Modelling Hospital Acquired *Clostridium difficile* Infections and its Transmission in Acute Hospital Settings

by

Biao Wang

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirement for the degree of

Doctor of Philosophy

in

Management Sciences

Waterloo, Ontario, Canada, 2015

©Biao Wang 2015

Author’s declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

The thesis explored a number of fundamental issues regarding the development of predictive models for hospital acquired *Clostridium difficile* infection (HA CDI) and its outbreaks. As predictive modeling for hospital acquired infection is still an emerging field and the ability to analyse HA CDI and potential outbreaks are in a developmental stage, the research documented in this thesis is exploratory and preliminary.

Predictive modeling for the outbreak of hospital acquired infections can be considered at two levels: population and individual. We provide a comprehensive review regarding modeling methodology in this field at both population level and individual level.

The transmission of HA CDI is not well understood. An agent based simulation model was built to evaluate the relative importance of the potential sources of *Clostridium difficile* (*C. difficile*) infection in a non-outbreak ward setting in an acute care hospital. The model was calibrated through a two stage procedure which utilized Latin Hypercube Sampling methodology and Genetic Algorithm optimization to capture five different patterns reported in the literature. A number of aspects of the model including housekeeping, hand hygiene compliance, patient turnover, and antibiotic pressure were explored. Based on the modeling results, several prevention policies are recommended.

One widely used tool to better understand the dynamics of infectious disease outbreaks is network epidemiology. We explored the potential of using network statistics for the prediction of the transmission of HA CDIs in the hospital. Two types of dynamic networks were studied: ward level contacts and hospital transfers. An innovative method that combines time series data mining and predictive classification models was introduced for the analysis of these dynamic networks and for the prediction of HA CDI transmission. The results suggest that the network statistics extracted from the dynamic networks are potential predictors for the transmission of HA CDIs.

We explored the potential of using the “multiple modeling methods approach” to predict HA CDI patient at risk by using the data from the information systems in the hospital. A range of machine learning predictive models were utilized to analyse collected data from a hospital. Our results suggest that the multiple modeling methods approach is able to improve prediction performance and to reveal new insights in the data set. We recommend that this approach might be considered for future studies on the predictive model construction and risk factor analysis.

Acknowledgements

I would like to thank my advisor Professor Kenneth N. McKay for providing me the opportunity to conduct research under his supervision. This thesis would not have been possible without his support, guidance, and encouragement. I would like to give my special thanks to Dr. William Ciccotelli. Without his “special lectures”, guidance, and coordination, this thesis would never have existed.

Thank you to my advisory committee members Professor Qi-ming He, Professor Fatih Safa Erenay, and Professor Ali Elkamel for their advice and support during my PhD.

Thank you to Ruth Schertzberg, Josh Mores, Karen Straus, Chris Mitchell, Margie Foster, Barbara Merry, and Rhonda Costello for all their time and assistance and training. I could not have completed this thesis without the help of these wonderful people.

Lastly, thanks to my beloved family.

Ethics clearance

This research was reviewed and approved by the Office of Research Ethics of the University of Waterloo (ORE #: 19156) and the participating hospital (THREB #: 2013-0519).

Table of contents

List of figures

List of tables

Chapter 1. Introduction

This thesis explores several aspects of developing predictive models for the prevention of hospital acquired *Clostridium difficile* infection (HA CDI) outbreaks.

1.1 Hospital acquired *Clostridium difficile* infection and its outbreaks

Hospital acquired infections (HAIs), also known as nosocomial infections (NIs), are infections that patients acquire during the course of receiving healthcare treatment for other conditions. Today, HAIs are one of the most serious patient safety issues in hospitals. Annually, approximately 220,000 cases of HAIs occur in Canadian hospitals, leading to at least 8,000 deaths each year (Zoutman et al., 2003). It is estimated that one out of 10 adult patients and one in 12 children patients will possibly contract an infection while in a Canadian hospital (Gravel et al., 2007). Also, HAIs usually cause patients to have prolonged stays, occupying scarce bed-days and requiring additional diagnostic and therapeutic interventions (Annex, 2013), which add extra economic burden for both patients and hospitals. The direct costs of hospital acquired infections in Canada are estimated to be $1 billion annually. *Clostridium difficile* (*C. difficile*) is one type of bacteria that can cause HAIs. According to the Provincial Infectious Disease Advisory Committee (PIDAC) in Ontario, an HA CDI case is defined as (Annex, 2013):

> *"The symptoms of CDI were not present on admission (i.e., onset of symptoms > 72 hours after admission) or the infection is present at the time of admission but is related to a previous admission to your facility within the last four weeks."*

And the definition of HA CDI outbreak is (Annex, 2013):

"For wards/units with ≥20 beds, three (3) new cases of nosocomial CDI identified on one ward/unit within a seven-day period OR five (5) new cases of nosocomial CDI within a four-week period,

OR

For wards/units with <20 beds, two (2) new cases of nosocomial CDI identified on one ward/unit within a seven-day period OR four (4) new cases of nosocomial CDI within a four-week period,

OR

Facilities that have a facility nosocomial CDI rate that exceeds their annual nosocomial baseline rate for a period of two consecutive months."

HA CDI has become a critical problem in recent years (McCollum & Rodriguez, 2012). For instance, in 2003 the outbreak in Quebec hospitals was estimated to be associated with approximately 2000 deaths of patients during a one year period (Eggertson, 2005). One outbreak in an Ontario hospital between May 2006 and December 2007 led to more than 200 infected patients and 91 of them died (Eyre et al., 2013). This increase of incidence and severity of *C. difficile* also imposes a great cost on the health care system. A recent estimation (McGlone et al., 2012) of the CDI cost shows that hospitals with an incidence of 4.1 CDI cases per 100 000 discharges would incur costs greater than $3.2 million; an incidence of 10.5 would lead to costs greater than $30.6 million.

Therefore, preventing HA CDI along with other HAIs and their outbreaks has become one of the important priorities in hospitals.

1.2 The research problem and the challenges

Detecting infections in advance and preventing them from occurring has been an effective strategy widely adopted by many hospitals in practice (Annex, 2013). The development of predictive models for the

detection purpose is the essential component of this strategy. This thesis concentrates on the development of predictive models for the prediction of HA CDI and its outbreaks.

Intuitively, perfect predictions can be made if we know:

1. how HA CDI is transmitted from one patient to another
    - the population and environmental level information
2. what determines the development of a HA CDI
    - the individual level information

Unfortunately, there are many unknowns about HA CDI regarding these two types of information.

First, the transmission of HA CDI at population level is complicated and not well understood. Patients with infections or who are carriers of pathogens admitted to hospital are the potential sources of infection for other patients and healthcare workers (HCWs). Patients who become infected in the hospital are a further secondary source of infections. Pathogenic organisms may contaminate objects, devices, and materials which subsequently contact susceptible body sites of patients. Crowded conditions within the hospital, transfers of patients from one unit to another, and concentration of patients highly susceptible to infection in one area all contribute to the potential transmission of HA CDIs. As shown in Figure 1.1 (Adapted from Donskey, 2010), approximately one-third of the patients who acquire *C. difficile* colonization develop CDI, whereas the remaining two-thirds become asymptomatic carriers. Both CDI patients and asymptomatic carriers have the ability to transmit the pathogen to others through various contacts. However, it is very difficult to observe the transmission from asymptomatic carriers to others. Moreover, *C. difficile* forms spores to be transmitted (Kachrimanidou & Malisiovas 2011; Missaghi, Valenti, & Owens, 2008), and symptomatic *C. difficile* carriers usually start to shed spores before the onset of their symptoms (Donskey, 2010).

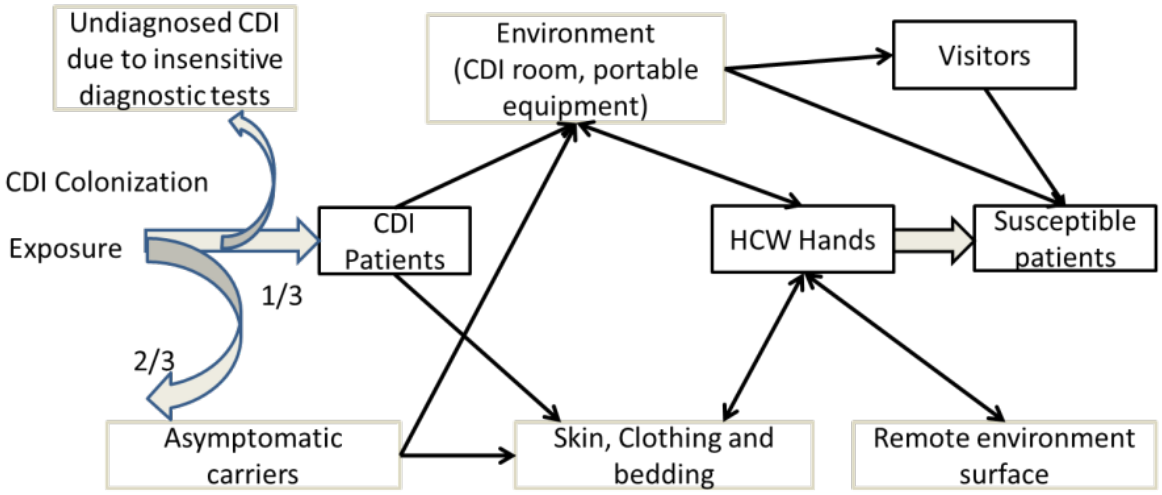


Figure 1.1 Overview of *C. difficile* transmission in hospitals

There is not a consensus about the transmission route of HA CDI. It has been traditionally believed that symptomatic patients are the major source of *C. difficile* spores and the most common transmission mechanisms are through person to person contact, as well as contact with contaminated environment surfaces in hospitals (Donskey, 2010). However, recent results in the literature using advanced genetic analysis have challenged these beliefs. For example, the study from Walker et al. (2012) found that ward contact transmission of *C. difficile* between symptomatic patients accounted for less than 25% of the new case acquisitions in their studied hospitals. The results from Eyre, Griffiths, et al. (2013) further suggested that transmission events from asymptomatic carriers were also likely to be rare, although this result was inconclusive because of some problems of the analysis as pointed out by the authors. Another recent study (Eyre, Cule, et al., 2013) also suggested that diverse sources might exist for the *C. difficile* transmission in hospitals.

Second, the interactions between patient and pathogen that lead to the development of HA CDI are not clear. HAIs are the result of interactions between patients and pathogens. However, each of components, patient and pathogen, is further affected by many factors. For example, patients offer subsistence and lodging for a pathogen and may or may not develop the disease, which is determined by

their immunity against the disease. However, the immunity of a patient against *C. difficile* itself is possibly influenced by two significant risk factors. Studies have shown that two potential major risk factors of affecting the immunity against *C. difficile* are old age and antibiotic therapies that alter the normal gut flora of the patients (Kachrimanidou & Malisiovas, 2011). However, no conclusive result is established at present. There are other key factors as well. For example, an infection will not occur if a pathogen does not invade the patient. Pathogens also have many aspects related to infections (Bennett, Jarvis, & Brachman, 2007; Filetoth, 2008; Merrill, 2012), such as pathogenicity, dose, specificity, infectivity, and etc. Moreover, these properties of a pathogen are not constant, as the pathogen will evolve itself overtime. Studies (Clements, Soares, Tatem, Paterson, & Riley, 2010; Hookman & Barkin, 2009; Missaghi et al., 2008; Riddle & Dubberke, 2009) have shown that the recent increase of incidence and severity of CDI is due to the emerging of the new super-virulent *C. difficile* strain ribotype 027. It is believed that this changing epidemiology has been one of the major reasons responsible for the marked increase of the incidence and severity of HA CDI (de Blank et al., 2013).

1.3 Overview of thesis structure and contributions

As can be seen from the above discussion and the literature review that follows, the ability to analyse HA CDI and potential outbreaks can be considered to be in an early developmental phase of maturity. The predictive power of current methods and approaches does not appear to have a strong foundation and it is the goal of this thesis to explore potential methodology and contribute insights; incrementally contributing at the base level. With this perspective, the research documented in this thesis can be considered exploratory and preliminary in nature. The dissertation is structured as follows.

