naturally, we also added the condition that a badge can only be inside the closest room. If the RSSI of another bedmote was higher, then the visit had to end.

The thresholds chosen, in principle, were $R_{OUT} = 43$, $R_{DOOR} = 49$, $R_{IN} = 52.5$, $t_{MIN} = 16$. However, we changed these values later, as the next subsection explains.

2.3.8. Hand hygiene adherence

(Adapted from [83].) We define a *hand-hygiene opportunity* as the event corresponding to a healthcare worker entering or leaving a patient room. We associated an opportunity to an alcohol-dispenser activation if the two events occurred within 30 seconds of each other and were associated with the same badge (i.e., the same HCW). This informs our measure of *observed adherence*, central to our work, which we define as the fraction of opportunities associated with alcohol-dispenser activations, where both the opportunity and the activation correspond to the same individual.

2.3.9. Consistency maximization

(This work was briefly mentioned in [83] without any reference to the details presented here.) One can still have doubts regarding whether the visits are properly detected. Using different RSSI thresholds can easily lead to the splitting and destruction of the visits detected, as well as changes in their duration. In fact, initially, there was a great mismatch between opportunities and dispenser usage, leaving many activations unrelated to

opportunities, and most opportunities were left unsatisfied.

These discrepancies lead to a multiobjective optimization problem. The solution space consists in the parameters that serve to identify *events*: the parameters for visits (R_{OUT}, R_{DOOR}, R_{IN}, t_{MIN}) and dispenser activations (R_{HMIN}). After detecting events in the data, we get three numbers: n_{OPP} for opportunities, n_{HH} for activations, and n_{MATCH} for matched opportunities (similar to the number of matched activations). We are interested in maximizing the fractions $f_{OPP} = n_{MATCH}/n_{OPP}$ and $f_{HH} = n_{MATCH}/n_{HH}$. Then, a solution σ is *worse* (*Paretto inferior*) than a solution σ' iff $f_{OPP}(\sigma) \leq f_{OPP}(\sigma')$ and $f_{HH}(\sigma) \leq f_{HH}(\sigma')$, and at least one of the inequalities is strict. The set of all the solutions that are not *worse* than any other (*Paretto front*) is a 2-D curve in the space of f_{OPP} and f_{HH}. Of this curve, I picked the solution that had R_{HMIN} and R_{IN} closest to the values from *ground truth*.

If evaluating a single combination of parameters takes a couple of hours, then trying each parameter combination would take an enormous time. So, I proceeded like this. First, I computed the set of all alcohol dispenser activation events, associating them to the *winning* RSSI. Getting the activations with $RSSI \geq R_{HMIN}$ then reduced to parsing this set. Second, I reduced the readings from the bedmotes to a list $\{..., (t, RSSI_t, bedmote_t), (t+1, RSSI_{t+1}, bedmote_{t+1}), ...\}$ per badge, where $RSSI_t$ was the highest RSSI read from a bedmote at time t, and $bedmote_t$ is that bedmote. Then, for each value of R_{DOOR}, I parsed the previous list creating a list of intervals. The interval where $RSSI_t \geq R_{DOOR}$ and $bedmote_t$ remained the same was summarized as $(t_0, t_f, \max\{RSSI_{t\in[t_0,t_f]}\}, bedmote)$. Conversely, the interval where $RSSI_t < R_{DOOR}$ and $bedmote_t$ remained the same was summarized as $(t_0, t_f, \min\{RSSI_{t\in[t_0,t_f]}\}, bedmote)$. Then, for each value of R_{MIN}, a new list was created, merging all the adjacent intervals with $RSSI \geq R_{MIN}$ and associated to the same bedmote; the resulting intervals are associated to the maximum RSSI. Finally, for each value of R_{IN}, a new list ignoring the intervals with $RSSI < R_{IN}$ is created. Then it came the filtering according to t_{MIN}, and matching the activations to the resulting opportunities.

The parameters explored satisfied $40 \leq R_{OUT} \leq R_{DOOR} \leq R_{IN} \leq 52$, $8 \leq t_{MIN} \leq 24$, and $30 \leq R_{HMIN} \leq 45$. I also ranged the parameters for the RSSI interpolator $\mathcal{P}$ independently. A very smooth filter privileges very high R_{OUT} while a very weak filter privileges low R_{OUT}. In the end, the best solution came from a very weak filter ($b = 2^2$).

Two main solutions can be derived from the data. One comes from using the whole dataset. The second one comes from using only the *night shift* data, which is arguably the best choice. During day shifts, you have dispenser activations explained by people not wearing badges: relatives of the patients, religious figures, physicians from other units, etc. In contrast, in-unit physicians almost exclusively used the alcohol dispensers during the night.

2.3.10. Sanity checks

I cannot say that the processing of the data has been properly validated. We do not have a reference dataset with tagged events (dispenser activations, contacts, and room entries and exits) to use as a gold standard for validation. Therefore, we resorted to a few, simple *sanity checks*, to see that our processing *made sense*.

The first sanity check, published in [82], consisted in inferring the rooms that did not have patients. We have ground truth for this: at the beginning and end of each shift, a researcher walked around the MICU annotating the rooms that hosted patients. This allows for informed guesses: that rooms that were occupied at the beginning and end of the shift were probably continuously occupied, that rooms that were empty at the beginning and end of the shift were probably empty throughout the shift, and that rooms that started empty but ended occupied were occupied at some time during the shift. We call these categories *occupied, c.empty*, and *nc. empty*, respectively. We plotted the frequency these rooms were visited throughout the day in Fig. 2.6-(a). As expected, *c.empty* rooms received scant visits, *nc. empty* rooms received more visits as the day progressed, and *occupied* rooms received visits regardless of the time of the day.

The second sanity check consisted in evaluating whether the RSSI from the bedmotes related to distance similarly. I tested this by plotting a room-adjacency graph shown in

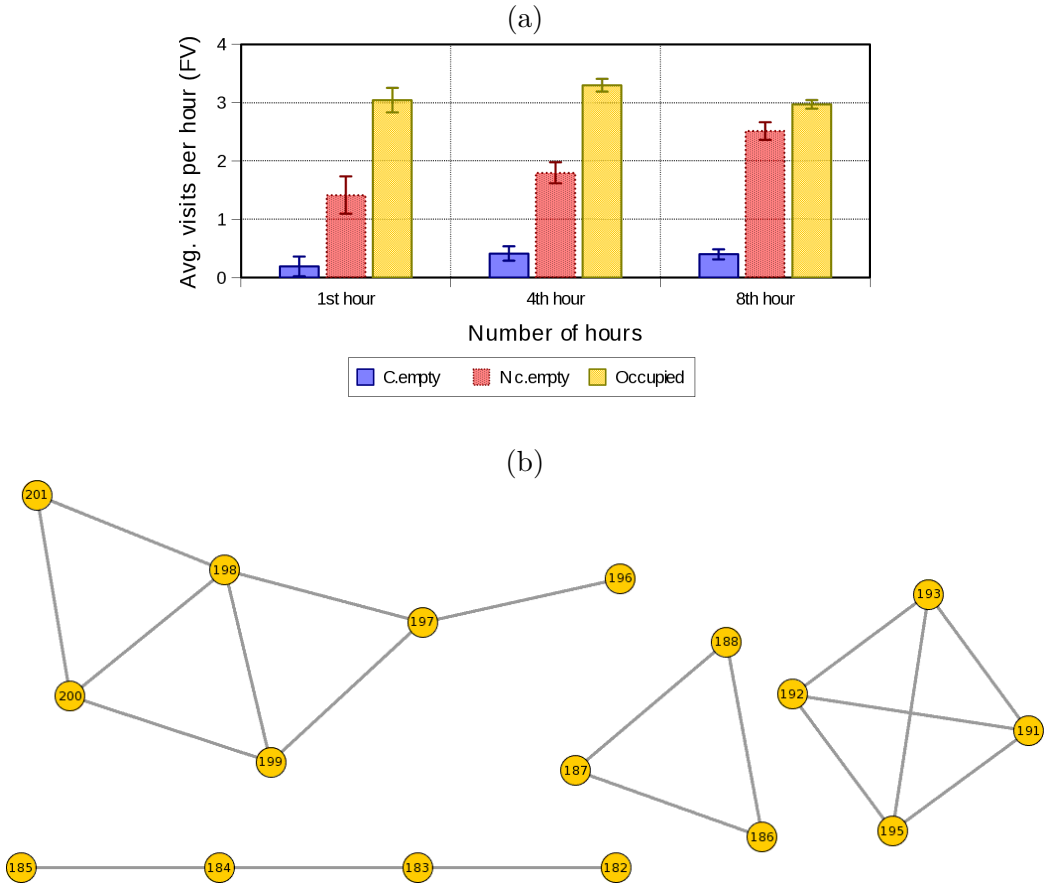


Figure 2.6: Plots depicting the sanity checks: (a) frequency of visits according to room occupancy status, and (b) graph of adjacent bedmotes. Both are in accord to expectations.

Fig. 2.6-(b), which matches which rooms had doors at close distance, for the bedmotes that were working during the walkthrough to get $RSSI_{IN}$ (a few bedmotes were off). For each pair of bedmotes A, B, I defined their *RSSI-proximity* r_{AB} as follows. Let $x_A(t)$ denote the RSSI received from bedmote A at time t. Then, r_{AB} is defined as

$$r_{AB} = \max\{r : (\exists t) r = x_A(t) = x_B(t)\},$$

i.e., the maximum RSSI to either bedmote over all equidistant points. Then, I created the adjacency graph by taking the bedmotes as vertices and the edges as the pairs of bedmotes A, B with $r_{AB} \geq 45$.

2.4. Predicting interactions from room visits

One of the questions we wanted to answer was whether knowing when workers visit rooms can be enough to predict their interactions (i.e., close proximity contacts). In a sense, this work was meant to validate previous work done in the compepi group in which contacts between healthcare workers were estimated by the spatial and temporal proximity of their respective computer logins, creating *login networks* [18, 17, 16]. We confirmed this hypothesis by being effectively able to predict contacts from visits to rooms. This work resulted in a publication in SocialCom 2012 [81].

2.4.1. Link prediction methodology

The purpose of the experiment was to predict the interactions between healthcare workers by knowing when they visited patient rooms. To do so, we used the following methodology:

1. For each shift, create a labeled graph G_L of healthcare workers, where each vertex has a label (badge and job type) and each edge has a weight determined as follows: for a pair of workers, identify their room entries, and count the entries followed by an entry of another worker within t time and d architectural *hops* (the minimum number of room-changes necessary to go from one room to another).

2. For each shift, create a labelled graph G_C of healthcare workers, where each vertex has a label (badge and job type) and each edge has a weight telling the amount of time the workers were in contact.

3. Train classifiers that predict G_C by using G_L. Each classifier f is trained using only one *link prediction score* at a time (see Table 2.1). Training is done using day and night shifts independently. Since the vertices are unique but identifiable if G_L and G_C belong to the same shift, f predicts edge $(u, v) \in G_C$ if $(u, v) \in G_L$, and using weight w_{uv} as well. Every possible pair of job types used a different classifier f.

4. Testing follows a cross validation scheme naturally suggested by the deployment: use

Name	*Acronym*	*Formula*
Common neighbors	CN	$CN_\alpha(u,v) = \sum_{z\in\Gamma(u)\cap\Gamma(v)} (w_L(u,z)^\alpha + w_L(z,v)^\alpha)$
Common neighbors extended	CN^+	Same as CN, but using Γ^+ instead of Γ
Adamic-Adar	AA	$AA_\alpha(u,v) = \sum_{z\in\Gamma(u)\cap\Gamma(v)} \frac{w_L(u,z)^\alpha + w_L(v,z)^\alpha}{\log(1+s_\alpha(z))}$
Adamic-Adar extended	AA^+	Same as AA, but using Γ^+ instead of Γ
Jaccard's index	JI	$JI_\alpha(u,v) = \frac{CN_\alpha(u,v)}{s_\alpha(u)+s_\alpha(v)}$
Jaccard's index extended	JI^+	Same as JI, but using Γ^+ instead of Γ
Login hypothesis	LH	$LH(u,v) = \begin{cases} 1, & \{u,v\} \in E_L \\ 0, & \sim \end{cases}$
Login hypothesis improved	$LH^\star$	$LH^\star(u,v) = w_L(u,v)$

Table 2.1: Similarity scores used to compare the topological similarity of two vertices. The formulas are defined for graph $G_L = (V, E_L)$, and vertices $u, v \in V$. The neighborhood function Γ is defined as $\Gamma(v) = \{u \in V : \{u,v\} \in E_L\}$, and $\Gamma^+(v) = \Gamma(v) \cup \{v\}$. Weights w_L are such that $w_L(u,v) = 0$ when $\{u,v\} \notin E$.

one shift as a hold-out and train using the rest. This led to 10-fold cross validation during the day and 9-fold cross validation during the night.

The *link prediction scores* are graph theoretical scores that assess the similarity of two vertices by using their shared neighborhood [73]. For example, the *common neighbors* score counts the number of common neighbors between two vertices. In weighted graphs, it is possible to sum the weights of the edges towards the common neighbors instead of counting them. An approach that interpolates between both possibilities is to apply an α-power on the weights, to account for the relative importance of *weak ties* [72]. If $\alpha = 0$, then the score is equivalent to counting the common neighbors. If $\alpha = 1$, then the score is equivalent to summing the weights. The scores used in this work (Table 2.1) include weights and the weak ties α-power in their definitions.

A couple of special scores are the *login hypothesis* and *login hypothesis improved*, which are intended to prove the adequacy of the *login networks* as proxy of actual contact networks. The login networks were generated much in the style of G_L, except that using computer logins instead of room entries.

To simplify classification, we applied thresholds to the edges of the G_C graphs, hence removing their weights. An edge $(u,v) \in G_C$ was left if $w_{u,v} \geq \theta$, where $\theta \in \{0.01, 0.02, 0.03\}$ (meaning being in contact 1%, 2% and 3% of the time). Such seemingly small thresholds

were chosen so that the graphs generated were not sparse.

The classifiers f used the scores and only allowed two types of classification by threshold: *homophily* and *heterophily*. In homophily, an edge is predicted if the similarity between the vertices exceeds a threshold (attraction by similarity). Heterophily works conversely.

2.4.2. Experiment results

Table 2.2 shows the results of the classifiers after cross validation. The Common Neighbors (CN) and Adamic Adar (AA) scores were the ones which gave consistently better performance. Otherwise, most classifiers performed nearly as well. Classification was heavily dominated by homophily. Heterophily was mainly present in Jaccard's index scores JI and JI^+, normally in *criticalcare-doctor* relations, together with negative α. The *login hypothesis* scores performed similarly and just slightly worse than the rest, except when predicting day G_C, with interaction at least 1% of the time. Note that LH either predicts: $\{u, v\} \in E_C$, $\{u, v\} \notin E_C$, $\{u, v\} \in E_L \Leftrightarrow \{u, v\} \in E_C$, and $\{u, v\} \in E_L \Leftrightarrow \{u, v\} \notin E_C$, but the latter possibility never happened in practice. The results basically demonstrate that room entries positively relate to contacts. (We may have achieved better accuracy by combining the scores in a more complex classifier, but this would have made it difficult to provide direct evidence that *login networks* approximate actual contact networks.)

2.5. *Peer effects* in the MICU

I now explain the main findings associated with the objective of the experiment: to unveil the existence of *peer effects* in the dynamics of hand hygiene adherence. One of hypotheses behind the deployment was that a healthcare worker was more likely to comply with hand hygiene when others were around (if no one watched them, they could get away with not washing their hands) and, when workers were not alone, some were more influential in making others wash their hands, directly or indirectly.

The data verified these hypotheses: workers were more prone to adhere with hand hygiene when they were accompanied, and some workers were more influential than others, by job type. This work resulted in an article accepted and under revision in ICHE [83].