Chapter 2 and Chapter 3 review the literature. The foundational issues underlying the outbreak of infectious diseases are at two levels: population and individual. At the population level, the focus of modeling is on the characterization of the transmission of infectious diseases in the study population. Chapter 2 reviews the literature regarding the predictive modeling at population level. At the individual

level, the prediction is mainly on the identification patients at risk. Chapter 3 reviews the literature related to the predictive modeling of HA CDI at individual level.

In chapter 4, an agent based simulation model is described that was used to evaluate the importance of the potential sources of the *C. difficile* spores and the impacts of various parameters involved in the transmission of HA CDI. Based on the calibrated baseline model, several key measures such as bacteria age and source of infection were used to understand the roles played by different entities in the transmission of HA CDI. Several aspects of the model including housekeeping, hand hygiene compliance, patient turnover, and antibiotic pressure were further explored.

Chapter 5 explores the potential of using network statistics as predictors for the transmission of HA CDI. Network analysis appears to be increasing in popularity in the analysis of HAI with the growing availability of the relevant data. Two types of dynamic networks were explored: ward level contacts and hospital transfers. An innovative method that combines time series data mining and predictive classification models was introduced for the analysis of these dynamic networks and for the prediction of HA CDI transmission. The results suggest that the network statistics extracted from the dynamic networks might be good predictors for the transmission of HA CDI.

Chapter 6 studies the predictive modeling of HA CDI at individual level. A range of machine learning predictive models were utilized to perform the prediction on the collected data set. Eight algorithms in three categories were analyzed. Each of the methods has strengths and weaknesses with respect to exploiting data patterns and information in the dataset. Using the insights obtained from the eight algorithms and their analysis, a ‘super learner’ model was created that combined the eight models to provide better coverage of the data characteristics. The results suggest that the ‘super learner’ multiple modeling methods approach is able to improve prediction performance and to reveal new insights in data set.

Chapter 7 provides final remarks and discusses several directions for future research.

Chapter 2. Literature review – predictive models at the population level

2.1 Introduction

The modeling of HAIs at the population level started in the early 1990s (van Kleef, Robotham, Jit, Deeny, & Edmunds, 2013), although the modeling of community acquired infections (CAIs) has a history of more than a century (Hethcote, 2000). There are fewer models to be found for HAIs in the literature compared with CAIs, and most of the HAIs modes appear to have been strongly influenced by the CAIs models. Because of this observation, we extended the review to include predictive models for CAIs. However, before we present the results, it is necessary to emphasize two major differences between the HAI population and the CAI population (Meng, 2009), which are:

1) Rapid turnover of the HAI population

   A key difference between HAI and CAI population modeling is the rapid turnover of patients in a hospital population. Patients in hospitals normally just stay in hospitals for a few days or weeks. For example, the average length of stay in Ontario was approximately 6.5 days during 2010-2011 (CIHI, 2012), whereas people in the community at large change very slowly. Therefore, the change of the population is usually not considered by CAIs models when the modeling period is short (months or few years). However, we cannot ignore the population dynamics in HAI settings.

2) Small size of the HAI population

   The size of a HAI population in a hospital is relatively small. The populations studied in HAI models are usually associated with hospital units such as a ward, ICU center or the whole hospital. The patient population size in the HAI models ranged from dozens to a few thousand. Whereas, the populations studied by CAI models are usually towns, cities, nations, continents, or

the entire planet. The population size in the CAI models is at least in the order of tens of thousands, or even billions. Therefore, it is common to observe big fluctuations of infection prevalence in hospitals, and the stochastic chance effects may dominate the transmission dynamics (Keeling & Rohani, 2008).

2.2 Two frameworks of predictive modelling for infectious diseases at the population level

The population level predictive models roughly reside in two distinct frameworks, which we call the "what-if" framework (the analysis of the responses of a model under different scenarios is also considered as prediction) and the "anomaly-detection" framework.

2.2.1 "What-if" framework

The "what-if" framework appears to be a common approach used in the predictive modeling of infectious diseases at the population level. The major focus of the "what-if" approach is on the prediction of how infectious diseases will progress under different plausible parameter settings. Therefore, the studies under the "what-if" framework place significant emphasis on the modeling of the mechanism of the transmission of infectious diseases.

The typical context of the studies under this framework is where a new infectious disease is introduced to a susceptible population (Vynnycky & White, 2010). Predictive models are built to answer a line of questioning similar to:

- Will the new disease lead to an outbreak in the population? If yes, what is the size of the outbreak? What is the most effective prevention policy (e.g. vaccination strategy, hand hygiene compliance, antibiotic stewardship)?

We discuss the models under this framework in section 2.3.

2.2.2 "Anomaly-detection" framework

The "anomaly-detection" framework is the approach taken by syndromic surveillance (SS) in public health (Buckeridge, 2007; Huang et al., 2010; Zhang, Dang, Chen, Thurmond, & Larson, 2009). Over the last two decades, the concern over possible outbreaks due to bioterrorism or the spread of highly virulent viruses (e.g., SARS) has placed increased pressure on public health officials to monitor for abnormal diseases (Chen, Zeng, & Yan, 2010). SS has developed as one of the responses to this concern. The objectives of SS are to recognize illness clusters early, before diagnoses are confirmed, and to activate a rapid response, thus decreasing morbidity and mortality.

Under the "anomaly-detection" approach, it is believed that when an outbreak hits a population, the people's behaviors or symptoms will deviate from their normal routine (Buckeridge, 2007). Therefore, anomalies will emerge and can be used to as signs for the disease outbreak. The core of "anomaly-detection" in SS includes two major components: 1) the selection of representative data, and 2) the definition of anomalies and the corresponding detection algorithms. We discuss some of these methods in 2.4.

2.3 Predictive models under the "what-if" framework

Models under the "what-if" framework focus on the modeling of the mechanisms for the transmission of infectious diseases. Many factors are involved into the transmission of infectious diseases and depending on how these factors are incorporated, the models can be classified into one of the main model categories: compartment model, contact network model and agent based model.

2.3.1 Compartment model

2.3.1.1 Methodology overview of compartment models

Compartment models have been the major, historical method of modeling the transmission of infectious diseases (Hethcote, 2000; Keeling & Rohani, 2008; Vynnycky & White, 2010). In compartment models, individuals are grouped into different compartments according to their infections stages. For example, the simplest compartment model, susceptible— infectious—recovered (SIR) model, splits the population into susceptible, infectious, and recovered compartments. The dynamics of the disease transmission are usually analysed by a system of differential equations, which captures how the numbers in different compartments change. Equations (1) are the corresponding differential equations system for SIR model.

$$\left.\begin{aligned} \frac{dS}{dt} &= \lambda N - \beta SI - \mu S \\ \frac{dI}{dt} &= \beta SI - \gamma I - \mu I \\ \frac{dR}{dt} &= \gamma I - \mu Ro \\ N &= S + I + R \end{aligned}\right\} \quad (1)$$

In these equations, S, I and R are the variables that represent the number of individuals in their corresponding compartments. N is the total population. The parameters included in the model are λ, the birth rate or the arrival rate; μ, the death rate or the departure rate; β, the effective transmission rate; and γ, the recovery rate.

The important result of the compartment model is the basic reproductive ratio R_0. The basic reproduction number of an infectious disease is the number of secondary cases one initiating case generates on average over the course of its infectious period (Keeling & Rohani, 2008). Besides the basic reproductive ratio R_0, other information such as the outbreak size and duration can be also obtained from these equations. However, because nonlinear terms exist in this differential equation system, an analytical solution is hard

to find. Therefore, the solution is often obtained by numerical methods (Keeling & Rohani, 2008; Vynnycky & White, 2010).

2.3.1.2 Applications of compartment models in hospital acquired CDI

Two compartment models for hospital acquired CDI have been built in recent years. Starr et al. (2009) constructed a model in which the patient population in a ward was partitioned into the following classes: immune, susceptible-colonised, susceptible-colonised and toxin positive. Their results suggested patient susceptibility to CDI was the most important factor that determines the spread of the infection. Specifically, the result showed that doubling the rate at which patients become susceptible increased the incidence rate by 63%. Conversely, doubling the environmental load made almost no difference, increasing infection incidence by only 3%. They concluded that reducing patient susceptibility to infection, especially reducing the use of intravenous cephalosporin, was more effective in reducing the number of infection cases than lowering transmission rates.

Lanzas et al. (2011) established a model that split patients into five compartments: 1) resistant to colonization, 2) susceptible to colonization, 3) asymptomatically colonized without protection against CDI, 4) asymptomatically colonized with protection against CDI, and 5) disease. The basic reproduction number obtained from simulation ranged from 0.55 to 1.99 with a median of 1.04. They also concluded that the transmission within the ward alone from patients with CDI cannot sustain new *C. difficile* colonisations. Therefore, the admission of colonized patients played an important role in sustaining the transmission in the ward. Their experiments also suggested that the most influential parameters were the proportion of admitted patients who are asymptomatically colonized and with protection against CDI.

These two models provided good insights into how CDI might develop and could be modeled. However, they both failed to justify why the assumptions made by compartment models (homogeneous contacts and homogeneous conditions, discussed below) held in their cases.

2.3.1.3 Comments on the compartment model

Compartment models have achieved great success in modeling CAIs (e.g. malaria, measles, SARS and HIV) throughout the last century. There have also been attempts at applying the compartment model approach to HAIs since the early 1990s (van Kleef et al., 2013). However, when applying compartment models to HAIs, we have to be aware of the assumptions made by the compartment models in CAIs' context.

One assumption of compartment models is the homogenous contact assumption (mass action principle), which means that individuals from different compartments are fully mixed (Bansal, Grenfell, & Meyers, 2007; Meyers, 2007). In reality, the homogeneous contact assumption does not hold in hospital settings. Depending on their diseases, patients usually have considerable differences in their contacts with their environment. Their chances of contracting infections also vary accordingly. Although some compartment models in the literature take contact heterogeneity into account, many modifications made to the basic framework just further subdivide the S, I and R classification to reflect greater structure within the host population (Keeling & Eames, 2005). However, the assumption of full mixing in each pair of subgroups remains.

Another assumption of the compartment models is the homogenous condition assumption, which is saying that all the individuals in the same compartment have the same condition. For example, all the individuals in a susceptible compartment have the same susceptibility for a disease. However, this assumption might not be appropriate, especially for hospital populations, since there are many factors that impact a patient's susceptibility for infections.

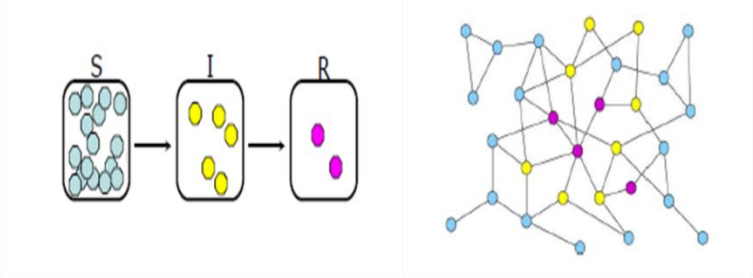
Contact network models and agent based models are the two major direct responses to the assumptions in compartment models. Contact network models explicitly take the contact structure of each individual into account, and deal with contact heterogeneity directly (Aggarwal, 2011; Keeling & Eames, 2005; Newman, 2002). Agent based models allow each agent in the model to have autonomy in order to

incorporate condition heterogeneity (Meng, Davies, Hardy, & Hawkey, 2010). In recent years, a trend of pairing contact network models and agent based models to study HAIs has emerged (Barnes, 2012).

2.3.2 Contact network model

2.3.2.1 Methodology overview of contact network model

Contact (social) network models (CNM) are grounded in two fields: social science and graph theory (Jackson, 2010; Jackson & Watts, 2002; Wasserman & Faust, 1994). These models attempt to model the local structure of contacts by explicitly constructing the links defined by "contact" among individuals. Therefore, it avoids the random-mixing and homogenous contacts assumptions (Meyers, 2007) found in the compartment model. In the network, individuals are usually represented by nodes, and the contacts among them are represented by links. Figure 2.1 (Adapted from Meyers, 2007) is a comparison of the contact assumptions between the compartment and CNMs.



Compartment model in the left assume patients in each compartment will mix randomly and their contacts are homogenous. Network work model in the right models the contact explicitly. The heterogeneity of contacts can be captured.

Figure 2.1 Comparison between compartment model and network model

Unfortunately, it is usually difficult to collect enough data to build a real contact network and a network subset results (Cusumano-Towner, Li, Tuo, Krishnan, & Maslove, 2013; Ueno & Masuda, 2008). At this point, contact network modeling of infectious diseases is still at the theoretic stage and many contact networks studied in the literature are generated artificially with respect to one or more properties. Based

on the nature of infection transmission, different types of networks have been explored. Some ideal networks, such as random networks, lattices, small-world networks, spatial networks, and scale-free networks, are well studied in literature (Newman, 2002). For these random networks, the probability generating function is the major mathematic analytic tool of obtaining the quantities that we are interested in, such as the distribution of outbreak size.

2.3.2.2 Application of contact network model in HAIs

Although a few applications of CNM have been reported for HAIs, we were unable to find any CNMs for *C. difficile*.

The first application of CNM for HAIs we found was the study on Mycoplasma pneumonia conducted by Meyers (Meyers, Newman, Martin, & Schrag, 2003). In their network, patients, caregivers and physical locations were modelled as “vertices”. “Edges” connected people to the wards in which they resided or worked. By manipulating the structure of the network, they simulated several scenarios to evaluate impact of different intervention methods, such as patient isolation or assigning fewer wards to caregivers. The results from their simulation suggested that the assignment of caregivers to patient groups is more critical to the course of an epidemic than the isolation of patients. Their model can also be used to calculate the probability and size of potential outbreaks. One problem of the model is that it did not further distinguish between patients and caregivers. In hospital settings, patients have different immunity to infections depending on their condition. Caregivers also fall into several subgroups that have different contact patterns with patients (Curtis, Kanade, Pemmaraju, Polgreen, & Segre, 2009).

Ueno and Masuda (Ueno & Masuda, 2008) also made a simulation on the transmission of MRSA based infections on the network of a hospital in Tokyo. The network was built from the medical records data collected from the hospital. In their network, they explicitly modeled the contacts between and within different roles in the hospital, which included patient, nurse, medical doctor, and environment. Their

results showed that intervention methods that restricted interaction between medical doctors and their visits to different wards shrunk the final epidemic size more than intervention methods that directly protected patients, such as isolating patients in single rooms. Their study also concluded that vaccinating doctors should be a priority rather than patients or nurses.

Barnes (2012) conducted an experiment using different network structures (determined by the ratio between the number of health care workers - HCWs - and the number of patients) at the ward level to see how the structures within the network affected the transmission rate of HAIs in a ward. The results showed that nurses were responsible for initiating more infections because they typically visited patients more often.

Because the above networks were limited and did not reflect the real contacts within the hospitals, the models were useful for understanding the situation, but were not capable of performing real-time prediction. With the information technology advances in modern hospitals, data about the hospital system has become more accessible and constructing more “realistic” contact networks for a hospital has become potentially feasible.

Curtis et al. (2009) constructed a contact network for healthcare workers in a hospital based on 12 million Electronic Medical Record (EMR) logins collected over 22 months. In their study, they first computed a spatial distribution (the probability that a HCW resided at a specific location in the hospital) for each HCW in the hospital in a time window. Then, contacts were established if two HCWs were encountered at the same location. As pointed out by the authors, this network suffered from systematic biases associated with worker login to the EMR system and the locations of their logins. Moreover, this contact network ignored patients. Therefore, in the subsequent study they enhanced the network by using wireless technology to track the contacts between the HCWs and patients (Herman et al., 2009). The enhanced system could recognize contacts between healthcare workers, and between healthcare workers and patients. Therefore, the latter network developed by combining EMR login data and data captured by

wireless devices was a more realistic estimation of contact patterns in the hospital. Based on wireless technology, they also developed a prototype system for monitoring hand cleaning in hospital clinics. They concluded that wireless sensor network is a promising area for addressing different aspects of nosocomial infection.

Cusumano-Towner et al. (2013) also developed a method to create a contact network of hospital inpatients by using approximately 70-days of EMR data from two hospital information systems (the admission-discharge-transfer system and the clinical documents system). They inferred the patient-to-patient contacts due to shared rooms from admission-discharge transfer data, and contacts with healthcare workers from clinical documents. They further built a probabilistic model and simulated the spread of MRSA on this network as a proof of concept.

2.3.2.3 Comment on contact network model

The above examples demonstrate that hospital information systems have great value when inferring the contact networks in hospitals and could possibly be used in a real time situation. However, we should note that different HAIs have different modes of transmission (e.g., MRSA and *C. difficile*). A universal network for all HAIs has not yet been found. Thus, we need specific data for a specific HAI to form the right links for its network.

2.3.3 Agent based model (ABM)

2.3.3.1 Methodology overview of agent based model

ABM is a class of computational models for simulating the actions and interactions of autonomous agents with a view to assessing their effects on the system as a whole (Railsback & Grimm, 2011). ABM has great flexibility to incorporate heterogeneity of each agent (Railsback & Grimm, 2011). For example, when modeling HAIs by ABM, we can easily associate many risk factors that relate to HAIs. Therefore,

ABMs can overcome the drawback of the homogenous condition assumption in compartment models. In addition, ABMs also have the ability to represent the spatial location and movements of different type of agents. This ability makes ABM a potentially better tool to model the transmission of infection between patients and their environment.

A strong connection exists between contact network models and ABMs. If we allow the nodes in a contact network model to incorporate more information and treat them as an agent, and let the links between the nodes represent the interactions among the agents, the contact network model can be turned into an ABM.

2.3.3.2 Application of ABM in HAIs

The application of ABM for the modeling of HAIs is relatively new. It is believed (Meng, 2009) that the first application was presented by Sébille et al. (1997). In this model, they constructed a hypothetical 10-bed intensive care unit (ICU) with 30 staff members. The impact of hand washing compliance on colonization was then explored. However, ABM has not widely adopted since then. As Meng (2009) pointed out, the second ABM on HAIs appeared in 2005. The second model also configured a hypothetical ICU with 24 beds. The model examined the effects of various factors on the transmission of HAIs, including pathogen transmissibility, duration of caregiver contamination, etc.

Although it is mentioned that ABMs are more likely to make reliable quantitative predictions in real time (Meng, 2009), the majority of the studies are still devoted to evaluating the effectiveness of different prevention polices. For example, Meng used the ABM approach to describe the transmission dynamics and evaluated the intervention policies of HAIs in general and MRSA in particular. Meng justified why ABM is appropriate for HAIs by comparing different simulation principles. Moreover, the research results also included a proposed framework that describes how to apply ABM to HAIs.

In recent years, ABM appears to have become more popular for the modeling of HAIs. For example, Barnes (2012) built several ABMs and network models of the transmission of infectious diseases at scales ranging from hospitals to networks of medical facilities and community populations. In a recent review paper (van Kleef et al., 2013), the authors found that approximately 30% of the publications on the modeling of the transmission of HAIs were agent based models during the period between 2008 and 2011. Approximately 16 agent based models were constructed during that period of time. However, most of the agent based models constructed in the literature were not for specific HAIs but for generic HAIs.

Four agent based simulation models for HA CDI that were developed to evaluate the effectiveness of different prevention policies were found in recent literature. The objective of the first model (Jiménez, Lewis, & Eubank, 2013) was to compare different medical treatments against HA CDI. The purpose of the second model (Rubin et al., 2013) was to assess different strategies to control *C. difficile* transmissions and infections. The author concluded that a "bundled" intervention is likely to reduce HA CDI incidence rates. Similarly, Codella et al. (2014) developed a model to evaluate the combined effects of different control measures in a midsized endemic hospital setting. Their main result was the development of a generalized ABM framework for CDI control. The fourth model was developed by Lanzas and Dubberke (2014) to evaluate the effectiveness of one specific policy—admission screening. Their results showed that admission screening may reduce both the number of new colonization and HA CDI cases. All of the models added understanding and insights about the transmission of the HA CDI.

2.3.3.3 Comments on ABM

However, ABMs also have some drawbacks (Grimm & Railsback, 2005). The major disadvantages of ABM include: 1) the absence of a mathematical formulation and the corresponding loss of analytical power; and 2) there is less parsimony in the model structure compared to compartment models and this results in most ABM models being very demanding with respect to data.

2.4 Predictive models under "anomaly-detection" framework

Predictive models under the "anomaly-detection" framework can be grouped into two categories depending on the data type and the corresponding detection methods. The two categories are temporal and temporal-spatial methods. In general, the detection algorithms are based on statistical hypothesis testing. Various statistics have been developed for this framework, and the literature of these algorithms is relatively abundant. No intention is made here to cover all the algorithms. Several of the commonly encountered methods are selected to demonstrate the basic ideas. For further information about these algorithms, we refer to Chen et al. (2010).

2.4.1 Temporal analysis models

Aggregated data in public health, such as incidence rates, are often presented as time series data. The representative method of detecting anomalies in this type of data is the statistical process control (SPC) chart method (Buckeridge, Burkom, Campbell, Hogan, & Moore, 2005). Within the SPC approach, the cumulative summation (CUMSUM) and exponential weighted moving average (EWMA) methods are widely applied (Hutwagner, Browne, Seeman, & Fleischauer, 2005).

CUMSUM is calculated by taking the summation of the difference between each observation and in-control expectation. It is defined by formula (2).

$$S_r = \sum_{i=1}^{r}(x_i - \mu) \quad (2)$$

Where, S_r is the cumulative sum statistics at time period r; x_i is the individual sample (or sample mean, if the sample size is not one) at time period i; and μ is the estimated mean in-control. In a process that is under control, each measured value x_i should be reasonably close to the mean μ. Thus, a plot of each calculated value of S_r should be centered at zero with small fluctuations up or down. When the process

have changed (even slightly), CUMSUM is able to detect them, as the changes are amplified by the cumulative summation.

Another common approach is EWMA, which is defined by formula (3).

$$EWMA_i = \lambda x_i + (1 - \lambda) EWMA_{i-1} \quad (3)$$

Where $EWMA_0$ is the historical mean; x_i is the individual sample at time period i. A pre-defined threshold is used as the upper limit as in all other SPC control charts. EWMA is also sensitive to the change of the process, as the recursive formula forms a local estimation of expected mean.

Sometimes, pre-processing of data is needed when applying SPC to time series data as the data may contain temporal structures (e.g. trend and seasonality). The usual approach of pre-processing is first fitting a regression model to the time series data, then applying SPC to the regression residuals.

Many other methods that deal with time series data (e.g. ARIMA) can also be applied to the detection of outbreaks. For further information, we refer the reader to Chen et al. (2010).

2.4.2 Temporal-spatial analysis method

While many of the models focus on how infectious disease patients are clustered with respect to time, clustering can also be found in the space dimension. Therefore, detecting outbreak signals using space dimensional data can be fruitful in some situations.

The most widely used method in syndromic surveillance for detecting “hot spots” is the spatial scan statistic (Agarwal, McGregor, Phillips, Venkatasubramanian, & Zhu, 2006). The general goal of the spatial scan statistic is to detect and evaluate the statistical significance of a spatial cluster of events that cannot be explained by spatial randomness, which is defined by a null probability model (e.g., uniform or passion) (Glaz, Pozdnyakov, & Wallenstein, 2009). There are also spatial scan statistics for two, three and more dimensions. If the scanning is done over a three-dimensional area defined by both space and time,

we have a space-time scan statistic, which is an important special case of the three-dimensional spatial scan statistic.

In its original form, the spatial scan statistic consists of a rectangular scanning window with a fixed size and shape (Glaz et al., 2009). This window is continuously moved over the predefined study region, covering all possible locations, and the definition of the spatial scan statistic is the maximum number of points in the scanning window at any given time. The next step is to find the probability of observing at least that many points within the window, under the null hypothesis.

Other methods dealing with temporal-spatial data exist. Several examples of these methods are: what is strange about recent events (WSARE) (Wong, Moore, Cooper, & Wagner, 2003), small area regression and testing (SMART) (Bradley, Rolka, Walker, & Loonsk, 2005), and recursive least square (RLS) (Najmi & Magruder, 2005). However, the same principle—detecting anomalies—applies to these methods.

2.4.3 Application of the "anomalies detection" algorithms in HAI outbreaks

While a considerable number of methods under the "anomaly-detection" framework have been developed for the outbreak detection in large populations, applications of these methods to hospitals has been observed to be relatively few (Carnevale et al., 2011).