(a) Day G_C with $w_C \geq 0.01$

Score	Accuracy	G_L	t	d
AA	0.653		30	2
AA^+	0.653		30	2
CN	0.635		30	2
CN^+	0.642		10	2
JI	0.611		30	4
JI^+	0.611		30	4
LH	0.680		15	2
$LH^\star$	0.552		30	4

(b) Night G_C with $w_C \geq 0.01$

Score	Accuracy	G_L	t	d
AA	0.832		30	4
AA^+	0.831		15	4
CN	0.827		30	4
CN^+	0.829		30	4
JI	0.816		30	4
JI^+	0.815		30	4
LH	0.696		15	0
$LH^\star$	0.749		30	2

(c) Day G_C with $w_C \geq 0.02$

Score	Accuracy	G_L	t	d
AA	0.723		30	2
AA^+	0.733		30	2
CN	0.734		10	2
CN^+	0.734		10	2
JI	0.731		30	4
JI^+	0.731		30	4
LH	0.700		30	0
$LH^\star$	0.624		30	2

(d) Night G_C with $w_C \geq 0.02$

Score	Accuracy	G_L	t	d
AA	0.893		30	4
AA^+	0.888		30	4
CN	0.887		15	4
CN^+	0.888		30	4
JI	0.877		30	4
JI^+	0.877		30	4
LH	0.715		10	0
$LH^\star$	0.812		30	0

(e) Day G_C with $w_C \geq 0.03$

Score	Accuracy	G_L	t	d
AA	0.764		10	2
AA^+	0.792		10	2
CN	0.795		10	2
CN^+	0.795		10	2
JI	0.740		15	4
JI^+	0.783		30	4
LH	0.737		10	0
$LH^\star$	0.779		30	0

(f) Night G_C with $w_C \geq 0.03$

Score	Accuracy	G_L	t	d
AA	0.873		30	4
AA^+	0.869		30	4
CN	0.866		30	4
CN^+	0.871		30	4
JI	0.871		30	4
JI^+	0.874		30	4
LH	0.664		30	0
$LH^\star$	0.759		30	2

Table 2.2: Accuracy of link prediction scores in day shifts (insets (a), (c), (e)) and night shifts (insets (b), (d), (f)), obtained through cross validation.

2.5.1. Observed adherence

Our analysis is limited to *observed adherence*, which we define as the fraction of opportunities satisfied using alcohol dispensers outside rooms. Therefore, we ignored the alcohol dispensers placed inside the rooms. Figure 2.7 shows the map with the beacons considered in the analysis.

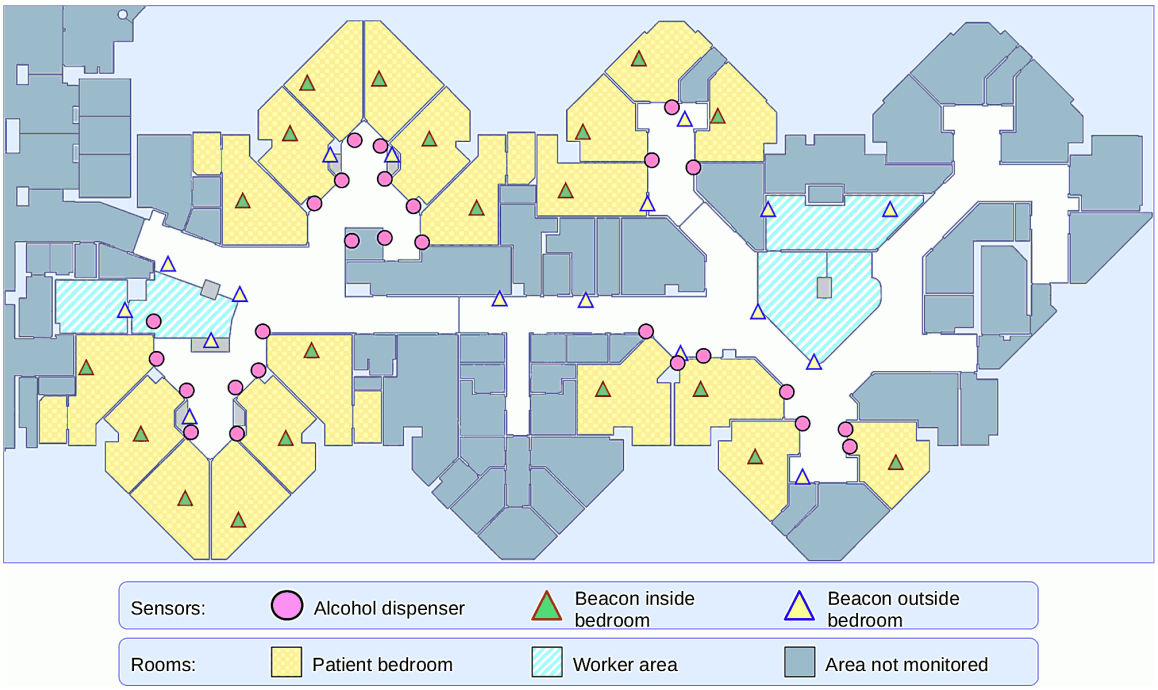


Figure 2.7: Placement of the sensors in the MICU, deployment of June 2011, relevant to the *peer effects* paper.

2.5.2. Measuring social proximity

We defined scores to characterize the social context of HCWs:

1. *Coworkers encountered within one minute (W1M).* This measure represents the number of different coworkers encountered within an interval of one minute centered on each hand-hygiene opportunity. An encounter is considered to occur within a distance of approximately 4 meters, distance at which a worker might be aware of coworkers. Measures alike have been used in other works [32, 120].

2. *Sum of RSSI (SRSSI).* The RSSI of a message serves as a measure of spatial proximity between the communicating radios. For a worker's badge, we can use the sum of the RSSIs (decibel-milliwatts) of messages received from other badges as a measure of the crowdedness of his/her social space at the time of each hand-hygiene opportunity. SRSSI increases as the worker gets closer to fellow coworkers, and it also increases when other badged coworkers enter the sensing range. This is a novel measure.

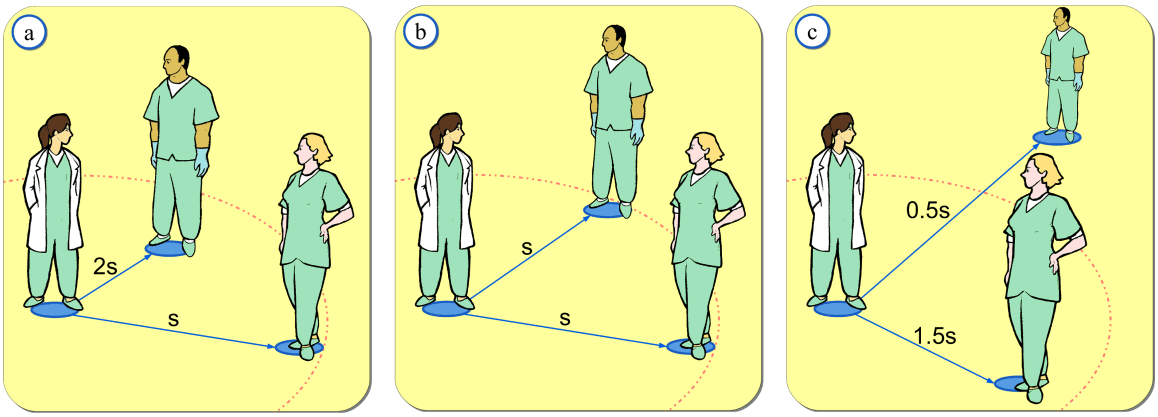


Figure 2.8: W1M and SRSSI compared in three different situations. Values of the variables in the insets: (a) $W1M = 2, SRSSI = 3s$, (b) $W1M = 2, SRSSI = 2s$, and (c) $W1M = 1, SRSSI = 2s$. W1M is the same only in insets (a) and (b), while SRSSI is the same only in insets (b) and (c). Besides differing in their nature (W1M is discrete and SRSSI is continuous), both variables also differ in the aspect of local crowding they measure.

While W1M provides a clear distinction between being *alone* ($W1M = 0$) and being *accompanied* ($W1M > 0$), SRSSI offers a smoother measure for the effect of the social context on adherence (Fig. 2.8).

2.5.3. Results

First, we found that when a worker was alone ($W1M = 0$), observed adherence was 20.85% (95%-confidence interval: $[19.78\%, 21.92\%]$). In contrast, we found that when other healthcare workers were present ($W1M > 0$), observed adherence was 27.90% (95%-CI: $[27.48\%, 28.33\%]$). This absolute increase of 7 percentage points is statistically significant, with an associated p-value P<0.001 (two-tailed t-test).

Second, we found that the observed adherence increased mainly with W1M, as shown in Fig. 2.9, with diminishing marginal returns. And when controlling for *confounding factors,* increases in W1M and SRSSI are independently associated to increases in *observed adherence.* The plots in Fig. 2.9 can be (are) biased by the presence of confounders that have an effect on adherence but are not associated to the social variables. For example, a patient with a contagious disease but demands little care is likely to receive more isolated healthcare workers who are very likely to wash their hands. Also, different workers may have different

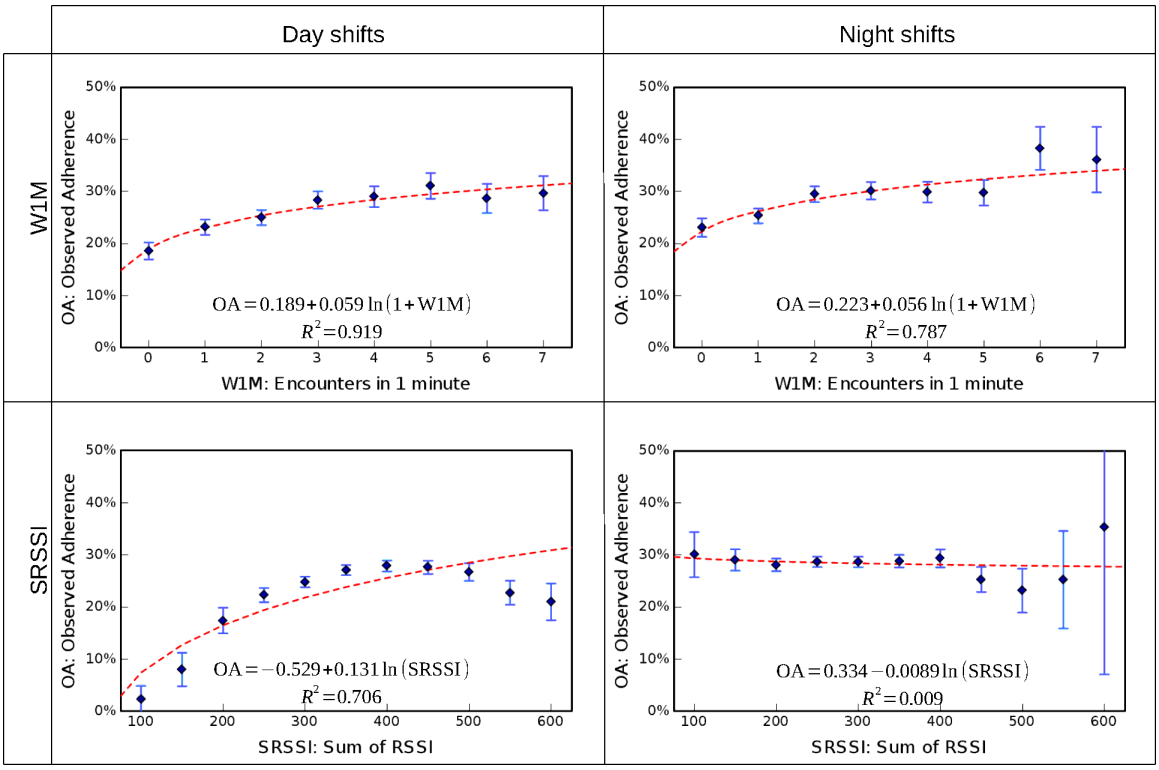


Figure 2.9: Associations between adherence and the social variables, for day and night shifts. Each diamond box represent an adherence rate, with its corresponding 95% confidence interval. Each red line represents the weighted linear regression model that associates adherence to the logarithm of the corresponding social variable. The width of the confidence intervals were used as weights. Adequacy of the regression model can be interpreted as diminishing marginal returns.

inclinations towards hand hygiene adherence. To remove their effects, I used *case control* and controlled for patients (represented by a shift number and the bedmote of the room) and workers (shift number and badge) independently (to create enough comparable groups), and found that increases in W1M and SRSSI were statistically significantly associated to increases in adherence (Wilcoxon's clustered-pairs signed rank test, $P \ll 0.001$).

And third, as Table 2.3 shows, not all job types seemed to exert the same effect on coworkers. Overall, if any worker was accompanied by a critical care worker, their adherence would increase. Doctors and nurses were equally influential, but nurses were more likely to adhere with hand hygiene. However, these observations are limited in that healthcare workers may attend patients by forming groups, and adherence may be tied to the specific

Worker	Coworker	Adherence	95%-CI radius	Opportunities
Any	None	20.85%	1.07%	5521
	Any	24.30%	0.89%	8880
	CCare	29.43%	3.82%	547
	Doctor	24.03%	3.43%	595
	Nurse	23.96%	0.95%	7738
CCare	None	18.07%	4.78%	249
	Any	22.89%	3.57%	533
	CCare	13.04%*	13.76%*	23*
	Doctor	31.25%*	16.06%*	32*
	Nurse	22.80%	3.76%	478
Doctor	None	12.38%	6.30%	105
	Any	20.13%	4.52%	303
	CCare	42.86%*	25.92%*	14*
	Doctor	15.22%*	10.38%*	46*
	Nurse	19.75%	5.01%	243
Nurse	None	21.15%	1.11%	5167
	Any	24.55%	0.94%	8044
	CCare	29.80%	3.97%	510
	Doctor	24.37%	3.70%	517
	Nurse	24.18%	1.00%	7017

**Infrequent pairs of job types led to unreliable rates and wide 95%-CIs.*

Table 2.3: Observed adherence by job type, when the workers were alone ($W1M = 0$) or accompanied by another one ($W1M = 1$).

function each individual served in the group.

CHAPTER 3
CLOSTRIDIUM DIFFICILE RISK MEASUREMENT

3.1. Introduction

Clostridium difficile (c.diff) is a contagious health-care associated infection (HAI) that is resistant to antibiotics and is the main cause of antibiotic-associated diarrhea (AAD). It is a significant concern in health-care provision world-wide, since it affects delicate patients and increases hospital occupancy. In the United States alone, CDI affects an increasing number of patients, from 139,000 to 336,600 in 2000 to 2009, with an increasing mortality that disproportionately rose from 3,000 deaths/year (1999-2000) to 14,000 deaths/year (2006-2007) [11].

C.diff is an antibiotic resistant bacteria that naturally lives in the human intestines of many individuals [11, 2]. It is normally constrained by the intestinal bacterial flora. In the situation of antibiotic exposure, the bacteria that constrain c.diff disappear, allowing it to grow without constraints and causing harm because of the toxins it generates, situation known as Clostridium difficile infection (CDI). The diarrhea caused by the infection allows living c.diff cells and its spores to be expelled to the environment. While c.diff survives in the environment for a few hours, the spores can last for several months. The spores can be carried around in the health-care workers' garments and can reach other patients, who can become colonized through the oral route; the living bacteria can survive in a stomach with reduced acidity, and the spores can tolerate the normal stomach acidity. This colonized patient, if exposed to antibiotics, can then develop CDI, and the process repeats. This is how the spread of c.diff normally occurs in health-care settings.

But recent results have put into question the scientific knowledge with respect to c.diff [31, 39, 121, 27, 28, 99, 54, 103, 63, 41, 119]. Methodological changes in the laboratory test of c.diff have demonstrated that the diagnosis of CDI is and has been considerably inaccurate [98, 39, 97, 99, 54]. As already stated, rates of CDI cases increased substantially (~2.5 times) and became considerably deadlier (~5 times) from 2000 to 2009, but these cases

were diagnosed using the enzyme immunoassay (EIA) test, which is not very specific (misses cases of CDI, i.e., true positives) but is insensitive (very low false negative ratio). The introduction of the polymerase chain-reaction (PCR) test, which has increased specificity (high true positive ratio) but is more sensitive than EIA (introduces more false positives), doubled the number of diagnoses of CDI, introducing concerns in the medical community [31, 39, 97, 99, 54]. This disagreement has a serious implication: that the diagnosis of CDI is or was inaccurate (likely both) [97, 53, 113, 99, 54], and that the statistics and conclusions presented in the literature are not reliable [91]. Furthermore, recent experiments have questioned the belief that c.diff is highly contagious [121, 27]. In these works, researchers examined the genome sequences of the c.diff bacteria collected from diagnosed CDI cases, encountering abundant strains and finding that most cases (~2/3) of CDI were unrelated. In spite of the fact that CDI cases were diagnosed using EIA, these works give relevance to the role of community-acquired c.diff, i.e., due to contagion that occurred outside hospitals or clinics, and *asymptomatic carriers*, i.e., those patients who carry c.diff but do not express symptoms [27, 28]. Asymptomatic carriers are known to represent a risk to other patients in long-term care facilities [103, 63], however, if they are more prevalent than previously believed, they can pose a considerable risk to other patients in hospitals by transmitting the disease to them.

And while the accuracy of the knowledge about c.diff has been put into question recently, analytical research with respect to CDI has been historically underdeveloped [92, 91, 119]. In fact, c.diff transmission has seldom been modeled, in spite of the need to assist health-care provision and policy design [91, 119]. Similarly, the problem of accurately predicting whether a patient will develop CDI has received little attention; the most elaborate models have been developed by Dubberke et al [25] and Wiens et al [124, 125], which contrast with the simplicity of the dominant *scorecard* methods that consider only a limited number of variables [33, 1]. In particular, Wiens et al have been the only ones to study dynamically changing risk scores [124, 125].

In this chapter, I review some relevant, computational research on CDI and describe a

data base built using anonymized electronic medical records data (without including clinical notes, however) from the hospital that will be used for research on CDI in the next chapter.

3.2. Relevant research on CDI

There is considerable research around c.diff. As of 5 May 2014, searching for "clostridium difficile" in Google Scholar returns *about* 93,900 publications (preserving quotations and excluding patents and citations), with *about* 9,320 of them dated after 2013. Most of these works narrate medical and laboratory work, and also many are about guidelines and implementation practices. Still, several are of computational interest or relevance. I classify them as:

- *Risk measurement.* These works are concerned with developing measures that better identify patients at risk. Computationally, this is a *soft* classification[1] task.
- *Risk factors.* These works are concerned with better identifying which characteristics define patients at risk.
- *Modeling.* These works are concerned with dynamical system models of c.diff.
- *Detection.* These works are concerned with improving the identification of patients who later develop CDI.

In what follows, I briefly summarize some recent work falling in the categories mentioned above. Any research ideas, and their feasibility, are discussed in the following sections.

3.2.1. Risk measurement

Several risk measurement scores to identify patients at risk have been developed, with methods with ranging levels of sophistication, applicability and generalizability. The simpler methods are easily accessible to a healthcare worker without the need to use a computer; just paper and pencil are enough. Also, they can be easily shared across institutions and can be included in guidelines. More sophisticated methods require the use of software, and

[1] In hard classification, an instance is classified into a single class. In soft classification, the instance is given a score for each class, stating that .

Variable name	*Type*
Age	integer
CDI colonization pressure[1]	decimal
Modified Acute Physiological Score[2]	decimal
Days on high risk antibiotics[3]	integer
Admitted once in the last 60 days	binary
Admitted twice or more in the last 60 days	binary
Admission to ICU	binary
Received laxatives	binary
Received gastric acid suppressor	binary
Received antimotility drug	binary
Albumin less than 3.5 g/L on admission	binary

[1] Average number of CDI patients in the units visited [24].

[2] Physical wellness score unavailable to us.

[3] Cephalosporins, clindamycin, and amoxicillin.

Table 3.4: Independent variables of the logistic regression model by Dubberke et al.

the more data they use, the more customized they become to an institution, and sharing them becomes more difficult. But they can do better in terms of predicting which patients are going to develop CDI.

Let us classify these works into three categories: logistic regression (Dubberke et al [25]), dynamic risk metrics (Wiens et al [124, 125]), and simple, rule-of-thumb metrics [33, 1]. In terms of accuracy, the former two represent the state of the art. The approach of Dubberke et al [25] consists of a logistic regression model whose features were chosen through feature selection (PCA, DCA, cluster analysis, forward selection) and then were non-linearly transformed (polynomials, max/min). The logistic regression model used 14 features coming from the variables listed in Table 3.4. Their model had an AUC statistic close to 0.88, which is considered *good* and close to *excellent* [90].

Wiens, Guttag and Horvitz published a couple of papers on improving the risk measurement of CDI, first by creating an evolving "risk process" (a time series of risk) [124], and then by reducing this series to a static metric and comparing its performance to the state of the art [125]. For their first paper, they used the internal structure of the static SVM classifier of patients (that classifies them into developing CDI or not), and then used the

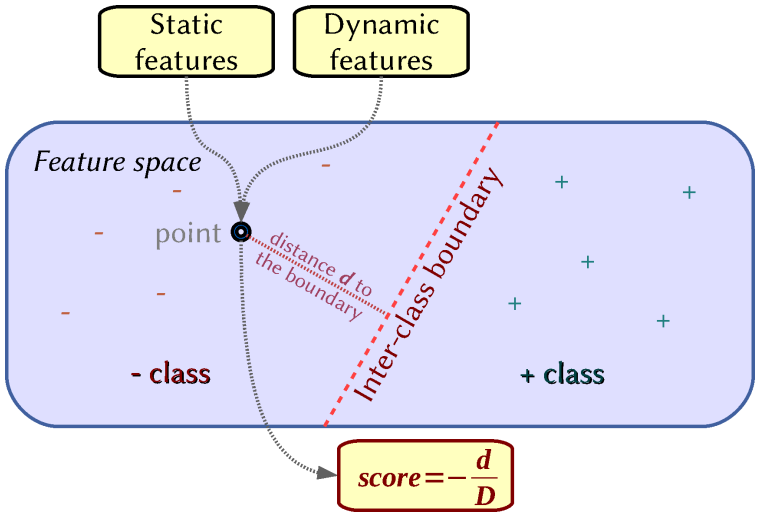


Figure 3.10: The extension to the SVM by Wiens et al that produces continuous output. A point at distance d to the class boundary and in class $\sigma \in \{-1, +1\}$ receives a score $\sigma \cdot d/D$, where D is a normalizing factor.

distance to the boundary between the classes to convert the binary {-1,+1} classification (the default positive and negative class prediction by SVMs) into a continuous [-1,+1] classification which also uses time-specific data, as shown in Fig. 3.10. Wiens et al compare their results to the state of the art classifier, i.e., the logistic regression model developed by Dubberke et al [25], finding that they improved the AUC[2] statistic significantly (from 0.69 to 0.79 in their dataset).

And the third group consists of the simple risk analysis scores. These simple risk scores are for in-situ use by nurses. They normally consider few risk factors (normally ranging from 3 to 6 in number), and serve to assess the risk of acquisition, complications, reincidence and mortality associated with CDI [33, 1]. Some of the risk factors used rely on specific laboratory test data unavailable to us (e.g., *white cell count* $> 20 \cdot 10^9/[\mathrm{L}]$), but they could still be estimated using the diagnosis data.

Other works use ideas from survival analysis [108]. Survival analysis is a statistical methodology that studies the time a disease (or risk) onsets (materializes).

3.2.2. Risk factors

Antibiotic prescription is one of the main risk factors in acquiring c.diff. It has been

[2] Area under the ROC curve, also known as C-statistic.

found that not all antibiotics are as dangerous, however. For example, metronidazole and vancomycin have been cited as antibiotics for combating c.diff [104, 61], even though other studies have not found significant differences in efficacy between antibiotics (fidaxomicin, vancomycin, metronidazole, usidic acid, nitazoxanide, rifaximin, teicoplanin) [87, 23]. Levy et al found two antibiotics as highly associated with CDI: cephalexin and cefixime [64]. Still, Dubberke et al found that adding the effect of antibiotics together was convenient as a feature in their risk model [25]; the antibiotics left in their model were: cephalosporins, clindamycin, and amoxicillin or ampicillin. Loss of protection has also been associated with consumption of gastric acid suppressors, i.e., antacids, proton pump inhibitors, and histamine H2 receptor antagonists [52].

Space and interactions are key elements of the c.diff epidemic. It has been found that physical proximity to a patient and antibiotic exposure are two strong independent predictors of CDI [13]. CDI pressure, defined as the daily average number of patients with CDI in a unit (or hospital) to whom a patient is exposed, is a significant risk factor for acquiring CDI [24]. It has also been found that using a room whose previous occupant was infected, significantly increases the risk of CDI and is more predictive than antibiotic use [108]. Patients with roommates with CDI were more likely to develop CDI as well [43]. However, the contagious nature of c.diff has been called into question; when comparing the genome sequences of the bacteria, it has been found that c.diff cases are mostly unrelated [121, 27]. Counter arguments refer to the use of low-sensitivity tests and ignoring asymptomatic carriers, but both effects have limited impact on the findings [28]. Still, the general critique of these works has been the excessive reliance on laboratory data, for the practical diagnosis of CDI does not rely exclusively on test results but other criteria as well.

Sanitization has limited effect in containing the spread of c.diff. In a study of heathcare workers treating CDI patients in a French university hospital, it was found that, after hand sanitization, CDI persisted in the gloves of workers 24% of the time [58]. Regarding sanitization of rooms used by patients with CDI, Manian et al showed that better sanitization practices reduced the rate of CDI from 0.088% to 0.055% [77].

An important couple of findings regarding the transmission of c.diff are, first, that patients expell ten times more vegetative cells than spores, and, second, that vegetative cells survive in moist surfaces for up to six hours [52]. Since an ingested cell is likely to keep producing more spores after swallowed, this suggests that more immediate contacts imply a considerably larger risk of developing CDI.

3.2.3. Modeling

Literature on mathematical models of CDI has been called *scant* [91], minor when compared to other HAIs such as MRSA and VRE [119], and in need to cover more situations [92]. This is consistent with the actual uncertainty with respect to the epidemic dynamics of c.diff, in which even basic parameters such as the transmissibility rate and incubation periods remain unknown, largely due to the always changing nature of c.diff [27, 91]. The small family of CDI models consist of compartmental models which are then simulated stochastically. These models evolved from considering simple exposure [110], to loss of immunity and complex environmental exposure [111], colonization and subsequent loss of immunity [61], and the possibility of recovering immunity [128]. Fig. 3.11 summarizes the last two models.

3.2.4. Detection

Some works have considered the risk associated with patients who carry c.diff but do not exhibit symptoms of c.diff. These *asymptomatic carriers* have been hypothesized to represent a potential risk to other patients, as it has been found that they are spreaders in long-term care settings [103]. It has been argued that their role as spreaders is of less importance than symptomatic carriers, since they are less contagious (release less spores) [63, 41].

Recent studies have also evaluated the differences in PCR and toxin detection, and have showed that PCR-positive but toxin-negative are misdiagnosed asymptomatic carriers [99, 54]. This comes to complement the findings of previous works expressing concerns with the increased sensitivity of PCR, which doubled the apparent cases of CDI [31, 39, 53, 113].

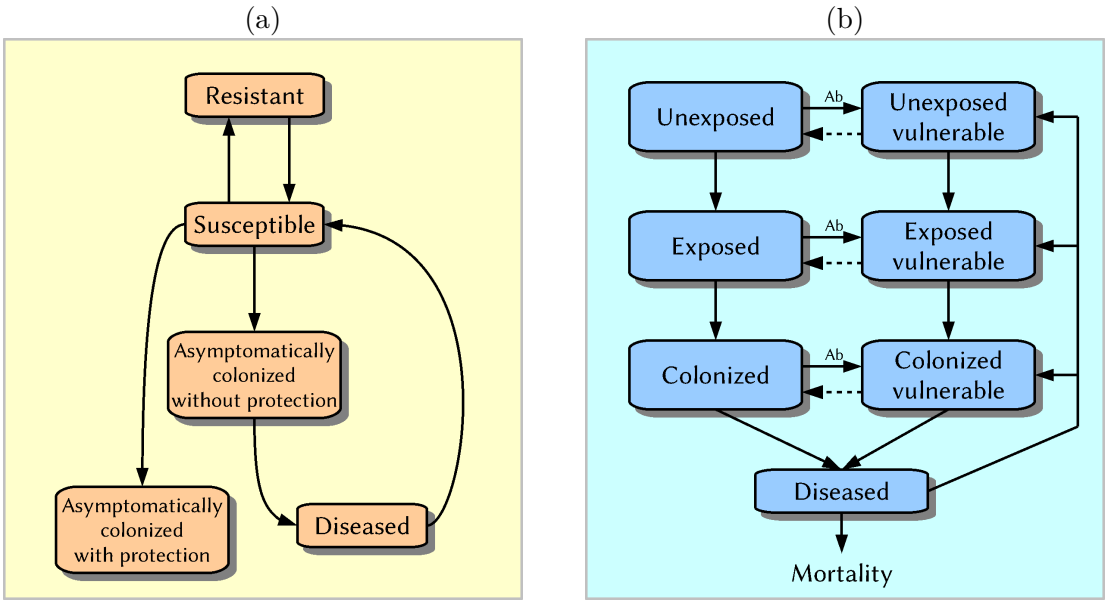


Figure 3.11: Two epidemiological models of CDI, by (a) Lanzas et al, and by (b) Yakob et al. In (b), *Ab* stands for antibiotic intake.

3.3. Building a data base of medical records

I spent the beginning of 2014 organizing and de-identifying data to be added to the *uihc* data base at the compepi server. This data I was working with was collected by the UIHC hospital, for internal use and for reporting. It basically consisted in patient data, containing diagnoses, prescriptions, procedures, physicians, etc., associated with each patient from October 2006 to December 2011.

The existing uihc data base at the compepi server contained architectural information from the hospital, logins of healthcare workers in the computer system, and patient admission-discharge-transfer (ADT) data, which accounted for the rooms visited by patients from 2005 to 2009.

The integration of the new data with the existing data base created a rich dataset that enables the exploration of medical data in a variety of ways.

3.3.1. Data sources merged into the UIHC data base

The new version of the UIHC data base in the compepi server combines information from the following datasets:

1. Hospital architecture. Room locations, aliases and distances, as derived from the hospital's architectural plans, and represented as a graph. The nodes represent small spatial units, like office rooms. If a room or corridor is too big, it is split into several nodes. We still know when several nodes represent a room or corridor. The edges of the graph represent adjacent an directly accessible nodes. For example, two adjecent offices that share a corridor but do not share a door, do not have an edge associated. But they share edges with the corridor. Finally, all the shortest distances between any two nodes have also been stored in the data base. The distance metrics include graph distance (hops, edges), physical distance, and a heuristic metric that penalizes going to upper floors through stairs more than going to lower floors.

2. AHRQ "Quality" data. This is a summary of the data submitted by the UIHC to the federal government (to the AHRQ agency, see http://www.ahrq.gov). The information collected in the dumps (last quarter of 2006 to 2011) contains details of each in-patient's visit, containing:

   (a) Patient demographics: age, ethnicity, zip code, etc.

   (b) Condition at the time of admission: the first diagnosis (ICD-9 code), the severity of the condition (how intense it is) and the risk of mortality, as estimated by the APR-DRG software (3M's All Patient Refined Diagnosis Related Groups software, as used by the AHRQ), and whether a condition was present at admission.

   (c) Complications: number of complication types and types of complications, major and minor surgery complications, healthcare-associated conditions (HACs, generalizing HAIs), etc.

   (d) Diagnoses: the condition diagnosed (ICD-9), its order (primary, secondary, etc.), whether it was present on admission, whether it is a HAC[3].

   (e) Procedures: the procedure performed, when it was performed, and the physician performing it (id).

[3]HACs are interesting because they represent avoidable healthcare costs.

(f) Physicians associated with the patient: the physicians who performed procedures on the patient (as mentioned right above) as well as the physician discharging the patient and other physicians involved in the care of the patient. We may regard this as patient-physician interaction data for social network analysis. This data may be incomplete–quite incomplete actually. In some units, it is said that, to save work, physicians only register the id of their lead physician instead of the team. The *EMR login* data (see later) can be used to more completely estimate all or most of the physicians who truly worked with the patient.

(g) Prescription or "Pharmacy" data: all the medications prescribed to the patient, when they were prescribed, the cost of the medications, whether they were used to palliate the side-effects of other medications, the routes (oral, injection, etc.) used to administer them, their form (tablet, solution, etc.), their dosage, their recommended dosage, and whether they are categorized as "high impact". Medications are described through a detailed hierarchical structure consisting of four levels, allowing for queries of variable selectivity (e.g., we could search for anti-fungals or lotions of aloe vera with coconut).

3. Admission-discharge-transfer "ADT" data. The ADT data contains precise information about each in-patient's visit (since it comes from the UIHC data base), from 2003, yet complete since 2005, until 2011, specifically about:

   (a) Precise visit information: exact admission and discharge dates, where the patient was admitted (the room), the route of admission (e.g., UIHC clinic, emergency room, transfer from another hospital, court, etc.), the priority of admission (e.g., routine, emergency, urgent, etc.), the main service category in charge of the patient (e.g., dermatology), where the patient was discharged to, the type of care after discharge (with a nurse, to an intensive care facility, etc.), the condition at admission, etc.

   (b) Room transfers: when (day and time) a patient was transferred from one room

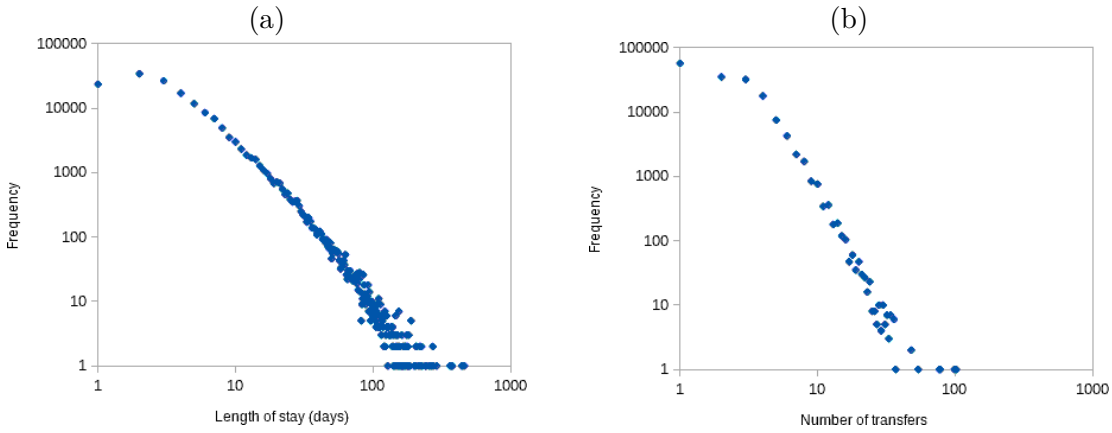


Figure 3.12: Histograms of (a) length of stay and (b) room transfers. The plots are in log-log scale, depicting almost *power-law* distributions.

to the next, the rooms involved in the transfer, the units hosting the rooms, whether the transfer was urgent or critical, etc. Location data can be linked to hospital architectural data.