Carnevale et al. (2011) evaluated the utility of four algorithms for detecting clonal hospital infection outbreaks. These methods are CUSUM, EWMA, space-time scan statistic (STSS), and WSARE. The data used by the algorithms was the daily case counts for each organism taken from all microbiologic culture data from a university hospital. The results showed that each of the four algorithms generated a list of interesting suspect clusters substantially different from the others. Moreover, the sensitivity values of these algorithms were very low, just ranging from 21% to 31%, which means more than 70% of the real

outbreaks were missed. Simultaneously, the positive predictive values were also low, ranging from 5.3% to 29%, which means the majority of the results were false positive.

2.5 Summary

Models in the “what-if” framework are of three types: compartment model, contact network model, and agent based models. Compartment models can be considered to be the classical approach of modeling the spread of an infection. Its two assumptions, contact homogeneity and condition homogeneity, are restrictive in hospital settings. Network models can overcome the homogenous contact assumption by explicitly modeling the contacts as links in the network. However, this approach demands considerable amounts of data which might be an obstacle in practice. Recently, data recorded by modern information systems has been exploited to establish the hospital contact network. However, specific infectious disease requires specific data to build its corresponding network, as diseases have different transmission modes. Agent based models have the advantage of modeling the heterogeneity condition easily, which might be very important in hospital settings. However, this type of model also demands a great amount of data, as more parameters are used. ABMs also have high computation cost, especially when the number of agents in the model is large.

Under the “anomaly-detection” framework, there are many algorithms and most of the detection algorithms are based on statistical methods. As its goal is to find early sign of outbreaks, the “anomaly-detection” approach places little emphasis on understanding the mechanisms of transmission of the infection diseases. Therefore, these methods often lack interpretability. Depending on the type of data used for the detection, algorithms in “anomaly-detection” framework fall roughly into two groups: temporal data analysis algorithm and temporal-spatial data analysis algorithms. Only a few applications of the algorithms from “anomaly-detection” framework exist for HAIs and the performance of these algorithms is not strong.

Chapter 3. Literature review – predictive models at individual level for hospital acquired *Clostridium difficile* infection

The predictive models at the individual level consider how to identify patients at risk. Finding good risk factors and using proper modeling methods are very important for having a high performance model. We first review the risk factors for HA CDI considered in literature. Then, we present the methods employed by various predictive models for the prediction of HA CDI. In order to have a full view of this field, we also provide a review of the predictive modeling for HAI in general. A full, detailed review can be found in Appendix A. Not surprisingly, most of the literature is coming from medical community.

3.1 Risk factors

The risk of contracting an infection for a patient is determined by the chance of having contacts with the pathogen and by the condition of the patient. Therefore, the risk factors can be divided into two groups: contact factors and patient condition factors.

3.1.1 Contact factors

Contact factors are mainly determined by the transmission routes of infectious pathogens. In HAI prevention practices, transmission routes are one of the most heavily investigated subjects, since knowing the relative importance of the transmission routes is crucial for designing prevention strategies.

Several epidemiologic studies have been conducted for CDIs to investigate the transmission routes. Data used in these studies is diverse and is very detailed and granular. Generally, the data can be classified into four categories: temporal data, spatial data, biological data and patient medical data. Temporal data is usually related to events, such as patient admission/transfer/discharge events, physician and nurse visit events. Spatial data reflects the relative geographic positions of all entities in the hospital. Spatial data

includes hospital layout, wards and beds arrangement, equipment positions etc. Biological data contains the information of the infectious pathogen. Biological data is usually produced by labs in medical institutions, such as genotyping data of the pathogenic organism. Patient medical data includes patient demographic data, diagnostic data, treatment data, and etc. Statistical methods (e.g. visualization, cluster analysis) are the main tools to analyze these data.

Samore et al. (1996) made an early attempt to define the frequency of nosocomial CDI patient-to-patient transmission in an urban tertiary referral hospital. They cultured 99 prospectively identified patient contacts with index cases during a six-month study period. *C. difficile* was found in 31 of the 99 contacts, including 12 with diarrhea and 19 who were asymptomatic. Molecular typing data analysis showed that 5 of 12 from symptomatic contacts matched the corresponding index case, and only 1 of the 19 from asymptomatically colonized contacts matched. They concluded that *C. difficile* did not result from the transmission from the presumed index case. However, in this study it is not clear how patient contacts were defined.

Walker et al. (2012) recently made an investigation on ward-based transmission of *C. difficile* by subdividing outbreaks into distinct lineages defined by multi-locus sequence typing (MLST). In their study, sequence types (STs) were combined with admission and ward movement data. Networks of cases and potential transmission events were constructed for each ST. Potential infection sources for each case and transmission timescales were defined by prior ward-based contact with other cases sharing the same ST. Their results suggested that no more than 25% of cases could be linked to a potential ward-based transmission. The result of this study might be considered impressive. However, Harbarth and Samore (2012) point out that the conclusion of this study may not be well justified due to the bias already mentioned by the authors (e.g., selection, misclassification, and information biases). Other potential limitations that jeopardized validity of the result as further pointed out by Harbarth and Samore include the possibility of inter-ward transmission and the poor sensitivity of the Enzyme Immune-Assay (EIA) testing method for *C. difficile* diagnosis. The poor sensitivity might lead to a significant exclusion of CDI

cases from the sample. Although Walker et al.'s initial research is possibly subject to a number of limitations, Didelot et al. (2012) confirmed the result of Walker et al.'s work by using phylodynamics. This method can estimate the times back to common ancestors of bacterial lineages with sufficient resolution to distinguish whether direct transmission is plausible or not. However, phylodynamics inference is still subject to challenges for the application in complex systems (Frost et al., 2014), and the new study still did not address the problem of the possible significant of exclusion of patients undetected pointed by Harbarth and Samore (2012).

The two studies from Walker et al. and Didelot et al. showed that ward-transmission may not be as important as has previously been claimed. However, this result seems to contradict the observation that CDI outbreaks or clusters usually happen at the ward level (Alfa et al., 2000; Kristjánsson et al., 1994). Moreover, the research did not answer the question which transmission routes are important. Many other possible transmission routes are still waiting to be investigated Several other studies have suggested the possible importance of the transmission routes through environment surfaces (Mutters et al., 2009; Weber, Rutala, Miller, Huslage, & Sickbert-Bennett, 2010), asymptomatic carriers (Riggs et al., 2007) and air (Best, Fawley, Parnell, & Wilcox, 2010). However, it is not yet clear how important these routes are.

The above research suggests that contact heterogeneity appears to exist in the transmission of hospital acquired CDI. However, it is still not clear how variable and how relevant the contacts are for each individual.

3.1.2 Patient condition factors

In HAI prevention practices, the patient condition is also examined. Recent literature shows that the epidemiology of HA CDI has been changing in recent years (Honda & Dubberke, 2014). Relevant

literature was reviewed to have a more complete picture about the risk factors for HA CDI and 27 risk factors are extracted from the literature in PubMed and are shown in Table 3.1.

Table 3.1 Risk factors of CDI from literature

Category	Risk Factors	Articles
Conditions (13)	Severity of underlying diseases (aspecific)	de Blank et al., 2013; Dubberke, Reske, Yan, et al., 2007; Henrich, Krakower, Bitton, & Yokoe, 2009; Kutty et al., 2010
	Prior hospitalization	Dubberke et al., 2007; Henrich et al., 2009; Kutty et al., 2010; Linsky, Gupta, Lawler, Fonda, & Hermos, 2010; McFarland, Clarridge, Beneda, & Raugi, 2007
	Surgery procedures (abdominal/ gastric)	Brown, Talbot, Axelrod, Provencher, & Hoegg, 1990; Johnson et al., 1990; Thibault, Miller, & Gaese, 1991
	comorbidity scores	Dubberke et al., 2007; McFarland et al., 2007; Stevens, Concannon, van Wijngaarden, & McGregor, 2013
	Immunosuppression	Henrich et al., 2009
	Chemotherapy	de Blank et al., 2013
	ICU	Kutty et al., 2010
	Hospitalization prior surgery	Campbell, Phillips, Stachel, Bosco, & Mehta, 2013
	Rehabilitation stay	Henrich et al., 2009
	Max glucose level >150 mg/dL	Henrich et al., 2009
	Max leukocyte count >20,000/μL	Dubberke et al., 2007; Henrich et al., 2009
	Max creatinine level >2 mg/dL	Henrich et al., 2009
Antibiotics use (6)	Antibiotics (aspecific)	Campbell et al., 2013; Louie & Meddings, 2004; McFarland et al., 2007; Owens, Donskey, Gaynes, Loo, & Muto, 2008; Thibault et al., 1991
	Cephalosporins	Brown et al., 1990; de Blank et al., 2013; Zimmerman, 1991
	Quinolones	Kutty et al., 2010; Yip, Loeb, Salama, Moss, & Olde, 2001
	Penicillin	Kutty et al., 2010; McFarland et al., 2007
	Aminoglycoside	de Blank et al., 2013; McFarland et al., 2007
	Clindamycin	McFarland et al., 2007; Owens et al., 2008
Drug use (4)	Proton Pump Inhibitor	Campbell et al., 2013; de Blank et al., 2013; Dubberke et al., 2007; Linsky et al., 2010; Louie & Meddings, 2004
	Antimotility medications	Dubberke , et al., 2007; Kutty et al., 2010
	H2 blockers	Dubberke et al., 2007; Kutty et al., 2010
	Storied	Leekha, Aronhalt, Sloan, Patel, & Orenstein, 2013
Demographics (2)	Old Age	Brown et al., 1990; Dubberke et al., 2007; Henrich et al., 2009; Linsky et al., 2010
	Gender	de Blank et al., 2013
Others (2)	Long stay in hospital	McFarland, Surawicz, & Stamm, 1990; Thibault et al., 1991; Zimmerman, 1991
	Strain type	Kachrimanidou & Malisiovas, 2011

3.2 Predictive models for HA CDI

Six CDI predictive models were found in the literature. Five of the models predicted the contraction of CDI at time of admission:

1. Garey et al. (2008) developed a risk index model for CDI based on the patient medical records provided by a hospital. They first made a logistic regression model to identify important variables. As a result, age (50-80, >80), haemodialysis, non-surgical admission, ICU length of stay were selected to form the risk index. Two risk indices, simple and odds ratio (OR) index, were formed for their model. The index score were organized into four categories: 1) very low risk; 2) low risk; 3) medium risk; and 4) high risk. The model achieved a reasonable performance; the area under curve (AUC) is 0.73.
2. Tanner et al. (2009) in UK developed a score system to target high risk patients of developing CDI by the similar procedure. Only two variables, waterlow score and proton pump inhibitors were the significant variables, and were included in their model. The AUC of their model is 0.827.
3. Dubberke et al. (2011) built a logistic regression model to predict the patients' risk of getting CDI. They also used medical records provided by a hospital. The variables used in the model are age, CDI pressure, times admitted to hospitals in the previous 60 days, modified Acute Physiology Score, days of treatment with high-risk antibiotics, whether albumin level was low, admission to an intensive care unit, and receipt of laxatives, and gastric acid suppressors or anti-motility drugs. The AUC of the model is 0.88.
4. Cooper et al. (2012) proposed a logistic regression model to identify patients who are at risk of CDI at their admission. They employed six variables in their model, including antibiotic usage, age, admission from other facility, stool history, recent hospital stay within 90 days and prior positive *C. difficile* toxin assay. Their model's AUC is 0.929, a better result compared with the above two.
5. Steele et al. (2012) built a Bayesian Network Model to estimate patient risk for CDI after colon resection surgery and their model's AUC is 0.75. They obtained the data from the Nationwide Inpatient Sample in the US. The variables included in the model have four categories, which were

demographic variables, pre-existing medical conditions, patient disease severity and antibiotic use.

As noted, the five models predicted the risk of patients of contracting CDI at their admission. However, the risk of contracting CDI changes over time during the hospital stay period of a patient. As the patient stays longer in this hospital, the risk is higher (Loo et al., 2011).