4. "CDI" laboratory order data. For each patient tested and diagnosed with clostridium difficile, we know the exact day the laboratory test was ordered, the unit where it was ordered from, and the type of test performed, from 2005 to 2011. We do not have data on negative results.

5. EMR login or "keyboard" data. This data consists of all the computer logins performed by healthcare workers, telling who logged into which computer (its location) and when (day and time). We have data from 1 August 2006 to 21 June 2008.

Table 3.5 reports general statistics (counts) of the loaded data and Table 3.6 reports the average statistics per patient visit.

3.3.2. Profile of the arriving patients

Admitted patients stay for an average of 5.8269 days, standard deviation of 9.6653, minimum of 1 and maximum of 462 days. Patients change rooms 2.04 times in average, with a minimum of 0 (likely missing data) and a maximum of 102 times. Details on the distribution of length of stay and room transfers are depicted in Fig. 3.12.

Item	*Subitem*	*N*
Admissions	*all*	208,902
	per year	29,783
Patients		126,265
Zip codes		3,106
Diagnoses		5,361
Prescriptions		7,788,703
Medications	*all (known)*	2,491
	prescribed	935
	high impact	765
	mid impact	1,045
	low impact	681
Physicians	*all*	25,065
	EMR logins	14,596
	Quality	11,455
	Procedures (Quality)	1,449
	Discharge (Quality)	851
	Other (Quality)	10,764
EMR data	*logins*	19,800,874
	computers	17,520
	rooms	4,379
Transfers		426,155
Rooms	*all*	19,559
	bedroom	561
	observation	17
Units		186

Table 3.5: Miscellaneous statistics of the collected data: number of entries by data type.

The *source of admission* tells us whether there is a risk a patient came with a HAI or with a depressed immune system, for instance. For each source, Table 3.7 displays the number of patients, average length of stay, and the first and last times the source type was used. But sources change over time. For example, *emergency room* was removed[4] as an admission source after 1 July 2010, and *transfer from another unit* began being used by 27 December 2007. Similarly, several sources stopped being used during 2008.

The *type of admission* also describes the overall condition of the admitted patient. Table 3.8 shows the statistics associated with the five admission types in the data base. Many are associated with the *emergency* type, which demands urgent care for life threatening

[4]Not physically, just as an entry for *admission source*.

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

Table 3.6: Miscellaneous statistics of the collected data: statistics of the data linked to a visit. Most of the zeroes in the table are explained by incomplete (missing) data.

conditions. The *urgent* type is for other priority admissions that are not life threatening.

One in five patients (41,825, 20% of the admissions) were associated with conditions (diagnoses) that were not present on admission. These patients had 3.2030 such conditions on average, standard deviation of 2.91, minimum of 1 and maximum of 19. On complications, only 10,751 patients (5.15%) were associated with them. These patients had an average of 1.3619 *types* of complications, standard deviation 0.711, minimum 1, maximum 7.

Regarding HAIs, several patients were diagnosed with drug resistant organisms: 1,654 were diagnosed with CDI (however, from laboratory data, we know that 1,851 had CDI); 1,001 had methicillin-resistant staphylococcus aureus (MRSA) and 395 were suspected of having it; 482 had vancomycin-resistant enterococci (VRE); and 873 were infected with other drug resistant microorganisms.

3.3.3. Network data

The main type of interaction recorded in the data are patient-physician relations, where a *patient* specifies a patient but still distinguishes between different admissions. There are 614,079 such relations in the data base (putting together the physicians who performed procedures, discharged, or were involved in any direct way with the patient). Patients were associated with a an average of 2.9396 physicians, with standard deviation of 3.0395, a maximum of 15 physicians, and a minimum of 1 physician (of course). A more detailed depiction

Admission source	*N*	*Average LOS (days)*	*First use (yyyy-mm-dd)*	*Last use (yyyy-mm-dd)*
UIHC clinic	53963	0.4080	2004-01-29	2011-12-30
To acute hospital	51628	1.7509	2004-04-05	2011-12-31
Non healthcare facility	35247	0.5241	2005-04-01	2011-12-31
UIHC emergency room	34297	0.4538	2003-12-13	2010-07-01
Born in UIHC	7919	3.5981	2007-05-18	2011-12-29
Court order	2919	0.0277	2004-10-05	2011-12-01
Other healthcare facility	1427	0.8381	2004-12-25	2011-12-30
Unit transfer (UIHC)	1372	0.1130	2007-12-27	2011-12-27
Skilled care facility	1156	1.0164	2004-12-06	2011-12-26
Prison, jail, halfway house	744	0.5242	2005-01-08	2011-12-24

Table 3.7: Top 10 most common sources of admission (that were still in use by 2009). LOS stands for length of stay.

Admission type	*N*	*Average LOS (days)*	*First use (yyyy-mm-dd)*	*Last use (yyyy-mm-dd)*
Emergency	75329	4.6419	2003-12-13	2011-12-31
Urgent	58769	5.7096	2004-01-29	2011-12-30
Elective or routine	57414	3.3986	2004-09-17	2011-12-31
Newborn	12680	7.7865	2004-08-08	2011-12-29
Trauma center	4710	6.2146	2005-01-24	2011-12-30

Table 3.8: The five admission types, and their statistics. LOS stands for length of stay.

of the number of physicians associated with an inpatient is depicted in Fig. 3.13, which shows that patients are not associated with too many physicians (inset a), but physicians are associated with many more patients, in a heavy tailed distribution (inset b, quasi-linear distribution in log-log scale, i.e., *power-law*).

As argued in the introduction, few works have used contact networks to study epidemics in healthcare settings. Such networks are expensive to generate. The cheapest solution is to use medical record data (as in the uihc data base). Two studies used contact networks generated from records data to study the spread MRSA in a neonatal intensive care unit [37] and, through simulations, in a whole hospital [19], the latter even including ADT data. And not in the intention of studying epidemics but studying contact patterns, Barnett and Landon et al constructed large scale patient-sharing networks, i.e., physicians who worked with the same patients, with data from 51 hospitals and using survey and medical records

[5, 59, 4, 60].

3.4. CDI at the UIHC hospital

3.4.1. CDI diagnosis at the UIHC

The testing methodology for CDI at the UIHC has changed over the years. Until 2009 (inclusive), tests consisted in the detection of toxins A and B, with results coming out in 24 hours, using an EIA test (*VIDAS' Toxin A & B Detection*) [21]. This changed during December 2009, when PCR started being used in the detection of CDI, reducing the result time to within 2 hours and the added specificity of the test [20]. But restrictions were introduced in 2012, limiting testing of a patient to once every 7 days, and to 10 days if the previous result was positive, since the PCR test (*Xpert C.difficile*) would detect c.diff even after cessation of symptoms [22]. This is in line with testing recommendations to reduce the rate of false positives, especially because of the high prevalence of colonized but not infected carriers [98].

We have 2,103 positive tests results (both tests) from 2005 to 2011 (inclusive), spanning 1,851 visits, but only totalling 1,651 diagnoses (0.79% of the admissions). Redundant testing (less than 10 days between positive results) occurred 84 times, spanning 76 visits. The cases roughly doubled after the introduction of the PCR test, from an average of 0.6076 to 1.2268 cases per day, as shown in Fig. 3.14.

3.4.2. Who is diagnosed with CDI

As it is well known, CDI cases tend to be more frequent for the elderly [11, 2]. Fig. 3.15 shows the rate of CDI in patients by age group, in bins of 5 years. However, our data suggests that children between ages 1 and 10 are also at increased risk of developing CDI, and that the risk grows continuously with age.

Community-acquired c.diff infection, i.e., CDI present on admission to the hospital, totals 486 cases (26.25% of visits associated with CDI). These cases can be disagreggated by direct admission at the UIHC (287 admissions, 59.05% of community-acquired CDI), referral from UIHC clinics (150, 30.86%), and transfers from other healthcare facilities (49,

Condition/diagnosis (ICD9-CM)	*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 [89].

Table 3.9: The 10 most frequent conditions diagnosed to patients diagnosed with CDI, listed before the diagnosis of CDI.

10.8%), which include long-term care (nursing homes). 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.

Regarding the characteristics of the medical treatment experienced by the patients who developed CDI, Tables 3.9, 3.10, and 3.11 show the most common comorbidities, antibiotics prescribed and procedures performed on patients diagnosed with CDI. The conditions presented in Table 3.9 are both associated with CDI and general acute condition. Items 1, 2, 9, and 10 are associated with diarrhea and intestinal failure, which are very consistent with CDI [89]. The rest of the items depict a scenario of acute condition; items 4 to 6 are associated with respiratory problems, and items 3 and 6 are associated with sepsis (an infection taking over the host). These conditions, however, are known to co-occur with CDI [109].

The treatment of patients who develop CDI is mostly consistent with those in acute condition. But the presence of CDI introduces changes in their care. Antibiotic prescription increases roughly 2.7 times after the development of CDI. Table 3.10 shows the most common antibiotics prescribed to patients who develop CDI, both before and after the diagnosis. Metronidazole and vancomycin are associated with the largest increases in prescription rates after the diagnosis of CDI. These two antibiotics are known to target c.diff

	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

Table 3.10: The 5 most frequent antibiotics prescribed to patients after being diagnosed with CDI, since 2010, ranked by increase of use after CDI and overall frequency of use. Coincidentally, these antibiotics are the most commonly prescribed before and after the onset the disease, independently. The first two antibiotics, metronidazole and vancomycin, are used for treating 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.

Table 3.11: The 5 most frequent procedures performed on patients after being diagnosed with CDI, since 2010. The first, third and fourth items are consistent with the condition of CDI.

[87, 64]. With respects to procedures, items 1, 3 and 4 of Table 3.11 are consistent with CDI; items 3 and 4, in particular, can be explained by intestinal problems. 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.

3.4.3. Example of a case

Fig. 3.16 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 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).

3.4.4. Ubiquity of the cases

We may wonder whether there is a relation between occupancy rate and cases of CDI in a unit. In an attempt to explore this question visually, we plotted Fig. 3.17, in which we see the number of cases of CDI in four different units over time, plotted againts the occupancy of the unit (just the number of patients who were there), by month. The units chosen correspond to those with the highest rates of CDI, and all of them provide acute care (they treat weakened patients). However, the low counts, together with the nonlinear dynamics of CDI, do not visually demonstrate associations between the series, except, perhaps, in the MICU (Fig. 3.17-(c)).

3.5. Conclusions

In this chapter, I reviewed the threat that c.diff poses to healthcare provision, some of the extensive literature about this *superbug*, and described a data base of anonymized medical records that I will use to study CDI in the hospital.

I am interested in the problem of identifying the patients who are at risk of c.diff infection. By knowing how c.diff spreads and develops, healthcare workers and officials are able to design prevention practices and containment strategies. In particular, knowing which patients are at greater risk of developing CDI enables healthcare workers to provide individualized care. This sets the importance of identifying the patients at risk of developing CDI, which corresponds to the computational problem of *classification*. Classification variants allow for different use-cases. Binary, hard classification (traditional classification) identifies the patients with the highest likelihood of developing CDI; this is the task for which we have ground truth in the data (the actual cases of CDI). Soft classification, the translation of classification into soft membership or ranging probabilities (calibration), informs about the actual exposure and susceptibility, and it is informed by the hard classification task. This is the task of *risk estimation*. Dynamic risk estimation (updating, contingent soft clas-

sification) allows for the timely identification of patients in need of preventive care, while it also provides a deeper understanding of the changing susceptibility to CDI and the ever changing exposure (contamination). Every classification variant is associated with use cases valuable in the context of CDI care.

In the next chapter, I present and validate a method for analyzing these medical records data for predictive cases of CDI, alongside with discovering useful information about risk factors of the disease. The classifiers I develop, which are ensembles of logistic regression models, effectively perform hard and soft classification, and are able to produce dynamically changing curves of risk of developing CDI.

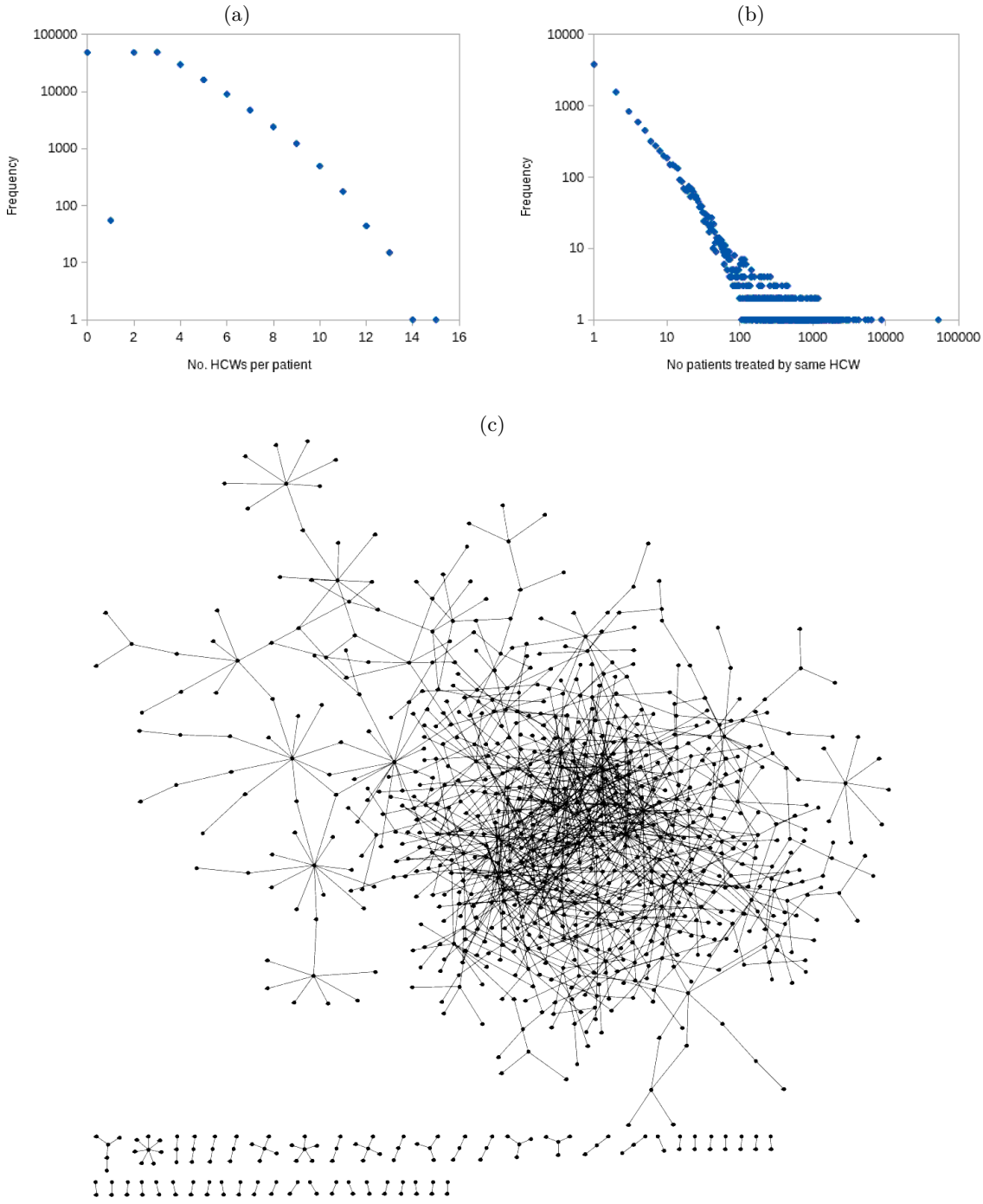


Figure 3.13: Patient-physician associations. Insets: (a) histogram of number of physicians associated with a patient, (b) histogram of the number of patients associated with a physician (log-log; shows a power-law), and (c) the network of patient-physician procedures during the first 7 days of 2007.