A sixth model by Wiens et al. (2012) proposed a two-stage model that predicts the risk of getting CDI for a patient over time (stay in the hospital). At the first stage, they employed SVM to estimate the risk of getting CDI for each day of the hospitalization for each patient. A time series was then formed for each patient for their stay in the hospital. They then evaluated several methods to extract features from the time series of each patient. The extracted features were further fed to a SVM to do the final classification. Several issues with their model are:

1) The risk computed in the paper was actually the risk for *C. difficile* colonization and not for CDI. In practice, the risk for colonization is not a major result as colonization is quite common; approximately 30% (Kachrimanidou & Malisiovas 2011) in hospital settings and most of colonisations will not turn into infections (Kachrimanidou & Malisiovas 2011).
2) Their model was restricted to patients who stayed in the hospital longer than 7 days. Conceptually, it may not be appropriate to apply the model to patients who stay in hospital less than 7 days. However, the average length of stay in hospital is approximately 6.5 days based on Canadian data, which implies that the model might have quite limited application in practice.
3) When training their SVM at the first stage, they did not use real labelled data but generated the data randomly. Data, especially labelled data, is very scarce and expensive to obtain in practice. That is possibly why their model achieved moderate accuracy (AUC: 0.79) even when using approximately 10,000 features.

3.3 Summary

A few observations can be drawn from the review. First, the epidemiology of HA CDI appears to be not well understood. It is possible that more factors than those examined in the literature might have impact on the risk for HA CDI contraction. For the purpose of building predictive models in practice, using different data sources which have an extended coverage of risk factors might be crucial for the prediction performance. Second, literature on the predictive modeling of HA CDI often uses one single method and the number of the studies in the literature is small.

The reviews from Chapter 2 and Chapter 3 show that there is a lack of understanding and modeling (especially, empirical modeling) of HA CDI regarding its transmission and development at both population and individual levels. In the next three chapters, we make an attempt to address some of these perceived gaps by collaborating with a medium sized acute care hospital. In Chapter 4, an agent based simulation model based on a ward in the hospital is built to evaluate the importance of the potential sources of the *C. difficile* spores and the impacts of various parameters involved in the transmission of HA CDI in order to have a better understanding of the transmission of HA CDI. Next, in Chapter 5, we focus on the prediction of HA CDI transmission. We explore the potential of using network statistics as predictors for the transmission of HA CDI by using the real data from the hospital. Chapter 6 provides an empirical study on the predictive modeling of HA CDI at the individual level based on the data from the hospital. A range of machine learning predictive models are utilized to improve the prediction performance.

Chapter 4. Evaluating the relative importance of different *Clostridium difficile* sources in acute care hospital settings with an agent based simulation model

4.1 Introduction

In order to prevent the spread of any infectious disease effectively and efficiently, it is important to understand its reservoirs and the transmission routes. Unfortunately, the transmission of HA CDI is not well understood and recent research has challenged traditional assumptions. It has been traditionally believed that symptomatic patients are the major source of *C. difficile* spores and the most common transmission mechanisms are through person to person contact, as well as contact with contaminated environment surfaces in hospitals (Donskey, 2010). For outbreaks, this implied that cases were related to the transmission of 1 strain/clone at the genetic level. However, recent results in the literature using advanced genetic analysis have challenged these beliefs. For example, the study from Walker et al. (2012) found that ward contact transmission of *C. difficile* between symptomatic patients accounted for less than 25% of the new case acquisitions in their studied hospitals. The results from Eyre et al. (2013) further suggested that transmission events from asymptomatic carriers were also likely to be rare, although this result was inconclusive because of some problems of the analysis as pointed out by the authors. Another recent study (Eyre et al., 2013) also suggested that diverse sources might exist for the *C. difficile* transmission in hospitals. Specifically, 45% of the CDI cases in the studied hospitals were genetically distinct from all the previous cases. Moreover, only 13% of the cases were genetically related to and involved in ward contacts, and 19% of the cases were genetically related to and involved in some sort of hospital contacts.

However, because of the limitations (Eyre, Cule, et al., 2013; Eyre, Griffiths, et al., 2013; Walker et al., 2012) of these studies (including generalizability), the diversity of *C. difficile* sources could be explained via two opposing hypotheses: 1) the diversity is mainly due to the import of the bacteria from outside

(community or other healthcare facility); 2) the diversity is mainly due to the evolution of the bacteria and the new cases are caused by pre-existing bacteria circulating within the hospitals.

This dichotomy and the confusion mentioned above lead to a few questions regarding the transmission of HA CDI:

- What causes the HA CDI: the newly imported bacteria, or the pre-existing bacteria circulating within the hospital? Or both? Are outbreaks by syndromic surveillance always clonal?
- What the roles do different entities, such as environmental objects and healthcare workers (HCWs), play in the transmission of HA CDI?
- How important are these entities for the transmission of HA CDI?

In this chapter, we present an agent based simulation model of a typical hospital ward to explore these questions. The model assumes the spread of the disease is on a genetically related strain.

The rest of the chapter is organized as follows. After a brief review of the related work in Section 4.2, the construction of the simulation model is described in Section 4.3. Then, the model calibration methods and results as well as model uncertainty and sensitivity analysis are presented in Section 4.4 and 4.5 respectively. Then, model exploration based on the calibrated baseline model and the results are described in Section 4.6. This chapter concludes with a discussion about the importance of the various entities that are involved in the transmission of HA CDI and a few recommendations for its prevention.

4.2 Model description

An inpatient ward in the participating hospital was modelled by using NetLogo 5.0.4 (https://ccl.northwestern.edu/netlogo/). In this inpatient ward, we have five private rooms, six semi-private rooms, four four-bed rooms, and a nurse station which also included a storage room. Therefore, 36 beds are available in this ward. Based on the real operation of the studied hospital, the bed occupancy rate

was set to be 100% (Kaier, Luft, Dettenkofer, Kist, & Frank, 2011). The patient-nurse ratio is set to be 9:1 for the night shift and 6:1 for the day shift. There are always four physicians in this ward during the day shift and one physician during the night shift.

The basic operations are as follows. Patients are admitted into the ward as long as there are empty beds in it. Admitted patients stay in the ward for a period of time according to a distribution of patient length of stay (LOS) fitted from the hospital data. During their stay in the ward, patients interact with their environment objects and HCWs. These interactions could lead to the transmission of the *C. difficile* bacteria among these entities and may cause patients to get infected. Eventually, patients are discharged after the period of their LOS, although some patients may have a prolonged stay because of the infection. Newly released beds then can be used to admit new patients. In this sense, the simulation is not a terminal-simulation but a steady state simulation. We have set the warm up period of the simulation to be one year, and the simulated data collection period to be ten years.

In this model, the behaviors of six types of entities that are involved in the transmission of *C. difficile* bacteria are considered, and twelve processes are constructed to model the interactions among these entities that can potentially lead to the transmission of *C. difficile* bacteria. The remaining part of this section describes the entities and the processes in detail, as well as the key performance measures devised for the evaluation of the model for our study purpose.

4.2.1 Entities and their states in the model

In this model, six types of entities are considered. Four entity types are modeled explicitly as agents: patients, HCWs, visitors and environmental objects. These entities are the potential reservoirs (sources) for the *C. difficile* spores.

- **Patients**: patient agents are always in one of three states: non-colonized, asymptomatic (colonized), or symptomatic. Asymptomatic and symptomatic patients are both potential sources

for the *C. difficile* spores. Non-colonized patients will have a probability to become asymptomatic patients if they have contacts with *C. difficile* spore contaminated sources and also have antibiotic exposure. Asymptomatic patients will have a probability to become symptomatic patients as the disease naturally progresses. In this model, by definition (Annex, 2013), all the patients who are discharged as symptomatic patients but originally admitted as non-colonized or asymptomatic patient and had their onsets of the symptoms after at least two days of their admission are considered to be HA CDI cases.

- **HCWs**: two types of HCWs, nurses and physicians, are modeled explicitly as agents in this model. They are also potential sources for the *C. difficile* spores. They are always in one of the two states: non-contaminated, or contaminated. The reason for modeling nurses and physicians separately is because they have different contact patterns with patients. Often, nurses contact patients more frequently than physicians do, with different "at risk" activities as it relates to personal care (e.g., toileting). The nursing assignments also tend to be clustered geographically. Conversely, physicians' interaction tends to be shorter in duration, but with more diversity in terms of the frequency of contact and the corresponding contact locations.

- **Visitors**: visitors are modeled similarly to HCWs in terms of contact behavior. However, visitors just interact with one patient and in room objects only. An assumption is made that they are less compliant with hand hygiene than HCWs are.

- **Environmental objects**: four types of environmental objects that are also potential sources of *C. difficile* bacteria are included in the model. They are high-risk-in-room-objects (HRIROs, e.g. toilets) (Eckstein et al., 2007), low-risk-in-room-objects (LRIROs, e.g., chairs) (Eckstein et al., 2007), high-risk-out-room-objects (HROROs, e.g., commode chairs, portable blood stations; other shared portable medical devices), and low-risk-out-room-objects (LROROs, e.g., computers). They are also always in one of the two states: non-contaminated, or contaminated.

There are also two types of entities modeled implicitly in the model. They are *C. difficile* bacteria and housekeepers.

- ***C. difficile* bacteria**: *C. difficile* bacteria are the causes for patients to be asymptomatic or symptomatic and for others agents to be contaminated. The *C. difficile* bacteria are modeled implicitly through assigning a state variable called age-of-bacteria to other agents in the model. This variable is used to record how long the bacteria that the entities harbor have been circulating around in the ward.

- **Housekeepers**: housekeepers clean the environmental objects. Therefore, those contaminated objects have a probability to become non-contaminated again.

4.2.2 Model process and schedules

In the model, twelve processes are implemented to reflect the patient flow and the transmission interaction of the various entities described above. The twelve processes fall into three categories based on their frequency, namely: *daily*, *shift*, and *hourly* processes. The detailed process flow charts are provided in Appendix B.

The processes under *daily* category include time-advance, patient-discharge, patient-admission, patient-antibiotic-intake, patient-natural-progression, scheduled-housekeeping, patient-environment-interaction, and visitor-visit-patient.

- **Time-advance**: time advance process updates temporary variables in the model, such as patient day and age-of-bacteria.

- **Patient-discharge**: if a patient has stayed in the ward for the LOS determined by the distribution, the patient will be discharged. The distribution of patient LOS adopted in this model is lognormal distribution fitted from the hospital data. Right after the discharge, housekeeping process will be

executed to clean the bed and the room that the patient has stayed in. The room will have a "terminal clean" if the patient is symptomatic. The room will just receive an "ordinary clean" if the patient is not symptomatic. When the patient gets discharged, the evacuated bed is available for the admission of a new patient.

- **Patient-admission**: as long as an empty bed is available, the ward admits a new patient. The admitted patients are in one of the three states regarding CDI: non-colonized, asymptomatic, or symptomatic. If there are multiple beds available, the beds are occupied in the order from four-bed room, semi-bed room, to private room bed. This sequence of admission aligns with the observation that the hospital always reserves the private rooms as isolation rooms. The admitted symptomatic patients will be isolated immediately if isolation rooms are available. Otherwise, the patients will be transferred out of the ward.

- **Patient-antibiotic-intake:** another condition for the development of CDI also considered in the model is the intake of antibiotics. Assumption is made that all antibiotics will act in a similar fashion with respect to disease history.

- **Patient-natural-progression**: from acquisition of the *C. difficile* bacteria and the intake of antibiotics to having the onset of symptoms, there is an elapsed time period. During this time period, the patients will be in the asymptomatic state. We model this transitional process as patient natural progression. We assume all the symptomatic patients will be detected in this ward.

- **Scheduled-housekeeping**: housekeeping will be performed every day for each room. As *C. difficile* spores are hard to kill, special housekeeping procedures are needed to deal with them. Therefore, in practice there are two types of housekeeping that can be observed: terminal housekeeping and ordinary housekeeping. Terminal housekeeping is more thorough and happens when CDI patients are clearly identified. Otherwise, ordinary housekeeping is performed. During

the housekeeping process, the contaminated objects will have a probability to become uncontaminated again.

- **Patient-environment-interaction**: patients will have interactions with their environmental objects, including HRIROs, LRIROs, and HROROs during a day. The transmission of *C. difficile* spores will happen if at least one of entities harbours the spores. We assume all of the patients will not leave their room.

- **Visitor-visit-patient:** there will be a dedicated visitor for each patient. Each visitor will have contact transmission with one patient. Visitors may also have contacts with in room objects (i.e., HRIROs and LRIROs).

The processes under *shift* category have nurse-shift-change, physician-shift-change, and physician-visit-patient.

- **Nurse-shift-change**: the ward has two shifts of nurses: day and night shift. The day shift has a larger number of nurses than the night shift does. When the shift changes, all the nurses for the incoming shift are in non-contaminated state. The process will also reassign patients to the new nurses. The nurse assignment has a geo-cluster property based on the practice in the hospital: a nurse is assigned to a group of rooms that are physically close.

- **Physician-shift-change**: essentially, the process of physician-shift-change is the same as the nurse-shift-change. The only difference is that the physician assignment does not have a geo-cluster property.

- **Physician-visit-patient**: physicians will visit their patients in their shifts. Before the visits, physicians will have a probability to remove the bacteria through hand hygiene if they are colonized. In this model, hand hygiene is assumed to be with soap and water to reduce contamination. During the visits, the transmission of *C. difficile* bacteria between patients and

physicians may occur. After the visits, physicians will have a probability to perform hand hygiene again to remove the bacteria if they are colonized. Lastly, physicians will use LROROs (shared by HCWs, e.g. computers, patient charts). Therefore, the transmission interaction will happen between physicians and LROROs.

The *hourly* category just has one process, which is nurse-visit-patient.

- **Nurse-visit-patient**: this process is essentially the same as the physician-visit-patient. Specifically, before visits, nurses will have a probability to perform hand hygiene. During the visits, we assume nurses and patients will also have a probability to have interaction with HRIROs and HROROs. After visits, nurses will also have a probability to do the hand hygiene again and use LROROs. In this process, transmission of the bacteria between objects may occur. Because nurses visit patients more frequently, the process will be executed in an hourly scale.

4.2.3 Key performance measures

The purpose of the simulation model is to evaluate the importance of the various entities that are the potential sources of the *C. difficile* bacteria in the transmission of HA CDI. Two key measures are designed for this evaluation purpose.

The first measure is the number of HA CDI cases that can be attributed to the sources (reservoirs). This measure is straightforward. The larger the number of HA CDI cases that can be attributed to a source is, the more important this source is in the transmission of HA CDI.

The second measure is the age of the bacteria. It measures how long the bacteria have existed in the environment. If the age of the bacteria is high, we can reason that the bacteria must have circulated in the ward environment for a long time. Otherwise, the bacteria must be imported from outside of the ward.

4.3 Model calibration

Parameter values in this model were obtained through: literature, hospital data, expert opinion, and calibration. Table 4.1 summarizes the key parameters obtained through ways other than calibration. All parameters regarding contact transmission were calibrated, as they have great uncertainty and almost no knowledge exists in the literature. The remainder of this section describes the calibration procedure and the results.

Table 4.1 Parameters obtained from literature or hospital data

Parameters	Value	Reference
Patient admissions		
Asymptomatic admission rate (%)	4	Lanzas et al., 2011; Samore et al., 1994
Symptomatic admission rate (%)	0.9	Clabots, Johnson, Olson, Peterson, & Gerding, 1992; Lanzas et al., 2011
Patient Characteristics		
Natural progression rate of asymptomatic patients becoming symptomatic (per day)	0.25	Donskey, 2010; Rubin et al., 2013
Average of length of stay of non-symptomatic patients (days)	7	Information, 2005, Hospital×
Prolonged length of stay of symptomatic patients (days)	6	Miller, Hyland, Ofner-Agostini, Gourdeau, & Ishak, 2002, Hospital
Antibiotic pressure: probability that patients will have antibiotics (%)	30	Moss, McSwiggan, McNicol, & Miller, 1981; Scheckler & Bennett, 1970
Housekeeping		
Terminal cleaning rate (%)	90	Eckstein et al., 2007
Routine cleaning rate (%)	40	Rubin et al., 2013
HCW visits and hand hygiene		
Number of contacts a patient will have with a nurse per day	20	Cohen, Hyman, Rosenberg, & Larson, 2012; McArdle, Lee, Gibb, & Walsh, 2006
Proportion that the contacts involve HRORO (%)	10	Expert opinion
Number of contacts a patient will have with a physician per day	2	Cohen et al., 2012; McArdle et al., 2006
Nurse before visit hand hygiene rate (%)	70	Erasmus et al., 2010
Nurse after visit hand hygiene rate (%)	90	Erasmus et al., 2010
Physician before visit hand hygiene rate (%)	40	Erasmus et al., 2010
Physician after visit hand hygiene rate (%)	80	Erasmus et al., 2010
Visitor		
Probability of pathogen carriage (%)	4	Ryan & Ray, 2010

× We obtained length of stay data from a hospital. The data was fitted by lognormal distribution through the software EasyFit

4.3.1 Calibration objectives

We calibrate the model in multiple dimensions (Railsback & Grimm, 2011), including four criteria: 1) incidence rate case-per-1,000-admissions; 2) incidence rate case-per-10,000-patient-days; 3) cumulative risk for infection; 4) cumulative risk for colonization (Kim, 2007). We calibrated to five patterns derived from the four criteria (i.e., two from the first two criteria, and three from the last two) for calibration.

The target patterns for criteria 1) and 2) are obtained from the Canadian Nosocomial Infection Surveillance Program website for *C. difficile*. The website records *C. difficile* incidence rates from 2007 to 2011 in three regions (i.e., western, central, and eastern) of Canada. We use the rates from central region as our targets as shown in Table 4.2 (i.e., Ontario is in the central region of Canada).

Table 4.2 Incidence rate of HA CDI in central Canada

Year	Case per 1,000 admissions	Case per 10,000 patient days
2007	5.07	8.22
2008	5.48	7.36
2009	4.98	5.91
2010	5.13	6.76
2011	6.21	8.37
Average	5.37	7.32

Source: http://www.phac-aspc.gc.ca/nois-sinp/projects/cdad-eng.php

The target patterns for criteria 3) and 4) are the published results from paper (Loo et al., 2011), as shown in Figure 4.1.

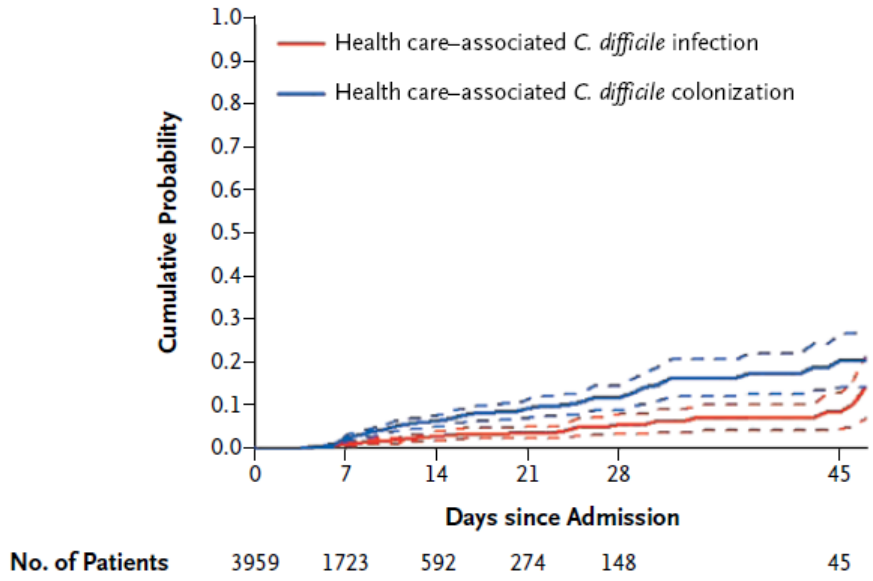


Figure 4.1 Cumulative probability of HA *C. difficile* colonization and HA *C. difficile* infection

Specifically, we try to match three patterns from Figure 4.1: 1) the cumulated risk of colonization is approximately two to three times larger than the cumulated risk of infections (Donskey, 2010); 2) the maximum cumulated risk of infection is approximately 10% (Loo et al., 2011); and 3) the infection risk increases with the increase of stay in the hospitals (Loo et al., 2011). Therefore, adding the previous two targets provided by the first two criteria, we have five patterns as our calibration targets in total.

4.3.2 Calibration methods and procedure

Three typical methods to calibrate a model include: 1) design of experiment (DOE); 2) Monte Carlo sampling methods; and 3) optimization (Stonedahl & Adviser-Wilensky, 2011).

DOE can be an effective method that can search the parameter space structurally and systematically. It is efficient when the dimensions of the model are small. However, when the dimensions of the model are large, it is not practical, as the computation cost rises dramatically with the increase of dimensions (i.e., “curse of dimensionality”).

Monte Carlo methods can be relatively effective to deal with the "curse of dimensionality" by using sparse sampling, although they also need considerable computation power. The simplest Monte Carlo method samples the parameter space randomly. However, this approach does not guarantee that the parameter space is properly sampled (Helton & Davis, 2003). One specific type of Monte Carlo method that potentially overcomes this problem is Latin Hypercube Sampling (LHS) (Helton & Davis, 2003; Marino, Hogue, Ray, & Kirschner, 2008). It is very popular for the exploration of computer simulation models for two main reasons (Viana, 2013): 1) the flexibility of including or dropping dimensions, and 2) the capability of covering both small and large parameter space.

The third method is optimization (Stonedahl & Adviser-Wilensky, 2011). If we know the reference patterns from the real system, we can construct an error measure which compares the model outputs to the reference patterns. Then, we can minimize the error measure to calibrate the model. As long as we have a good definition of the error measure, an optimization approach could be effective to calibrate the model. As the structure of a simulation model is often complicated, it is almost impossible to use exact methods to optimize the parameter sets to reach the targeted outputs. Therefore, in practice, stochastic heuristic methods, such as genetic algorithms (GA), are often employed as the optimization algorithm. There are also two problems with the stochastic heuristic methods. First, because of the built-in randomness, these methods (e.g. GA) might not be able to search the parameter space comprehensively. Second, they might not be efficient when the number of the reference patterns is large, as multiple criteria optimization is generally hard.

Therefore, after consideration of the advantages and disadvantages of the three different calibration methods, we did not rely on one method and we calibrated our model with a combination of the Monte Carlo method (i.e., LHS) and an optimization method (i.e., GA). We took a two stage approach for conducting the calibration by dividing and matching the five reference patterns separately. Specifically, in the first stage, we ran the LHS and GA, and chose the parameter sets that satisfied the first two criteria (the two incidence rates) to be fed into the second stage. The choice of the criteria in two stages is based

on the consideration of computation speed, as the program has to write more data out in order to compute the colonization and infection risk, which slows down the program dramatically. During the first stage, we were exploring and narrowing down the parameter space. Therefore, we did not take fully address the aleatory uncertainty (Marino et al., 2008) in the simulation model and we ran ten replications for each parameter set. In the second stage, we ran selected parameter sets with the required number of replications to get reliable results. We chose the parameter set that had the best fit to the patterns presented by the last two criteria. It is worth noting that the calibration process was not linear but iterative. Sometimes, during iterations, simulation code is adjusted if it is necessary. The calibration process flow chart is provided in Appendix C.

4.3.3 Calibration results

The calibration went through multiple iterations. In the final iteration's first stage, 20 sets of parameters that had a reasonable match with the incidence rates were found via LHS. We then obtained 20 sets of parameters from the GA approach. Therefore, 40 parameter sets were entered into the second stage. The final calibrated parameter values are in Table 4.3. This set of parameters achieved a good match to the patterns reported in the literature.

Table 4.3 Calibrated parameter value of baseline model

Parameters	Calibrated base value	Tested range	References
Contact transmission Parameters			
Patient-HCW			
Patient-to-HCW-transmission-rate (%)	20.12	9-25	Pittet et al., 2006
HCW-to-patient-transmission-rate (%)	9.12	9-25	Pittet et al., 2006
Patient—HRIRO (dirty sites, e.g. toilet)			
HRIRO-to-patient transmission rate (%)	1.72	0-10	Harrison, Griffith, Ayers, & Michaels, 2003
Patient-to-HRIRO transmission rate (%)	55.05	33-83	Dubberke et al., 2007
Patient—LRIRO (clean sites, e.g. phone)			
LRIRO-to-patient transmission rate (%)	2.73	0-10	Harrison et al., 2003
Patient-to-LRIRO transmission rate (%)	37.70	0-67	Dubberke et al., 2007
Patient—HRORO (portable share objects, e.g., commode chair, portable blood station)			
HRORO-to-patient transmission rate (%)	2.90	0-10	Harrison et al., 2003
Patient-to-HRORO transmission rate (%)	37.73	33-83	Dubberke et al., 2007
HRORO-clean-rate-after-use (%)	15.95	0-20	Hospital data, expert opinion
HCW– LRORO (HCW shared objects, e.g., computer, patients medication station)			
HCW-to-LRORO transmission rate (%)	10.00	10-15	Harrison et al., 2003; Mutters et al., 2009
LRORO-to-HCW transmission rate (%)	4.27	0-10	Harrison et al., 2003
Patient onset rate			
Patient-onset-rate (%)	29.14	10-33	Poutanen & Simor, 2004

4.3.4 Calibration summary

In summary, a multiple-target-oriented approach was utilized to calibrate the model. A calibration process that had two stages and utilized both LHS and GA methods was devised to overcome the high dimension problem for the calibration. After the intensive iterative calibration process, the model achieved high fidelity to the patterns found in the literature.