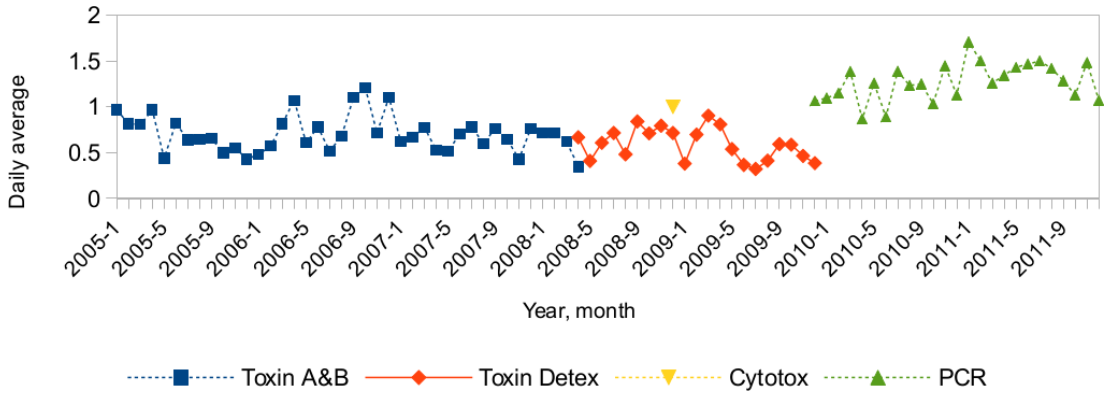


Figure 3.14: Daily rates of CDI by testing methodology in the UIHC. The number of cases is normalized to *daily averages* instead of the total number of cases per month, because of the inherent differences between months (30-31 days, or 28-29 for February) and because some tests were not performed during the entire month, such as December 2009, where the switch from toxin detection to PCR occurred. With the introduction of the PCR test, cases rose from an average of 0.6076 to 1.2268 cases per day.

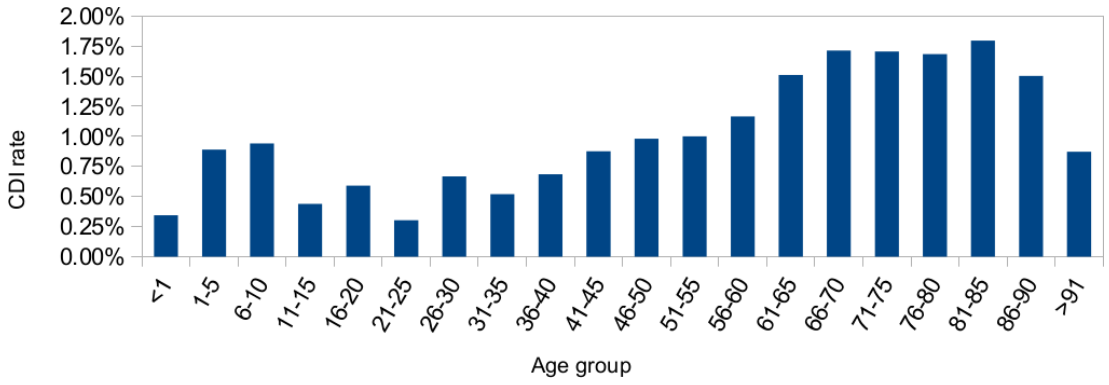


Figure 3.15: Rate of CDI cases by age in the UIHC. Each bar depicts the proportion of patients admitted that developed CDI within that age group.

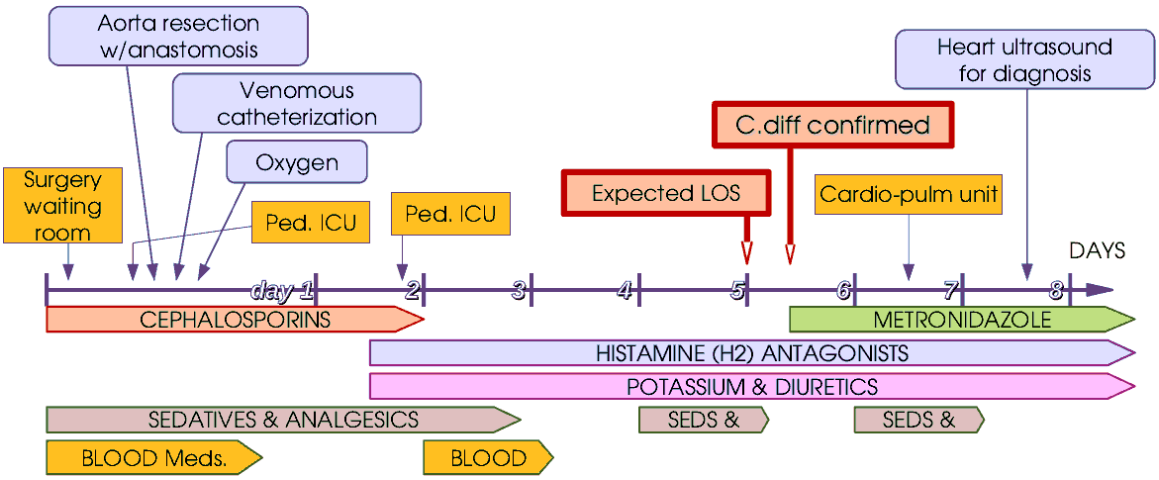
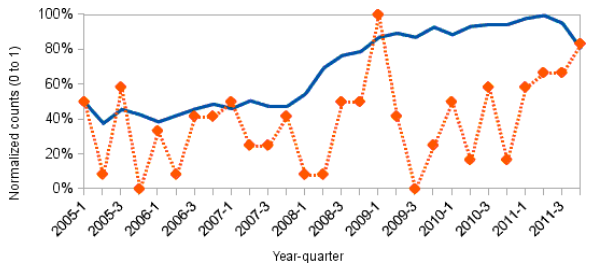
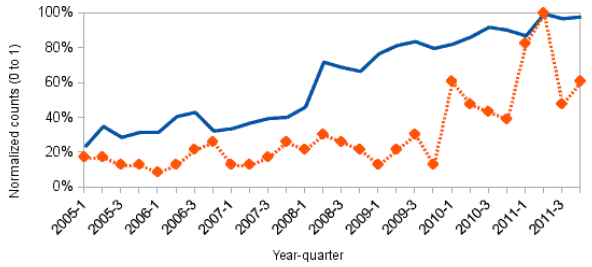


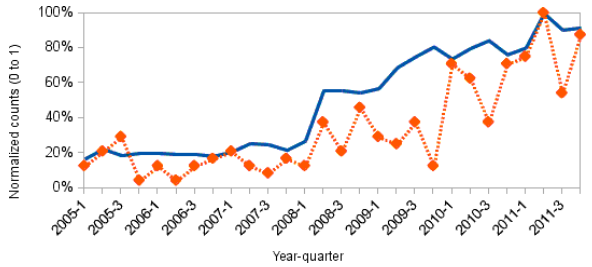
Figure 3.16: A case of CDI in the hospital. A child is admitted to the hospital for a scheduled cardiac surgery and develops CDI while in care.



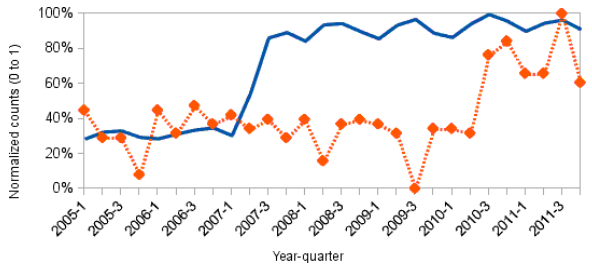
(a) 4 Roy Carver East (4RCE, Heart and Vascular Center)



(b) 6 Roy Carver East (6RCE, Cardiac Rehabilitation)



(c) Medical Intensive Care Unit (MICU, 5 Roy Carver East)



(d) Operating Room (OR, 5 John Colloton)

Figure 3.17: Normalized cases of CDI/quarter in the UIHC, in four different units: (a) 4 Roy Carver East (Heart and Vascular Center), (b) 6 Roy Carver East (Cardiac Rehabilitation), (c) the Medical Intensive Care Unit (5 Roy Carver East), (d) the Operating Room (5 John Colloton). The plots have been normalized to make the *occupancy* and *CDI rate* curves comparable. However, there is no clear association betweeen these curves.

CHAPTER 4
PREDICTING CDI THROUGH ORDERED PAIRS OF EVENTS

4.1. Introduction

As mentioned in the previous chapter, c.diff (Clostridium Difficile) is a bacterium that causes its carriers to undergo severe colitis when their intestinal flora and overall immunity is compromised, a situation referred to as CDI (Clostridium difficile infection). On top of that, c.diff is contagious through physical contact and its spores can last on surfaces for several months. So far, it can be effectively erradicated only through proper hygiene and the usage of bleach or UV rays; soap and alcohol are not effective means of killing the spores. Considering all these facts, preventing the occurrence of CDI and the spread of the bacterium is a desirable goal in health care provision.

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 [25]. Garey et al developed a CDI-prediction model, but only for patients receiving broad-spectrum antibiotics [35]. Wiens et al. have built models that compute evolving risk scores for patients [124, 125], and another model that only uses data that were available within 24 hours of a patient's admission to the hospital [123]. In addition, there are simplified risk scores that can be calculated using only 4-5 features [33].

Predicting outcomes from medical records is difficult for a number of reasons. Medical records consist a variety of data types [51], and medical records often contain inaccuracies, biases and censoring [95]. In addition, medical records evolve over time. Some researchers have mined frequent events or common patterns of clinical events using partial orders [93, 94], sequences [114], and temporal abstractions [62], but these approaches have shortcomings as well, including inflexible representations of time and poor interaction between the classifier and the patterns discovered [67, 47].

In this chapter, we propose a method 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 [68], fitted with regularization, as in [30]. 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 [93, 94]. 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 successfully make use of hierarchical relations among the features to produce more informative models. And fifth, we contribute with a method for building classifiers for the situation of class imbalance and high dimensionality.

4.2. Related work

Several works have focused on developing CDI risk estimators for in-patients using medical records data. Garey et al developed a model for predicting CDI in patients receiving broad-spectrum antibiotics [35], which they validated through out-of-sample testing. Later, Dubberke et al introduced a model for predicting CDI for any patient admitted to the hospital [25]. Two successive works of Wiens et al focused on building evolving risk scores for patients, and proved that their model outperformed the model of Dubberke et al [124, 125]. Finally, a new model by Wiens et al focused on building a logistic regression model trained on extensive medical records data, but limiting only to knowledge within the 24 hours of admission, outperforming all previous estimators [123]. (By structure, [123] generalized [35].) Also, besides the previous works, many have developed simplistic, *rule-of-thumb* risk scores for clinical use, most of which consider only 4 to 5 features [33].

Mining medical records to predict risk of undesirable outcomes is not new; Li et al was

mining risk patterns in medical data and warning of weaknesses of some mining approaches back in 2005 [67]. Jensen et al comment on the challenges and opportunities of mining medical records, including mining varied types of data [51]. Paxton et al discusses challenges of mining medical records, including the presence of inaccuracies, biases, confounders and censoring (incompleteness) [95]. Tran et al developed a system that mines patient visits by converting events (e.g., diagnoses, medications) to time series, and then supplying these series to a classifier [117]. Lin et al developed a system that predicts rheumatoid arthritis by performing classification on features derived from clinical texts (with the added difficulty of natural language processing) [69] and using laboratory values (continuous data) [70]. Lim et al used an ensemble of logistic regression models to evaluate the outcome of treatment for acute myeloid leukemia using gene expression data [68]. (The ensemble was used to classify under the restriction that the dimension of the data has to be smaller than the number of instances in it, not for feature selection, as we review at the end of this section.) Halpern et al developed a system to estimate the state of the patient by looking at *anchor variables* [42]. Several of the works mentioned reduce the dynamic, temporal data to a static expression suitable for traditional machine learning, describing visits as features with a fixed dimension (Fig. 4.18).

Several works concerned with mining medical records use, adapt, and extend techniques and ideas from *pattern mining*. Pattern mining is the task of mining frequent *patterns* from data, where a pattern describes a portion of the data. For example, if our data consisted of items bought together in the same sale, then the patterns would be subsets of those sales[5]. But medical records contain temporal data, in particular, *events*. Medical events patterns have been mined as event sequences [114] and, more generally, as partial orders [93, 94]. Besides events, clinical data also considers vital signs and other continuous information. Several works have address this problem by discretizing these multivariate time series and mining *time-intervals* through *temporal abstractions* [7, 8, 86, 85]. Temporal abstractions have also been used to mine medical events [62]. However, some have argued that this

[5]This is actually called *frequent itemset mining*, because items bougth together are called *itemsets*.

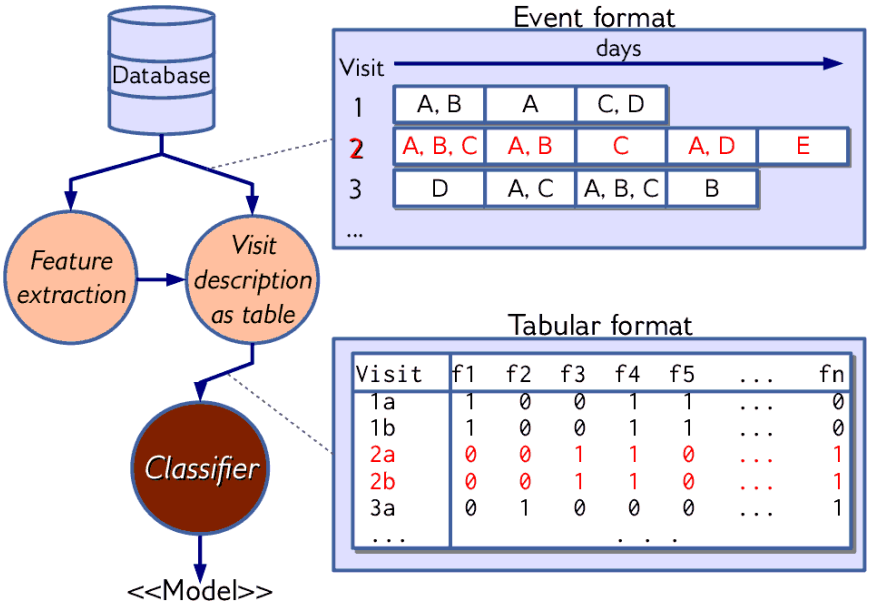


Figure 4.18: From temporal/event data to its static equivalent.

approach is inherently limited and have called for a better treatment of time [47]. Other approaches have also been considered, including convex optimization [38], dynamic time warping [125], and supervised mining using visualization [40].

Very related to pattern mining, association rule mining is about finding patterns that are *associated* to outcomes, even though these outcomes might not be predictive. For example, there is work on predicting prescriptions for patients based on their prescription history [14, 126] and ordering diagnostic tests [6]. Since association rules can be described as event sequences, sequence mining has also been used to discover associations [114], including antibiotic prescription and development of resistance [112], identifying subjects that meet the criteria for inclusion in further classification [66], or even treating different visits as events and mining associations between the diagnoses associated to those visits [44]. Other work has focused on the harder problem of extracting rules from clinical texts, which also has to address the problem of natural language processing [127]. However, it has been recognized early that pattern mining and classification do not necessarily work well together [67], which is consistent with the idea that an association might appear relevant by itself but, when considered together with other rules, it might lose its relevance.

Association rule mining also addresses the more general task of identifying salient or relevant features, which are often called *risk factors* in the medical literature or *biomarkers* if they are related to laboratory data (e.g., genes, vital signs), for diagnostics or early detection of undesirable outcomes. Traditionally, the association between a feature and an outcome has been measured with simple, pair-wise statistics (e.g., correlations, t-tests, χ^2 scores), so naturally several works have been devoted to improving these inexpensive measures by developing better measures and removing redundant features [12, 116, 132, 131]. Peng et al developed a system that iteratively discards irrelevant features (using simple pair-wise statistics) and then filters the best features through support vector machines and ROC curves [96]. (ROC curves can be used to compare and rank classifiers, and are indifferent to class imbalance [29].) Chen and Wasikowski developed a method that combines nearest neighbors and ROC curves for filtering features [15]. Related to ROC curves, there are works concerned on measuring the contribution of features to learning tasks as well [57, 76]. Zhou et al developed a regularization-based method (FeaFiner) for identifying biomarkers through feature selection and generalization, a challenging task due to the vast richness of medical data and the natural semantic association between medical data (e.g., hierarchies) [133].

The most seemingly challenging problem of finding relevant risk factors or biomarkers has been the high dimensionality of the data, a problem that lately has been addressed through ensemble methods. Several methods based on ensembles of logistic regression have been proposed, most of them feeding each model with a randomly chosen subset of features [122, 130], some even using the ensembles for feature selection while addressing class imbalance [129] or combining other methods for filtering features [10, 49], and there are methods that use other regression techniques as well [106]. Non regression-based methods have also been used [105], which serves as a reminder that Random Forests is a particular case of an ensemble classifier where each (unprunned) decision tree is fed a randomly chosen subset of features. These ensembles can be classified as bagging of classifiers, since all classifiers within them are treated equally [34].