4.4 Uncertainty and sensitivity analysis

Uncertainty analysis (UA) is used to investigate the uncertainty in the model outputs that is generated from the uncertainty in parameter inputs. Sensitivity analysis (SA) assesses how the variation in model outputs can be apportioned, qualitatively or quantitatively, to different input sources (Marino et al., 2008). It is important to perform UA and SA for computer simulation models to obtain representative results and to interpret them properly. In this section, we show the procedures and results of the UA and SA after the calibration of the model.

4.4.1 Methods and procedure

Simulation results might be affected by two types of uncertainty: aleatory uncertainty and epistemic uncertainty (Marino et al., 2008). Aleatory uncertainty is introduced by the inherent stochasticity implemented in the simulation model (e.g., random number generator). Therefore, the simulation model will not necessarily produce the same outputs when repeated with the same inputs. Epistemic uncertainty, in contrast, usually stems from the limited knowledge of the parameter values used in the system.

A number of methods exist in the literature for UA and SA. The rationale for the methods we used for this study is based on Alden et al. (2013). Three analyses have been performed, namely aleatory analysis, local sensitivity analysis, and global sensitivity analysis. The flowchart of our UA and SA analysis in this study can be found in Appendix D.

The aleatory analysis is done before the second stage of the calibration process to ensure we get representative results from the calibrated parameter sets. The key of aleatory analysis is to determine the number of replications needed to have representative results. We used the “confidence interval” approach (Nelson, Carson, & Banks, 2001) to determine a value of 180. The remaining part of this section is devoted to describe procedures and the results of local and global sensitivity analysis.

All of the analyses were performed in the statistic environment R (http://www.r-project.org/) using packages providing connection between R and NetLogo for UA and SA.

4.4.2 Local sensitivity analysis (robust analysis)

One way of examining the impact of a specific parameter on the outputs is to hold the other parameters constant and perturb the targeted parameter systematically from its baseline value. The baseline value can be the result of the model calibration or a point in which we are interested in the parameter space. We can also explore the impact of other parameters in the same fashion. This one at a time approach is also a robust approach by Alden et al. (2013).

A parameter has significant impact if the perturbation of the parameter leads to a significant change of the outputs. To quantify the "significance", the Vargha-Delaney A-Test score is employed (Alden et al,. 2013). The Vargha-Delaney A-Test is a non-parametric rank-based effect size test (Vargha & Delaney, 2000), which can handle both continuous and discrete ordinal data without prior assumptions on the distribution of the data.

For presentation purposes, we show how the incidence rate (case-per-10,000-patient-days) and the infection risk respond to the change of the 12 calibrated parameters, as the parameters have similar effects on the other two measures. The summary is provided in Table 4.4.

Table 4.4 Parameter sensitivity results from local sensitivity analysis*

Parameters	**Base Value**	**Case-per-10,000-patient-days**	**Infection Risk**
Contact transmission Parameters			
Patient-HCW			
Patient-to-HCW-transmission-rate	20		
HCW-to-patient-transmission-rate	9	X	
Patient—HRIRO (dirty sites, e.g. toilet)			
HRIRO-to-patient transmission rate	3	X	
Patient-to-HRIRO transmission rate	55	X	
Patient—LRIRO (clean sites, e.g. phone)			
LRIRO-to-patient transmission rate	3	X	X
Patient-to-LRIRO transmission rate	38	X	X
Patient—HRORO (portable share objects, e.g. commode chair, portable blood station)			
HRORO-to-patient transmission rate	3	X	X
Patient-to-HRORO transmission rate	40	X	X
HRORO-clean-rate-after-use	16	X	X
HCW– LRORO (HCW shared objects, e.g. computer, patients medication station)			
HCW-to-LRORO transmission rate	10		
LRORO-to-HCW transmission rate	4		
Patient onset rate			
Patient-onset-rate	29	X	

*If a parameter has significant impact on an output measure, the cell where the parameter and the measure intersect is marked with "X".

As indicated in Table 4.4, both incidence rate and infection risk are sensitive to the transmission rates between patients and LRIROs, and between patients and HROROs. The incidence rate is also sensitive to the transmission rates between patients and HRIROs, HCW to patient transmission rate, and patient onset rate.

4.4.3 Global sensitivity analysis

The one at a time approach cannot reveal the compounded impacts from two or more parameters together. In order to see the joint impacts, global analysis is needed. In this study, we employed LHS/ Partial Rank

Correlation Coefficient (PRCC) schema (Marino et al., 2008) for the global analysis. LHS has the ability to sample the parameter space in a wide range with reasonable small sample size, and can be used for the global uncertainty analysis. It has been described in Section 4.3 for the calibration of the model. PRCC is similar and provides a generalization of the correlation coefficient that measures the linear relationship of two variables. It is considered a robust sensitivity measure for nonlinear but monotonic relationships between two variables. Therefore, it can provide a vehicle to index and quantify the relationships between model parameters and outputs. By combining LHS and PRCC, the sensitivity of the model outputs to the parameter variation can be reasonably evaluated (Alden et al., 2013; Marino et al., 2008). Table 4.5 is a summary of the PRCCs between model parameters and output measures.

Table 4.5 PRCC between model parameters and output measures

Parameters	Case-per-10,000-patient-days	Infection Risk
Contact transmission Parameters		
Patient-HCW		
Patient-to-HCW-transmission-rate	-0.04	0.02
HCW-to-patient-transmission-rate	0.29	0.19
Patient—HRIRO (dirty sites, e.g. toilet)		
HRIRO-to-patient transmission rate	0.67	0.42
Patient-to-HRIRO transmission rate	0.26	0.07
Patient—LRIRO (clean sites, e.g. phone)		
LRIRO-to-patient transmission rate	0.48	0.21
Patient-to-LRIRO transmission rate	0.39	0.19
Patient—HRORO (portable share objects, e.g. commode chair, portable blood station)		
HRORO-to-patient transmission rate	0.93	0.82
Patient-to-HRORO transmission rate	0.51	0.30
HRORO-clean-rate-after-use	-0.61	-0.34
HCW– LRORO (HCW shared objects, e.g. computer, patients medication station)		
HCW-to-LRORO transmission rate	0.05	0.07
LRORO-to-HCW transmission rate	0.07	0.05
Patient onset rate		
Patient-onset-rate	0.77	0.09

Panel A and Panel B of Figure 4.2 are two representative plots that show detailed information between the parameters and the outputs through the LHS/PRCC analysis schema. In Panel A, we can see what appears to be a clear strong correlation (PRCC=0.93) between the parameter HRORO-to-patient-transmission rate and the incidence rate. In Panel B, there is no strong relationship apparent between patient-to-HCW-transmission rate and the incidence rate as indicated by PRCC, which is close to 0.

As indicated in Table 4.5, the parameter that has the strongest impact is the HRORO-to-patient-transmission-rate. Its PRCCs for the incidence rate and the infection risk are as high as 0.93 and 0.82 respectively. The incidence rate is also very sensitive to patient-onset-rate; the PRCC between them is 0.77.

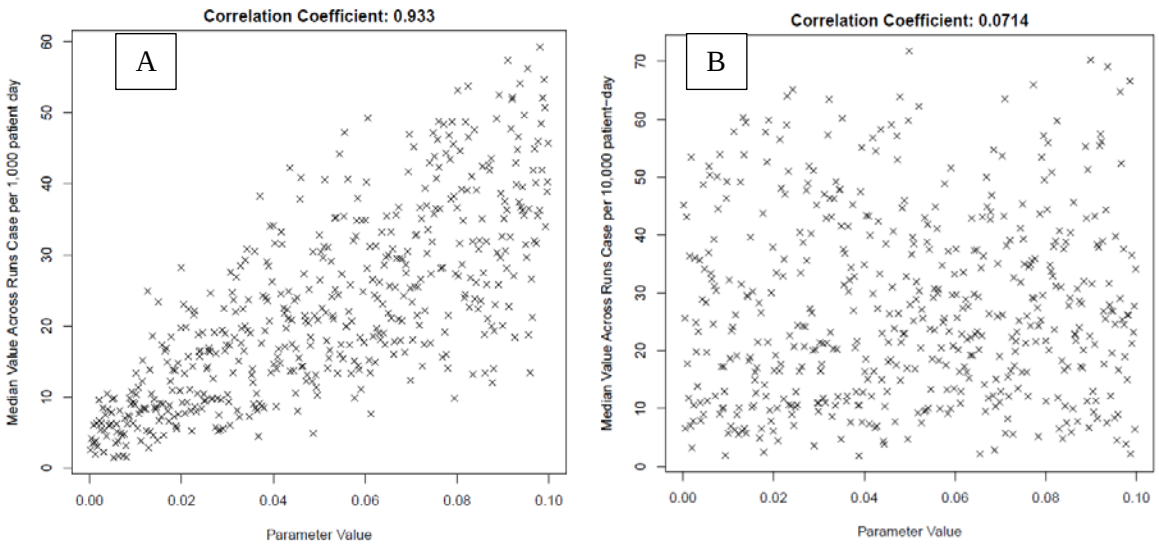


Figure 4.2 Example plots of LHS/PRCC analysis

4.4.4 UA and SA summary

The local and global sensitivity analysis results both suggest that the infection risk is relatively robust to the majority of the parameters except for the transmission rate from HRORO to patients. However, infection rates can be highly sensitive to a wide range of the parameters, especially to the rates related to HRORO, patient-onset-rate, and the transmission rate from HRIRO to patients.

4.5 Model explorations

In this section, we first look at the key performance measures produced by the baseline model to answer the questions proposed originally in 4.1. Then, we explore the model through the one at a time perturbation approach. The explored aspects include: 1) housekeeping; 2) hand hygiene compliance; 3) patient turnover; and 4) antibiotic pressure. The effects of these aspects are evaluated by the A-Test scores compared with the key performance measures obtained from baseline model. The performance measures compared include incidence rate, source of infection, and age of bacteria. All the detailed A-Test scores and performance measures results can be found in Appendix E.

4.5.1 Key performance measures of baseline model

The baseline model is the final calibrated model which satisfies the pattern of incidence rates and infection risk rates presented in the literature. In this subsection, we examine the following key performance measures: 1) sources of infection; 2) age of bacteria; and 3) effective transmission contact.

4.5.1.1 Sources of infection

Figure 4.3 shows the distribution of HA CDI cases attributed to different sources as derived from the baseline model. It shows that more than 95% percent of HA CDI cases can be attributed to the environmental objects other than HCWs. Among these objects, HRORO is the most important source. More than 50% of HA CDI can be attributed to it. This result appears to be consistent with the perception held by infection preventionists at the case study hospital that HROROs, like commode chairs, are the major entities that spread HA CDI.