4.3. Visit data

We use a particular subset of the UIHC database: the data describing the care associated with a patient admitted to the hospital, which we call *visit* data. (Visits are also known as *encounters*, even though the latter term is more general for it can include out-patient visits.) 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. *Clinical event data* represent what occurred during the visit. We consider four types of event data: prescriptions, procedures, transfers, and positive CDI laboratory tests.

Diagnoses, procedures, and prescriptions are associated with hierarchies, as illustrated by Fig. 4.19. Diagnoses and procedures are categorized into ICD-9 and CCS codes. We prefer CCS codes [26], 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.

4.4. Feature engineering

4.4.1. 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. 4.18, as has been previously reported [7, 8, 86, 85, 114, 93, 94, 62, 124, 125, 123]. 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. , consists of an ensemble of logistic regression models. The model produced by the

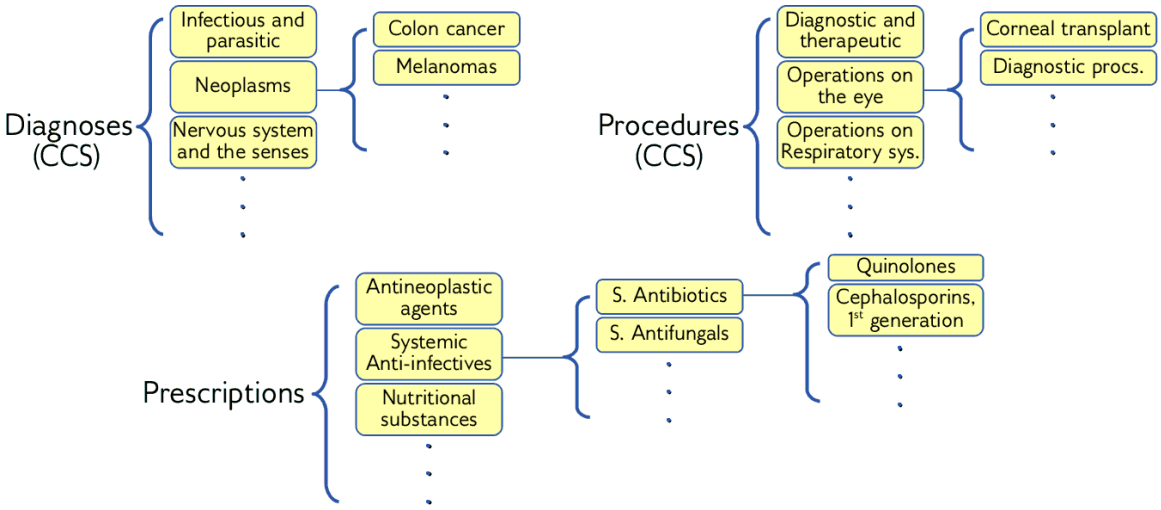


Figure 4.19: Partial view of the hierarchies associated with diagnoses, procedures, and prescriptions.

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.

4.4.2. Bare events and pairs of events

We now formally define how we describe events, visits, and translate them into features.

Definition 4.1. We define a *visit* V as a set of *events* corresponding to a patient's visit. Each *event* $(t, e) \in V$ is comprised of a time t and an *event action* e.

Example 4.2. The first visit shown in Fig. 4.18 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 represents days, where $t = 1$ means the first day of that patient's visit.

We described 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.

Definition 4.3. 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 4.4. 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 4.5. 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 4.6. 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*, corresponds to the basic representation of visits in previous research [35, 25, 123]. Feature *received laxatives* in Dubberke et al is an example of this [25]. Our proposal, however, consists in combining event actions in temporal order for representing visits.

Definition 4.7. 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 4.8. 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 [93, 94]. 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, but 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.

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

Definition 4.9. 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 4.10. As shown in Fig. 4.19, 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 (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 4.11. 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 4.12. 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 4.13. 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 description of the data.

Definition 4.14. Let D be a collection of tuples such that for every $(V, c) \in D$, V is a [partial] visit and c is a class (i.e., *CDI* or *non-CDI*), and let F be a function that takes a [partial] visit and returns a set. Then, *the tabular description of* D *induced by* F is defined as follows:

1. The totality of the features of the data set are $\mathcal{T} = \cup_{(v,c)\in D} F(v)$.

2. The headers $\mathcal{H} = (h_1, ..., h_m)$ introduce an ordering on $\mathcal{T}$, i.e., $m = |\mathcal{T}|$, $(\forall 1 \leq i \leq m)$ $h_i \in \mathcal{T}$, and $(\forall 1 \leq i, j \leq m)$ $h_i \neq h_j$ if $i \neq j$.

3. Each $(V, c) \in D$ defines an instance (x, c), where x is an m-dimensional binary vector defined as

$$(\forall 1 \leq i \leq m) \;\; x_i = \begin{cases} 1, & \text{if } h_i \in F(V), \\ 0, & \text{otherwise.} \end{cases}$$

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 descriptions, although we take advantage of the sparse description of visits internally in our code.

4.4.3. 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* [124, 125, 25, 123], 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.

4.4.4. 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 method is presented in the next section.

4.5. Classification

4.5.1. 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 [68]. As shown in Fig. 4.20, 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 [46].

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 method described in Section , 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.)

4.5.2. 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 [10, 49, 106, 122, 129, 130]. Instead, we perform feature selection inside each classifier. This produces an ensemble clas-

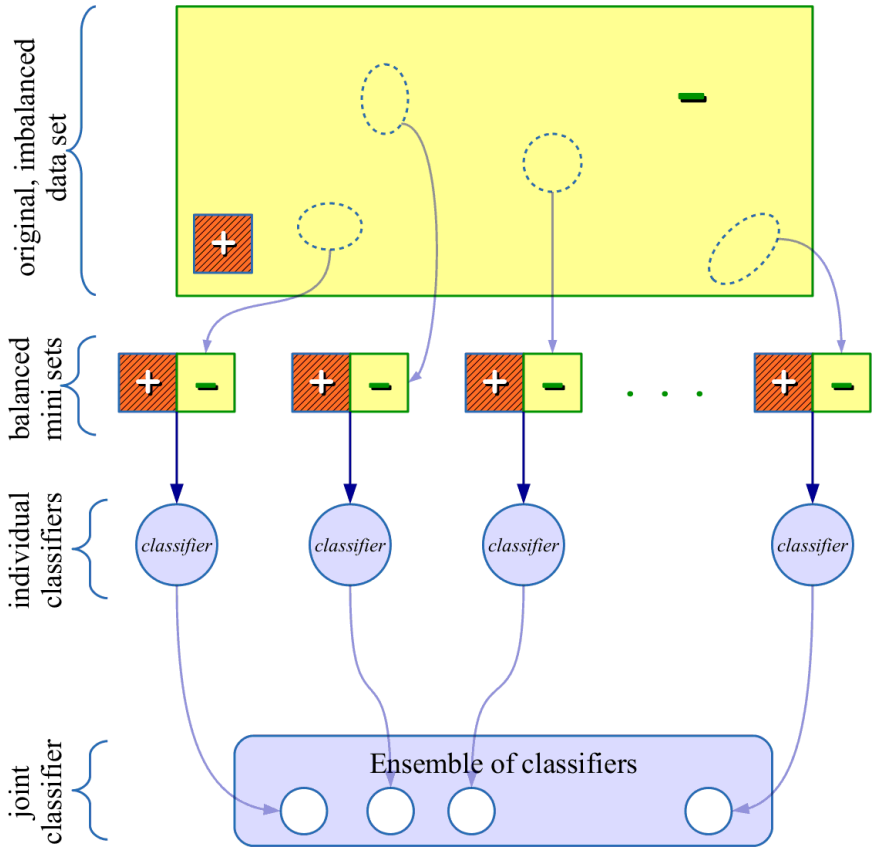


Figure 4.20: The ensemble approach to class imbalance.

sifier 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 [65].

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. 4.14) and returns the m "most likely" predictive features, with $m = 1000$[6]. This algorithm is also explained in Fig. 4.21.

In the second stage, we use L_1 regularization to further reduce the number of features.

[6]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.

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

Algorithm 1: Greedy randomized embedded feature filter

The L_1-regularized cost function for fitting logistic regression is

$$L(\alpha, \beta; \lambda) = \lambda|\beta|_1 + \sum_{(x,y)\in S} \ln\left(1 + \exp(-y(\alpha + \beta^T x))\right), \tag{4.5.1}$$

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. 4.14), 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 [30], choosing λ by minimizing the Bayesian Information Criterion (BIC) of L, which applied to our problem is

$$BIC = -2L + (1 + |\beta|_0) \ln |S|. \tag{4.5.2}$$

Using L_1-regularization for feature selection has theoretical backing. The $L_0 - L_1$ *equivalence* [71] 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.

4.5.3. Estimation

We predict CDI by converting each visit into a feature description (Def. 4.14), and supplying this description to each logistic regression model in the ensemble. Then, the

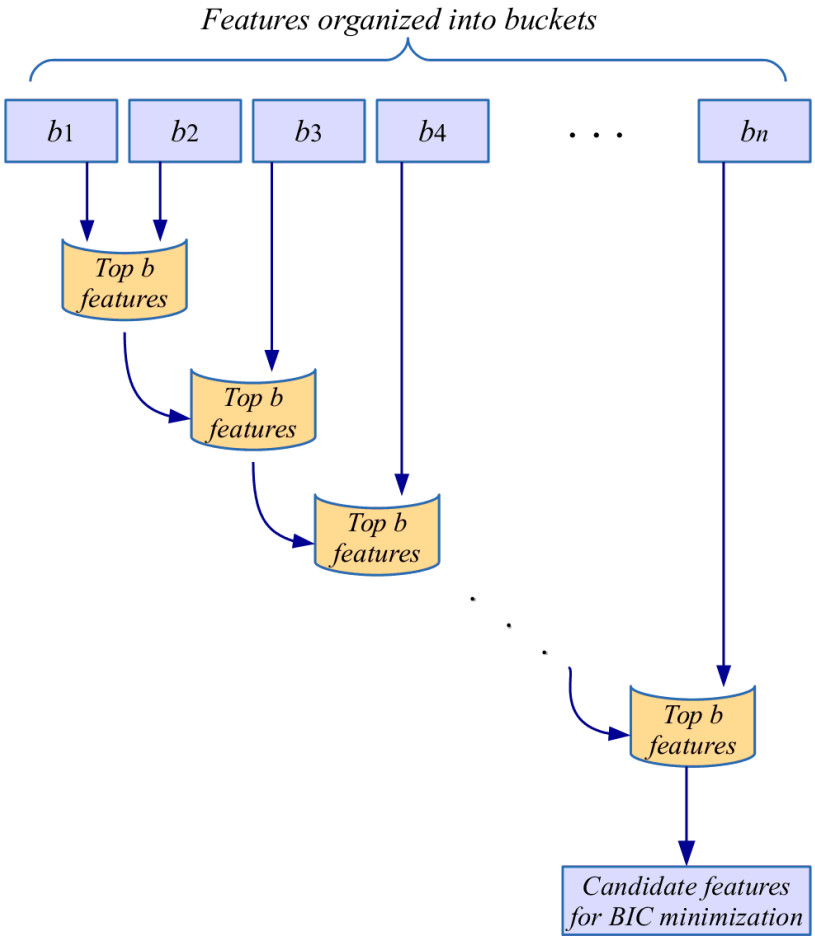


Figure 4.21: First pass of feature selection.

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.

4.6. Experiments

4.6.1. Experiments outline

Our first experiment compares our method to the state of the art: the work of Wiens et al [123]. 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.

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.

4.6.2. 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 method 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 [123]. Table 4.12 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 latter 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

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

Table 4.12: Characteristics of the three classifiers.

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 4.13: 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.

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 4.13 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. 4.22. Note that the AUC of SAC is similar to the one shown in Wiens et al [123]. The low sensitivity and high specificity 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.

4.6.3. 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

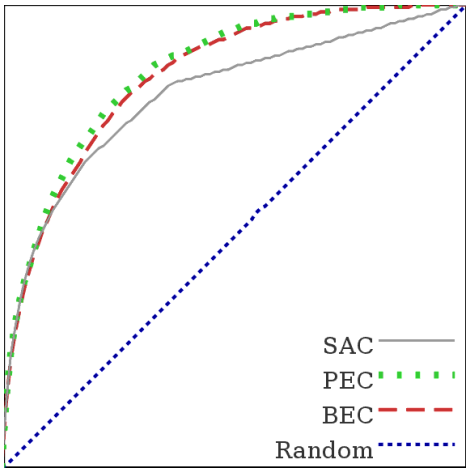


Figure 4.22: ROC curves of the classifiers in the task of predicting CDI, using 1 or 2 days of a visit. The ROC curves are averaged over the 10-fold cross validation tests. The *Random* curve stands for the uninformed classifier.

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 3.6 shows, most visits last only a few days.)

Table 4.14 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. Figure 4.23 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 4.13). 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. Fig. 4.24 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

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 4.14: 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.

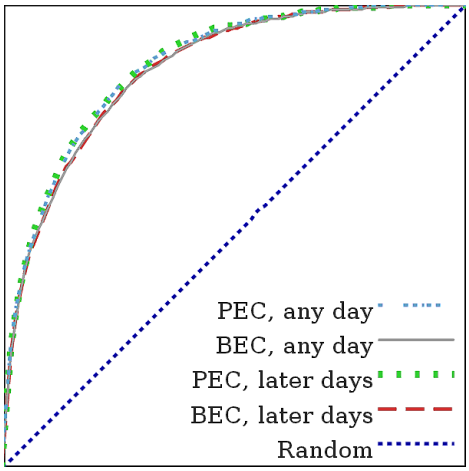


Figure 4.23: ROC curves of the classifiers in the task of predicting CDI, using any day of a visit. The ROC curves are averaged over the 10-fold cross validation tests. The *Random* curve stands for the uninformed classifier.

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 4.15 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.

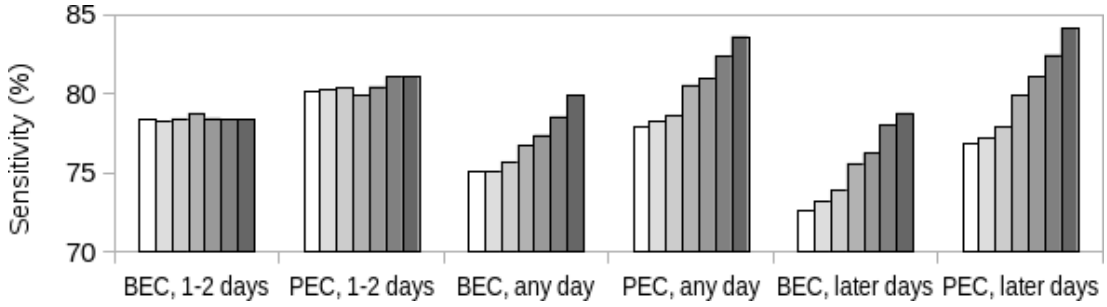


Figure 4.24: BEC and PEC classifiers as the time to CDI approaches. The bars represent the sensitivity of the classifiers from 7 days to the day before the onset of symptoms.

	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 4.15: 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.

4.6.4. 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 4.16 shows the performance of BEC and PEC classifiers trained using either admission data or clinical events while Fig. 4.25 compares their ROC. Classifiers using admission data clearly outperform classifiers using clinical events, which confirms that information available at admission time can indeed be used to predict whether a patient will develop CDI during the visit.

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 4.16: 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.

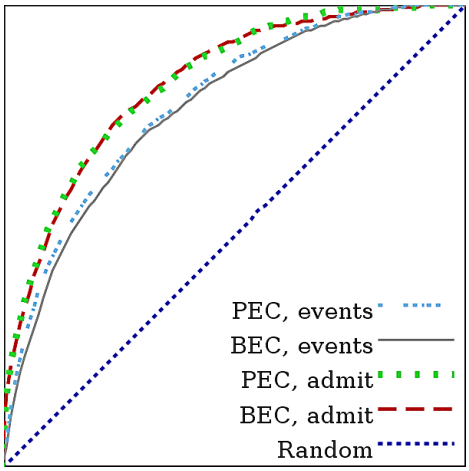


Figure 4.25: ROC curves of the classifiers in the task of predicting CDI, using admission data only (admit) or clinical events only (events). The ROC curves are averaged over the 10-fold cross validation tests. The *Random* curve stands for the uninformed classifier.

4.6.5. 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 4.17 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 4.17 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

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

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 4.17, include *intestinal infection* (@Diag=135), *osteoarthritis* (203), *pancreatic disorders except diabetes* (152), *nutritional deficiencies* (52), *abdominal hernia* (143), and *other lower respiratory disease* (133).

4.6.6. 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 4.18 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 4.18. 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.

Feature name	$\beta_{[x<y]}$	$\beta_{[y<x]}$
`[To=OR < To=PORR]`	-0.1831	-1.0547
transferred to OR no later than transferred to PORR		
`[To=OR < Proc=Diag/Therap]`	0.3556	-0.0004
transferred to OR no later than underwent miscellaneous diagnostic-therapeutic procedure		
`[Proc=223 < Proc=231]`	0.1982	-0.0019
underwent 'enteral and parenteral nutrition' no later than underwent 'other therapeutic procedures'		
`[Proc=231 < RxMin=812]`	-0.014	-0.2062
underwent 'other therapeutic procedures' no later than prescribed 'antibiotics systemic'		
`[Proc=Diag/Therap < RxSmin=81206]`	0.1781	0.0144
underwent miscellaneous diagnostic-therapeutic procedure no later than prescribed 'fourth generation cephalosporins'		
`[RxMaj=40 < RxSmin=562210]`	0.1362	0.0079
prescribed 'nutrients/nutritional agents' no later than prescribed '5ht3 receptor antagonists'		
`[Proc=216 < RxSmin=81219]`	0.0042	-0.1182
underwent 'respiratory intubation and mechanical ventilation' no later than prescribed 'extended-spectrum penicillins'		
`[To=6RCE < Proc=Diag/Therap]`	-0.0055	-0.112
transferred to 6RCE no later than underwent miscellaneous diagnostic-therapeutic procedure		
`[RxSmin=280892 < RxSmin=81203]`	0.1132	0.0096
prescribed 'misc analgesics systemic' no later than prescribed 'first generation cephalosporins'		
`[Proc=177 < RxSmin=280808]`	-0.0161	-0.1191
underwent 'computerized axial tomography (ct) scan head' no later than prescribed 'opiate agonists'		

Table 4.18: Top 10 ordered events where the order is relevant. For each ordered event, we include a human readable description of it as well as two log-odds: the individual log-odds of the ordered pairs $\beta_{[x<y]}$, and its converse $\beta_{[x<y]}$.

4.6.7. 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 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 [121, 27].

However, here we need to emphasize the limitations of using our classifier output to estimate the "importance" of features. Note that our classifier aims to produce a simple,

minimally redundant explanation of risk. For example, if two features (bare or pairs of events) have a similar explanatory power, the classifier will choose only one most of the time. 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 status on admission. Moreover, many procedures are associated with particular locations in the hospital, because of the medical speciality associated with the operation performed. 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.

4.7. 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 method 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.

The method introduced 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 method 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 method 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 $a \to b$ could be descrided 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.

There are other limitations that have more to do with the data itself: conclusions obtained from the UIHC may not hold well with the reality of other hospitals. Since different hospitals are culturally and architecturally different, we could expect that the relative importance of the risk factors might be different between hospitals as well. It could be possible that the CDI colonization pressure is predictive in other hospitals. Also, the purpose of cross validating the trained models, then, cannot be interpreted as proof of predictive power outside of the UIHC. Cross validation serves mostly to assert that the trained models are not overfit. We still note, however, that the statistics regarding CDI cases are consistent with those of other hospitals (Chapter 3).

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

The work presented in this chapter was funded in part by the University of Iowa's eHealth and eNovation Center.

CHAPTER 5
CONCLUSIONS

Healthcare associated infections are a considerable burden to the patients and to the health care system. The affected patients have their prognosis worsened and demand more resources from hospitals, with extended stays and requiring especial precautions, so they do not put other patients at risk. As the bacteria causing these infections are becoming increasingly resistant to antibiotics while also becoming more deadly and contagious, contributing with knowledge for stopping these infections is, therefore, important.

In this thesis, I reported on two projects centered on data collected at the University of Iowa Hospital and Clinics. The first project consisted in analyzing data collected by sensors that reported the location and hand washing behavior of health care workers. In this project, I analyzed radio signals to extract meaning from them; in particular, to determine location information of healthcare workers and determining when they washed their hands. After making sense of this data, I studied two epidemiologically relevant tasks associated to human behavior. In the first, I studied the problem of link prediction for the inference of contact networks, with the aim of predicting when a working will come in contact (close proximity) to others based on whom they are usually in contact with. This information can be used to build and simulate contact networks, whcih can be used to study the spread of infections in the hospital. In fact, there is a whole body of research called *contact network epidemiology*, which consists in studying epidemics in networks, including how to stop them. The other problem I addressed with this data was the study of associations between social pressure and hand washing. By doing so, I found that workers in proximity to others wash their hands more, but also that not all workers are as influential. Both of these findings are, perhaps, somewhat intuitive or expected, but they have implications for design of guidelines focused on preventing the spread of diseases.

In the second project, I developed a data mining method for analyzing medical records aimed at tackling the problems of class imbalance and high dimensionality, and applied it to predicting Clostridium Difficile infection (CDI). The developed method consisted in

building an ensemble of logistic regression models that, far different from existing developments, perform feature selection at the individual classifier level as opposed to the ensemble level. The models resulting from this methodology performed better than the state of the art models. They also improved prediction as the onset of symptoms approached, i.e., produced dynamically changing risk curves of the development of CDI. The main contribution, however, was in the information discovered: certain events in certain orders increased the risk of developing the infection, suggesting that reversing these orders could improve prognosis.

The overall line connecting my work has been the application of computational methods to epidemiological problems (i.e., *computational epidemiology*). So far, the most related discipline is biostatistics, which is concerned on analyzing data medical and public health data. So, there is room for the use of computational methods for analyzing such data, especially if they involve data that does not follow the usual statistical assumptions, such as radio signals, that follow the usual representations, such as dynamic networks and temporarily organized events (and partial orders), or falls in the big size or high dimensionality domain. Thus, perhaps, there is need for more research in *computational epidemiology* for the improvement of the quality of healthcare provision.

BIBLIOGRAPHY

[1] C. Abou, J. Pepin, and L. Valiquette. Prediction tools for unfavourable outcomes in clostridium difficile infection: A systematic review. *PLoS ONE*, 7(1):e30258.

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

[3] C. AS. Smartrf cc2420 preliminary datasheet (rev 1.2), June 2004.

[4] M. Barnett, N. Christakis, J. O'Malley, J. Onnela, N. Keating, L, and B. on. Physician patient-sharing networks and the cost and intensity of care in us hospitals. *Medical Care*, 50(2):152–60, 2012.

[5] M. Barnett, L, B. on, A. O'Malley, N. Keating, and N. Christakis. Mapping physician networks with self-reported and administrative data. *Health Services Research*, 46(5):1592–609, 2011.

[6] I. Batal and M. Hauskrecht. Mining clinical data using minimal predictive rules. In *AMIA 2010, American Medical Informatics Association Annual Symposium.*

[7] I. Batal, H. Valizadegan, G. Cooper, and M. Hauskrecht. A pattern mining approach for classifying multivariate temporal data. In *2011 IEEE International Conference on Bioinformatics and Biomedicine*, BIBM '11, 2011.

[8] I. Batal, H. Valizadegan, G. Cooper, and M. Hauskrecht. A temporal pattern mining approach for classifying electronic health record data. *ACM Transactions on Intelligent Systems and Technology*, 4(4), 2012.

[9] K. Benkic, M. Malajner, P. Planinsij, and Z. Cucej. Using rssi value for distance estimation in wireless sensor networks based on zigbee. *Proceedings of the 15th Conference on Systems, Signals and Image Processing (IWSSIP 2008). Jun 25-28, Bratislava, Slovak Republic. 2008.*

[10] A. Brahim and M. Limam. Robust ensemble feature selection for high dimensional data sets. In *2013 International Conference of High Performance Computing and Simulation (HPCS)*.

[11] Centers for Disease Control and Prevention (CDC). Vital signs: preventing clostridium difficile infections. *MMWR Morbidity and Mortality Weekly Report*, 61(9):157–62, 2012.

[12] B. Chandra and M. Gupta. An efficient statistical feature selection approach for classification of gene expression data. *Journal of Biomedical Informatics*, 44(2011):529–535.

[13] V. Chang and K. Nelson. The role of physical proximity in nosocomial diarrhea. *Clinical Infectious Diseases*, 31(3):717–22, 2000.

[14] J. Chen and R. Altman. Automated physician order recommendations and outcome predictions by data-mining electronic medical records. In *AMIA Joint Summits on Translational Science Proceedings*, 2014.

[15] X. Chen and M. Wasikowski. Fast: A roc-based feature selection metric for small samples and imbalanced data classification problems. In *The 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2008*.

[16] D. Curtis, C. Hlady, G. Kanade, S. Pemmaraju, P. Polgreen, and A. Segre. Healthcare worker contact networks and the prevention of hospital-acquired infections. *PLoS One*, 8(12):e79906, 2013.

[17] D. Curtis, C. Hlady, S. Pemmaraju, P. Polgreen, and A. Segre. Modeling and estimating the spatial distribution of healthcare workers. In *Proceedings of the 1st ACM International Health Informatics Symposium*, IHI '10, pages 287–296, 2010.

[18] D. Curtis, G. Kanade, S. Pemmaraju, P. Polgreen, and A. Segre. Analysis of hospital health-care worker contact networks. In *5th UK Social Networks Conference*, 2009.

[19] M. Cusumano-Towner, D. Li, S. Tuo, G. Krishnan, and D. Maslove. A social network of hospital acquired infection built from electronic medical record data. *Journal of the American Medical Informatics Association*, 20(3):427–34, 2013.

[20] Department of Pathology. Clostridium difficile toxin detection by pcr. *In Laboratory Bulletin, Laboratory Services Handbook (Web). Archived Bulletins 2009.*

[21] Department of Pathology. Delay in clostridium difficile results. *In Laboratory Bulletin, Laboratory Services Handbook (Web). Archived Bulletins 2009.*

[22] Department of Pathology. Frequency of clostridium difficile pcr testing. *In Laboratory Bulletin, Laboratory Services Handbook (Web). Archived Bulletins 2012.*

[23] D. Drekonja, M. Butler, R. MacDonald, D. Bliss, G. Filice, T. Rector, and T. Wilt. Comparative effectiveness of clostridium difficile treatments: a systematic review. *Annals of Internal Medicine*, 155(12):839–47, 2011.

[24] E. Dubberke, K. Reske, M. Olsen, K. McMullen, J. Mayfield, L. McDonald, and V. Fraser. Evaluation of clostridium difficile-associated disease pressure as a risk factor for c difficile-associated disease. *Archives of Internal Medicine*, 167(10):1092–1097, 2007.

[25] E. Dubberke, Y. Yan, K. Reske, A. Butler, J. Doherty, V. Pham, and V. Fraser. Development and validation of a clostridium difficile infection risk prediction model. *Infection Control and Hospital Epidemiology*, 32(4):360–6, 2011.

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

[27] D. Eyre, M. Cule, D. Wilson, D. Griffiths, A. Vaughan, L. O'Connor, C. Ip, T. Golubchik, E. Batty, J. Finney, D. Wyllie, X. Didelot, P. Piazza, R. Bowden, K. Dingle, R. Harding, D. Crook, M. Wilcox, T. Peto, and A. Walker. Diverse sources of c.difficile infection identified on whole-genome sequencing. *New England Journal of Medicine*, 369(13):1195–205, 2013.

[28] D. Eyre, M. Wilcox, and A. Walker. Diverse sources of c. difficile infection. *New England Journal of Medicine*, 370(2):183–4, 2014.

[29] J. Fan, S. Upadhye, and A. Worster. Understanding receiver operating characteristic (roc) curves. *Canadian Journal of Emergency Medicine*, 8(1):19–20, 2006.

[30] Y. Fan and C. Tang. Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society, series B*, 76(3):531–552, 2012.

[31] K. Fong, C. Fatica, G. Hall, G. Procop, S. Schindler, S. Gordon, and T. Fraser. Impact of pcr testing for clostridium difficile on incident rates and potential on public reporting: is the playing field level? *Infection Control and Hospital Epidemiology*, 32(9):932–3, 2011.

[32] A. Friggeri, G. Chelius, E. Fleury, A. Fraboulet, F. Mentré, , and J. Lucet. Reconstructing social interactions using an unreliable wireless sensor network. *Computer Communications*, 34(5):609–618, 2011.

[33] S. Fujitani, W. George, and A. Murthy. Comparison of clinical severity score indices for clostridium difficile infection. *Infection Control and Hospital Epidemiology*, 32(3):220–8, 2011.

[34] M. Galar, Fern, A. ez, E. Barrenechea, H. Bustince, and F. Herrera. A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions of Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 42(4):463–484, 2012.

[35] K. Garey, T. Dao-Tran, Z. Jiang, M. Price, L. Gentry, and H. Dupont. A clinical risk index for clostridium difficile infection in hospitalised patients receiving broad-spectrum antibiotics. *Journal of Hospital Infection*, 70(2):142–7, 2008.

[36] L. Gauvin, A. Panisson, C. Cattuto, and A. Barrat. Activity clocks: spreading dynamics on temporal networks of human contact. *Scientific Reports*, 3:3099, 2013.

[37] A. Geva, S. Wright, L. Baldini, J. Smallcomb, C. Safran, and J. Gray. Spread of methicillin-resistant staphylococcus aureus in a large tertiary nicu: network analysis. *Pediatrics*, 128(5):e1173–80, 2011.

[38] M. Ghalwash, V. Radosavljevic, and Z. Obradovic. Extraction of interpretable multivariate patterns for early diagnostics. In *2013 IEEE International Conference on Data Mining*, ICDM '13, 2013.

[39] S. Goldenberg. Public reporting of clostridium difficile and improvements in diagnostic tests. *Infection Control and Hospital Epidemiology*, 32(12):1231–2, 2011.

[40] D. Gotz, F. Wang, and A. Perer. A methodology for interactive mining and visual analysis of clinical event patterns using electronic health record data. *Journal of Biomedical Informatics*, 48(2014):148–159.

[41] D. Guerrero, J. Becker, E. Eckstein, S. Kundrapu, Deshp, A. e, A. Sethi, and C. Donskey. Asymptomatic carriage of toxigenic clostridium difficile by hospitalized patients. *Journal of Hospital Infection*, 85(2):155–158, 2013.

[42] Y. Halpern, Y. Choi, S. Horng, and D. Sontag. Using anchors to estimate clinical state without labeled data. In *AMIA 2014, American Medical Informatics Association Annual Symposium*.

[43] 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*, 38(3):173–81, 2010.

[44] D. Hanauer and N. Ramakrishnan. Modeling temporal relationships in large scale clinical associations. *Journal of the American Medical Informatics Association*, 20(2):332–41, 2013.

[45] P. Hartigan. Contact networks and the study of contagion. *Biometrics*, 36(3):473–85, 1980.

[46] H. He and E. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009.

[47] R. Henriques, S. Pina, and C. Antunes. Temporal mining of integrated healthcare data: Methods, revealings and implications. In *SIAM SDM International Workshop on Data Mining for Medicine and Healthcare*, DMMH '13, 2013.

[48] T. Hornbeck, D. Naylor, A. Segre, G. Thomas, T. Herman, and P. Polgreen. Using sensor networks to study the effect of peripatetic healthcare workers on the spread of hospital-associated infections. *Journal of Infectious Diseases*, 206(10):1549–57, 2012.

[49] K. Hwang, I. Lee, J. Park, T. Hambuch, Y. Choe, and et al. Reducing false-positive incidental findings with ensemble genotyping and logistic regression based variant filtering methods. *Human Mutation*, 35(8):936–944, 2014.

[50] L. Isella, M. Romano, A. Barrat, C. Cattuto, V. Colizza, den Van, F. Gesualdo, P, E. olfi, L. Ravá, C. Rizzo, and A. Tozzi. Close encounters in a pediatric ward: measuring face-to-face proximity and mixing patterns with wearable sensors. *PLoS One*, 6(2):e17144, 2011.

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

[52] 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*, 51(8):2883–7, 2007.

[53] A. Kaltsas, M. Simon, L. Unruh, C. Son, D. Wroblewski, K. Musser, K. Sepkowitz, N. Babady, and M. Kamboj. Clinical and laboratory characteristics of clostridium difficile infection in patients with discordant diagnostic test results. *Journal of Clinical Microbiology*, 50(4):1303–7, 2012.

[54] M. Kamboj, N. Babady, J. Marsh, J. Schlackman, C. Son, J. Sun, J. Eagan, Y. Tang, and K. Sepkowitz. Estimating risk of c. difficile transmission from pcr positive but cytotoxin negative cases. *PLoS One*, 9(2):e88262, 2014.

[55] M. Kazandjieva, J. Lee, M. Salathé, M. Feldman, J. Jones, and P. Levis. Experiences in measuring a human contact network for epidemiology research. In *The Sixth Workshop on Hot Topics in Embedded Networked Sensors*, HotEmNets '10, 2010.

[56] M. Keeling and K. Eames. Networks and epidemic models. *Journal ofRoyal Society Interface*, 2(4):295–307, 2005.

[57] K. Kipli and A. Kouzani. Degree of contribution (doc) feature selection algorithm for structural brain mri volumetric features in depression detection. *International Journal of Computer Assisted Radiology and Surgery*, 10(7):1003–16, 2015.

[58] C. Landelle, M. Verachten, Legr, P, E. Girou, F. Barbut, and C. Buisson. Contamination of healthcare workers' hands with clostridium difficile spores after caring for patients with c.difficile infection. *Infection Control and Hospital Epidemiology*, 35(1):10–5, 2014.

[59] B. Landon, N. Keating, M. Barnett, J. Onnela, S. Paul, A. O'Malley, T. Keegan, and N. Christakis. Variation in patient-sharing networks of physicians across the united states. *JAMA*, 308(3):265–273, 2012.

[60] B. Landon, J. Onnela, N. Keating, M. Barnett, S. Paul, A. O'Malley, T. Keegan, and N. Christakis. Using administrative data to identify naturally occurring networks of physicians. *Medical Care*, 51(8):715–21, 2013.

[61] C. Lanzas, E. Dubberke, Z. Lu, K. Reske, and Y. Gröhn. Epidemiological model for clostridium difficile transmission in healthcare settings. *Infection Control and Hospital Epidemiology*, 32(6):553–561, 2011.

[62] N. Lee, A. Laine, H. Hu, F. Wang, J. Sun, and 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.

[63] S. Leekha, K. Aronhalt, L. Sloan, R. Patel, and R. Orenstein. Asymptomatic clostridium difficile colonization in a tertiary care hospital: admission prevalence and risk factors. *American Journal of Infection Control*, 41(5):390–3, 2013.

[64] D. Levy, A. Stergachis, L. McFarland, V. Van, D. Graham, E. Johnson, B. Park, D. Shatin, J. Clouse, and G. Elmer. Antibiotics and clostridium difficile diarrhea in the ambulatory care setting. *Clinical Therapeutics*, 22(1):91–102, 2000.

[65] J. Lewis, L. Van Der Maaten, and V. de Sa. A behavioral investigation of dimensionality reduction. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, pages 671–676, 2012.

[66] D. Li, G. Simon, C. Chute, and J. Pathak. Using association rule mining for phenotype extraction from electronic health records. In *AMIA Joint Summits on Translational Science Proceedings*, 2013.

[67] J. Li, A. Fu, H. He, J. Chen, H. Jin, and et al. Mining risk patterns in medical data. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, KDD 2005*.

[68] N. Lim, H. Ahn, H. Moon, and J. Chen. Classification of high-dimensional data with ensemble of logistic regression models. *Journal of Biopharmaceutical Statistics*, 20(1):160–71, 2010.

[69] C. Lin, H. Canhao, T. Miller, D. Dligach, and et al. Feature engineering and selection for rheumatoid arthritis disease activity classification using electronic medical records. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012*.

[70] C. Lin, E. Karlson, H. Canhao, T. Miller, D. Dligach, and et al. Automatic prediction of rheumatoid arthritis disease activity from the electronic medical records. *PLoS ONE*, 8(8):e69932, 2013.

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

[72] L. Lü and T. Zhou. Role of weak ties in link prediction of complex networks. In *Proceedings of the 1st ACM international workshop on Complex networks meet information & knowledge management (CNIKM '09). ACM, New York, NY, USA, 2009, 55-58.*

[73] L. Lü and T. Zhou. Link prediction in complex networks: a survey. *Physica A: Statistical Mechanics and its Applications*, 390(6):1150–1170, 2011.

[74] J. Lucet, C. Laouenan, G. Chelius, N. Veziris, D. Lepelletier, A. Friggeri, D. Abiteboul, E. Bouvet, F. Mentre, and E. Fleury. Electronic sensors for assessing interactions between healthcare workers and patients under airborne precautions. *PLoS One*, 7(5):e37893, 2012.

[75] A. Machens, F. Gesualdo, C. Rizzo, A. Tozzi, A. Barrat, and C. Cattuto. An infectious disease model on empirical networks of human contact: bridging the gap between dynamic network data and contact matrices. *BMC Infectious Diseases*, 13:185, 2013.

[76] O. Mahmoud, A. Harrison, A. Perperoglou, A. Gul, Z. Khan, and et al. A feature selection method for classification within functional genomics experiments based on the proportional overlapping score. *BMC Bioinformatics*, 15(1):274, 2014.

[77] F. Manian, S. Griesnauer, and A. Bryant. Implementation of hospital-wide enhanced terminal cleaning of targeted patient rooms and its impact on endemic clostridium difficile infection rates. *American Journal of Infection Control*, 41(6):537–541, 2013.

[78] L. Meyers. Contact network epidemiology: Bond percolation applied to infectious disease prediction and control. *Bulletin of the American Mathematical Society*, 44:63–86, 2007.

[79] L. Meyers, M. Newman, M. Martin, and S. Schrag. Applying network theory to epidemics: control measures for mycoplasma pneumoniae outbreaks. *Emerging Infectious Diseases*, 9(2):204–10, 2003.

[80] M. Monsalve. The explanatory power of relations and an application to an economic network. In R. Menezes, A. Evsukoff, and M. C. Gonzalez, editors, *Complex Networks*, volume 424 of *Studies in Computational Intelligence*, pages 225–236. Springer Berlin Heidelberg, 2013.

[81] M. Monsalve, T. Herman, S. Pemmaraju, P. Polgreen, and G. Thomas. Inferring realistic intra-hospital contact networks using link prediction and computer logins. *SocialCom 2012, Amsterdam, The Netherlands, Aug 2012*.

[82] M. Monsalve, S. Pemmaraju, and P. Polgreen. Interactions in an intensive care unit: Experiences pre-processing sensor network data. In *Proceedings of the 4th Conference on Wireless Health*, WH '13, pages 5:1–5:8. ACM, 2013.

[83] M. Monsalve, S. Pemmaraju, G. Thomas, T. Herman, A. Segre, and P. Polgreen. Do peer effects improve hand hygiene adherence among healthcare workers? *Infection Control and Hospital Epidemiology*, 35:1277–1285, 2014.

[84] M. Monsalve, S. Tolentino, S. Pemmaraju, and P. Polgreen. Can we identify 'bellwether' states with respect to syphilis incidence? In *Proceedings of the 10th Annual Conference of the International Society for Disease Surveillance (ISDS 2011)*, volume 4, page 11702. Emerging Health Threats, 2011.

[85] R. Moskovitch and Y. Shahar. Medical temporal-knowledge discovery via temporal abstraction. In *AMIA 2009, American Medical Informatics Association Annual Symposium*, volume 2009, page 452, 2009.

[86] R. Moskovitch, C. Walsh, G. Hripcsak, and N. Tatonetti. Prediction of biomedical events via time intervals mining. In *ACM KDD on Workshop on Connected Health at Big Data Era*.

[87] R. Nelson, P. Kelsey, H. Leeman, N. Meardon, H. Patel, K. Paul, R. Rees, B. Taylor, E. Wood, and R. Malakun. Antibiotic treatment for clostridium difficile-associated diarrhea in adults. *Cochrane Database of Systematic Reviews*, 9:CD004610, 2011.

[88] M. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.

[89] NHS Choices (UK). Complications of clostridium difficile infection.

[90] N. Obuchowski. Receiver operating characteristic curves and their use in radiology. *Radiology*, 229:3–8, 2003.

[91] E. Otete, A. Ahankari, H. Jones, K. Bolton, C. Jordan, T. Boswell, M. Wilcox, N. Ferguson, C. Beck, and R. Puleston. Parameters for the mathematical modelling of clostridium difficile acquisition and transmission: a systematic review. *PloS One*, 8(12):e84224, 2013.

[92] A. Otten, R. Reid-Smith, A. Fazil, and J. Weese. Disease transmission model for community-associated clostridium difficile infection. *Epidemiology and Infection*, 138(6):907–14, 2010.

[93] D. Patnaik, P. Butler, N. Ramakrishan, L. Parida, and 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*, KDD '11, pages 360–368, 2011.

[94] D. Patnaik, N. Ramakrishnan, L. Parida, B. Keller, and D. Hanauer. Mining significant partial order patterns in electronic medical records (poster). In *AMIA 2011, American Medical Informatics Association Annual Symposium*.

[95] C. Paxton, A. Niculescu-Mizil, and S. Saria. Developing predictive models using electronic medical records: Challenges and pitfalls. In *AMIA 2013, American Medical Informatics Association Annual Symposium.*

[96] Y. Peng, Z. Wu, and J. Jiang. A novel feature selection approach for biomedical data classification. *Journal of Biomedical Informatics*, 43:15–23, 2010.

[97] L. Peterson, M. Mehta, P. Patel, D. Hacek, M. Harazin, P. Nagwekar, R. Thomson, and A. Robicsek. Laboratory testing for clostridium difficile infection: light at the end of the tunnel. *American Journal of Clinical Pathology*, 136(3):372–80, 2011.

[98] L. Peterson and A. Robicsek. Does my patient have clostridium difficile infection? *Annals of Internal Medicine*, 151(3):176–9, 2009.

[99] T. Planche, K. Davies, P. Coen, J. Finney, I. Monahan, K. Morris, L. O'Connor, S. Oakley, C. Pope, M. Wren, N. Shetty, D. Crook, and M. Wilcox. Differences in outcome according to clostridium difficile testing method: a prospective multicentre diagnostic validation study of c difficile infection. *The Lancet Infectious Diseases*, 13(11):936–45, 2013.

[100] J. Polastre, R. Szewczyk, and D. Culler. Telos: Enabling ultra-low power wireless research. *Proceedings of the 4th international symposium on Information processing in sensor networks (IPSN '05). Piscataway, NJ, USA. 2005.*

[101] P. Polgreen, C. Hlady, M. Severson, A. Segre, and T. Herman. Method for automated monitoring of hand hygiene adherence without radio-frequency identification. *Infection Control and Hospital Epidemiology*, 31(12):1294–1297, 2010.

[102] P. Polgreen, T. Tassier, S. Pemmaraju, and A. Segre. Prioritizing healthcare worker vaccinations on the basis of social network analysis. *Infection Control and Hospital Epidemiology*, 31(9):893–900, 2010.

[103] M. Riggs, A. Sethi, T. Zabarsky, E. Eckstein, R. Jump, and C. Donskey. Asymptomatic carriers are a potential source for transmission of epidemic and nonepidemic clostridium difficile strains among long-term care facility residents. *Clinical Infectious Diseases*, 45(8):992–8, 2007.

[104] D. Schwartz, R. Evans, B. Camins, Y. Khan, J. Lloyd, N. Shehab, and K. Stevenson. Deriving measures of intensive care unit antimicrobial use from computerized pharmacy data: methods, validation, and overcoming barriers. *Infection Control and Hospital Epidemiology*, 32(5):472–80, 2011.

[105] G. Serpen and S. Pathical. Classification in high-dimensional feature spaces: Random subsample ensemble. In *International Conference on Machine Learning and Applications, 2009, ICMLA '09.*

[106] J. Shankar, S. Szpakowski, N. Solis, S. Mounaud, and et al. A systematic evaluation of high-dimensional, ensemble-based regression for exploring large model spaces in microbiome analyses. *BMC Bioinformatics*, 16(31), 2015.

[107] M. Shapiro and E. Delgado-Eckert. Finding the probability of infection in an sir network is np-hard. *Mathematical Biosciences*, 240(2):77–84, 2012.

[108] M. Shaughnessy, R. Micielli, D. DePestel, J. Arndt, C. Strachan, K. Welch, and C. Chenoweth. Evaluation of hospital room assignment and acquisition of clostridium difficile infection. *Infection Control and Hospital Epidemiology*, 32(03):201–206, 2011.

[109] M. Siemann, M. Koch-Dörfler, and G. Rabenhorst. Clostridium difficile-associated diseases: The clinical courses of 18 fatal cases. *Intensive Care Medicine*, 26(4):416–21, 2000.

[110] J. Starr and A. Campbell. Mathematical modeling of clostridium difficile infection. *Clinical Microbiology and Infection*, 7(8):432–7, 2001.

[111] J. Starr, A. Campbell, E. Renshaw, I. Poxton, and G. Gibson. Spatio-temporal stochastic modelling of clostridium difficile. *Journal of Hospital Infection*, 71(1):49–56, 2009.

[112] N. Stenardo and A. Kalousis. Relationship-aware sequential pattern mining: results on medical practise on antibiotic treatment and resistance development. In *The 29th International Conference on Machine Learning (ICML 2012)*.

[113] W. Su, J. Mercer, H. Van, and M. Maley. Clostridium difficile testing: have we got it right? *Journal of Clinical Microbiology*, 51(1):377–8, 2013.

[114] N. Sundaravaradan, N. Ramakrishnan, and D. Hanauer. Factorizing event sequences. *IEEE Computer*, 45(12):73–75, 2012.

[115] G. Thomas, P. Polgreen, T. Herman, D. Sharma, B. Johns, H. Chen, G. Scranton, D. Naylor, M. Ireland, T. McCarty, et al. Improving patient safety with hand hygiene compliance monitoring. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 55, pages 823–827. SAGE Publications, 2011.

[116] L. Tolosi and T. Lengauer. Classification with correlated features: unreliability of feature ranking and solutions. *Bioinformatics*, 27(14):1986–1994, 2011.

[117] T. Tran, W. Luo, D. Phung, S. Gupta, and et al. A framework for feature extraction from hospital medical data with applications in risk prediction. *BMC Bioinformatics*, 15:425, 2014.

[118] T. Ueno and N. Masuda. Controlling nosocomial infection based on structure of hospital social networks. *Journal of Theoretical Biology*, 254(3):655–66, 2008.

[119] E. van Kleef, J. Robotham, M. Jit, S. Deeny, and W. Edmunds. Modelling the transmission of healthcare associated infections: a systematic review. *BMC Infectious Diseases*, 13:294, 2013.

[120] P. Vanhems, A. Barrat, C. Cattuto, J. Pinton, N. Khanafer, C. Régis, B. Kim, B. Comte, and N. Voirin. Estimating potential infection transmission routes in hospital wards using wearable proximity sensors. *PLoS One*, 8(9):e73970, 2013.

[121] A. Walker, D. Eyre, D. Wyllie, K. Dingle, R. Harding, L. O'Connor, D. Griffiths, A. Vaughan, J. Finney, M. Wilcox, D. Crook, and T. Peto. Characterisation of clostridium difficile hospital ward-based transmission using extensive epidemiological data and molecular typing. *PLoS Medicine*, 9(2):e1001172, 2012.

[122] S. Wang, X. Chen, J. Huang, and S. Feng. Scalable subspace logistic regression models for high dimensional data. In *Proceedings of the 14th Asia-Pacific International Conference on Web Technologies and Applications*, APWeb'12, pages 685–694, 2012.

[123] J. Wiens, W. Campbell, E. Franklin, J. Guttag, and E. Horvitz. Learning data-driven patient risk stratification models for clostridium difficile. *Open Forum Infectious Diseases Advance Access, June 2014*.

[124] J. Wiens, J. Guttag, and E. Horvitz. Learning evolving patient risk processes for c. diff colonization. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012*.

[125] J. Wiens, J. Guttag, and E. Horvitz. Patient risk stratification for hospital-associated c. diff as a time-series classification task. In *Advances in Neural Information Processing Systems 25 (NIPS 2012)*.

[126] A. Wright, A. Wright, A. McCoy, and D. Sittig. The use of sequential pattern mining to predict next prescribed medications. *Journal of Biomedical Informatics*, 53:73–80, 2015.

[127] Y. Xu, K. Hong, J. Tsujii, and E. Chang. Feature engineering combined with machine learning and rule-based methods for structured information extraction from narrative clinical discharge summaries. *Journal of the American Medical Informatics Association*, 19(5):824–32, 2012.

[128] L. Yakob, T. Riley, D. Paterson, and A. Clements. Clostridium difficile exposure as an insidious source of infection in healthcare settings: an epidemiological model. *BMC Infectious Diseases*, 13(1):376, 2013.

[129] P. Yang, W. Liu, B. Zhou, S. Chawla, and A. Zomaya. Ensemble-based wrapper methods for feature selection and class imbalance learning. In *Advances in Knowledge Discovery and Data Mining*, pages 544–555. Springer, 2013.

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

[131] X. Zeng and G. Li. Supervised redundant feature detection for tumor classification. *BMC medical genomics*, 7(Suppl 2):S5, 2014.

[132] Z. Zhao and H. Liu. Searching for interacting features. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, pages 1156–1161, 2007.

[133] J. Zhou, Z. Lu, J. Sun, L. Yuan, F. Wang, and J. Ye. Feafiner: Biomarker identification from medical data through feature generalization and selection. In *The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2013*.