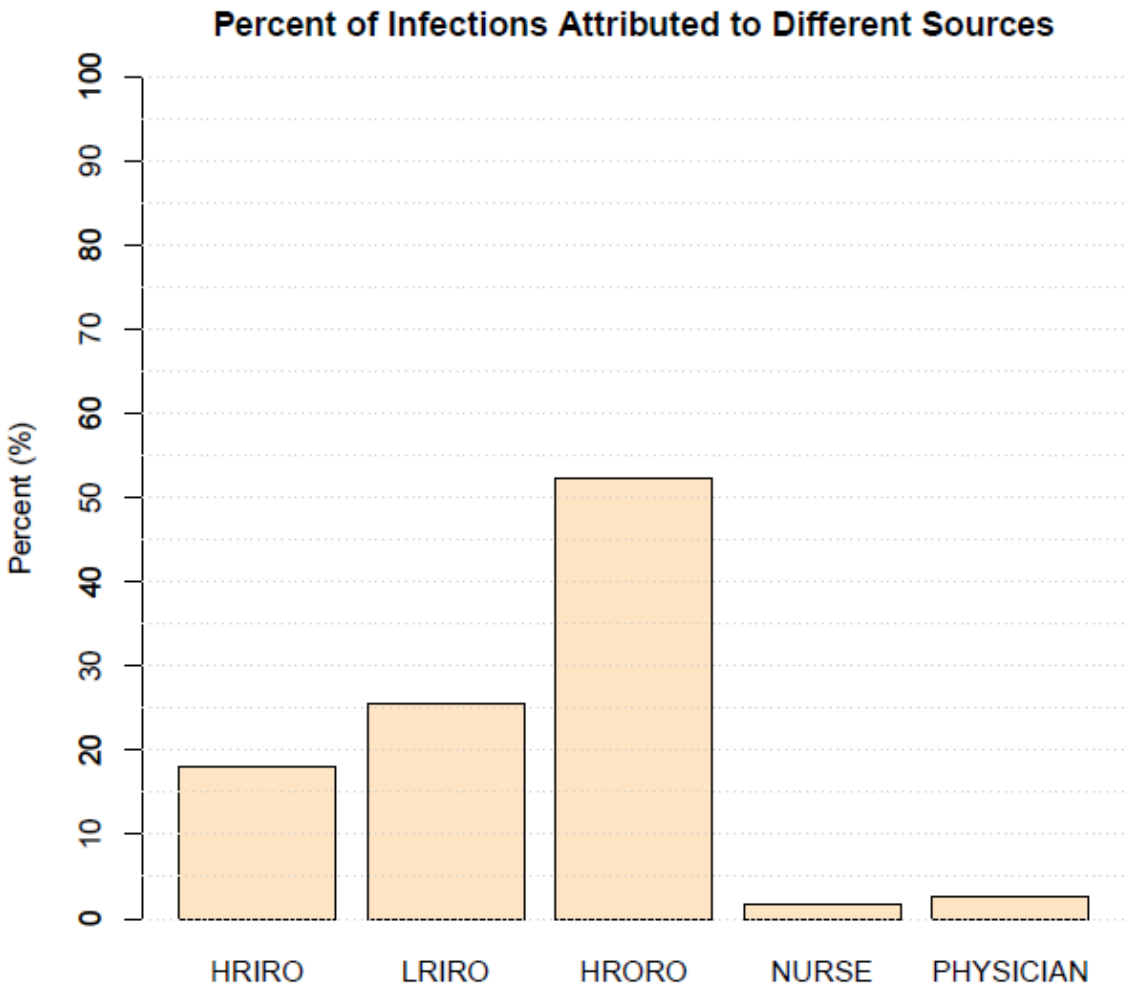


Figure 4.3 Distribution of infection sources

4.5.1.2 Age of bacteria

Figure 4.4 shows the distribution of max bacteria age on different entities. The largest median bacteria age is close to 90 days (three months), and the max of the bacteria age is approximately 150 days (five months). These numbers are consistent with the literature that *C. difficile* spores can survive in the environment for a long time (Ryan & Ray, 2010).

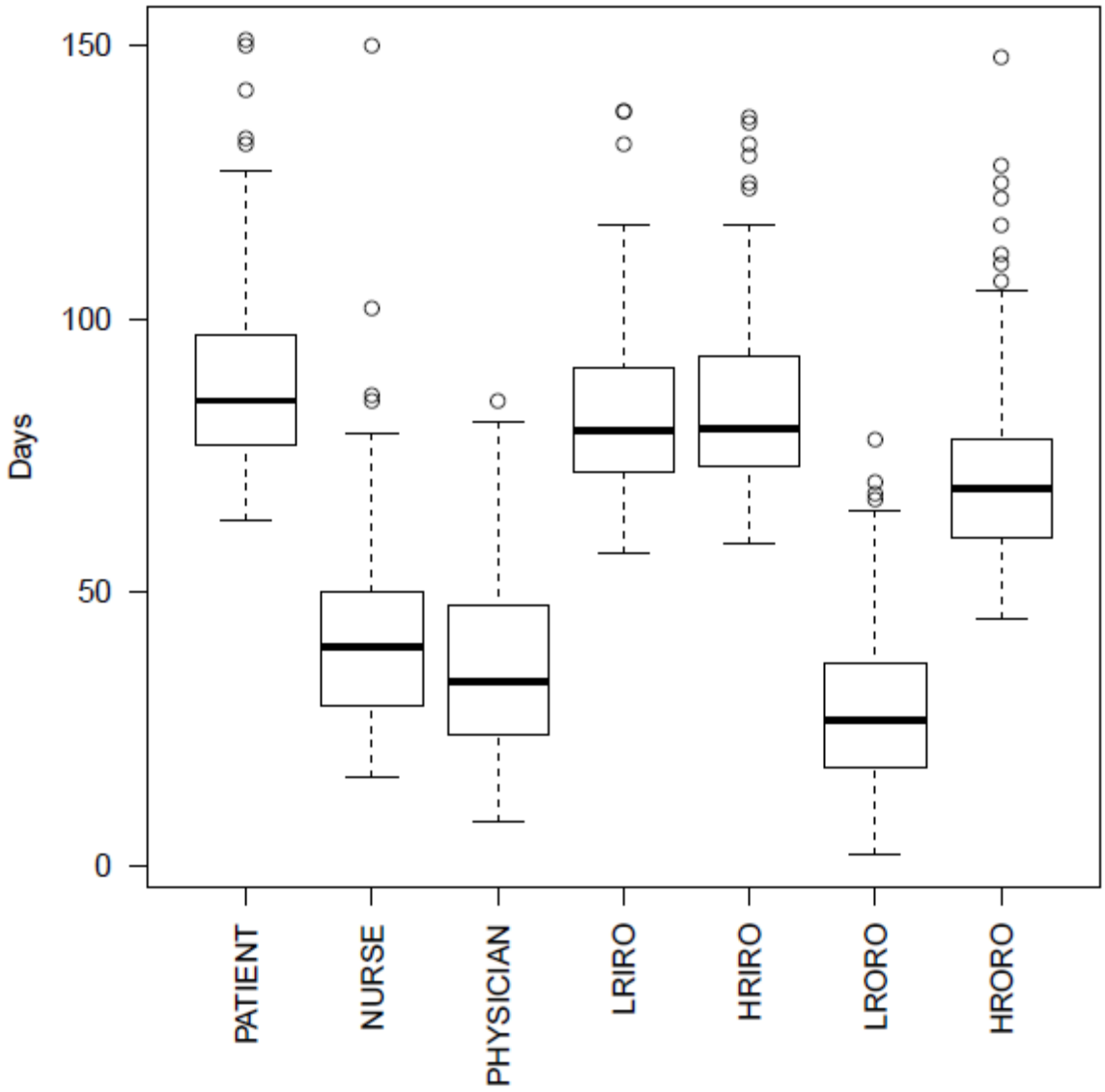


Figure 4.4 Distribution of max bacteria age on different sources

Figure 4.4 suggests that it is possible that the bacteria will circulate around the environment for a long time period. However, it is not clear whether old bacteria are a major contributor to HA CDI. Therefore, we collected the bacteria age data when HA CDI patients get colonized. Figure 4.5 is the empirical cumulative probability of the bacteria age at the colonization time of the HA CDI patients.

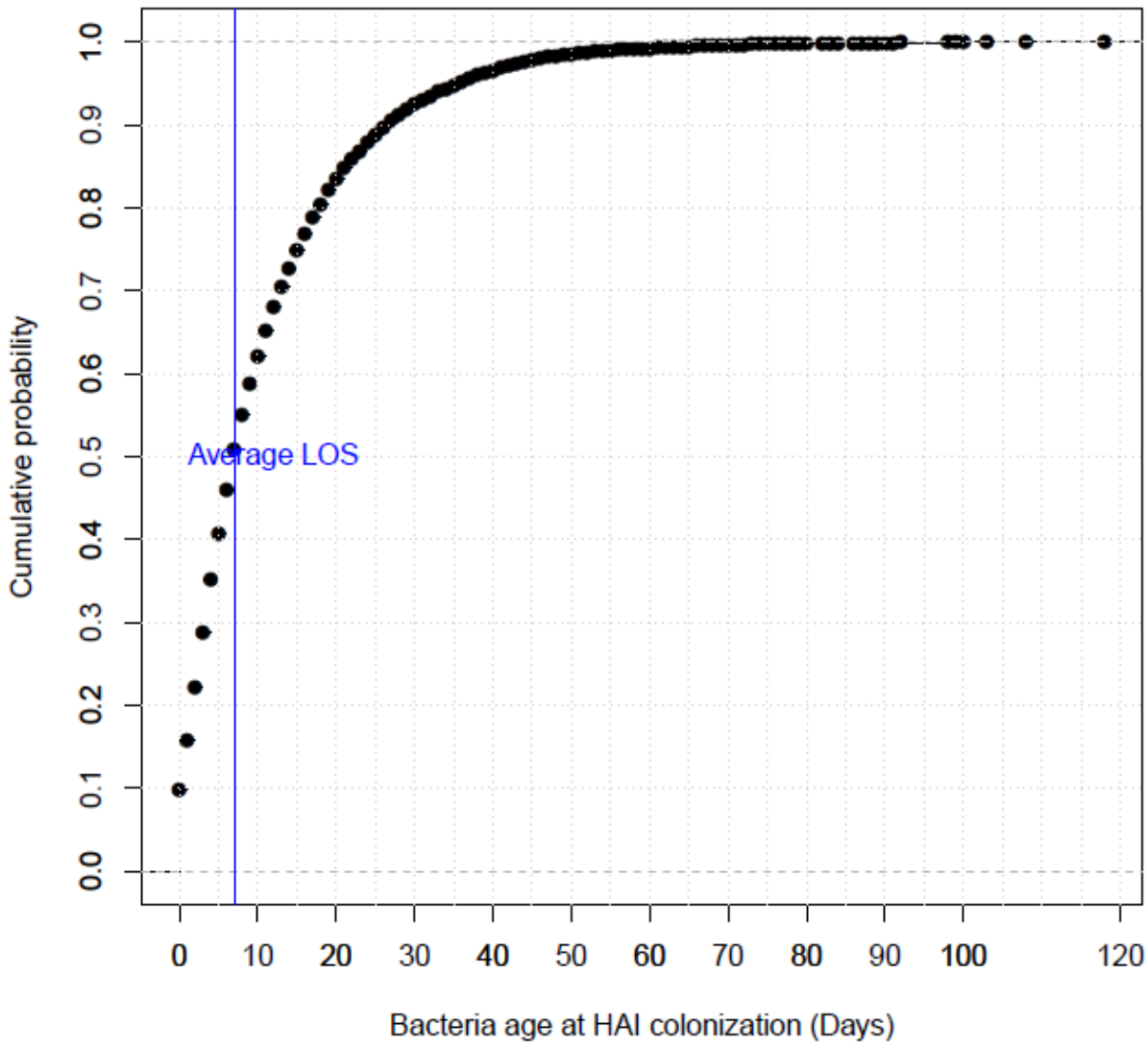


Figure 4.5 Distribution of bacterial age at colonization of HA CDI patients

Approximately 45% percent of the bacterial age at colonization is less than the average length of stay (LOS), and 90% of the ages are less than 30 days. This result implies that the majority of the HA CDI cases are probably caused by the newly imported *C. difficile* from outside of the ward, other than the bacteria that survived in the environment for a long time.

4.5.1.3 Effective transmission contact

We also looked at the number of effective transmission contacts among these sources and patients. An effective transmission contact is defined as a contact that leads to successful transmission of the bacteria from one object to another. This measure can serve as a cross-check for the first measure, the number of HA CDI cases attributed to different sources. Intuitively, more effective transmission contacts will occur between important sources and patients.

Figure 4.6 Panel A is a network plot that shows the number of effective transmission contacts (median value from multiple replications) that occur among the potential sources. Figure 4.6 panel B shows the percentage of the contacts "from" and "to" patients respectively. The "to" patients percentages match with Figure 8, the distribution of number of HA CDI cases attributed to different sources. This match is a reasonable cross check for the model logic.

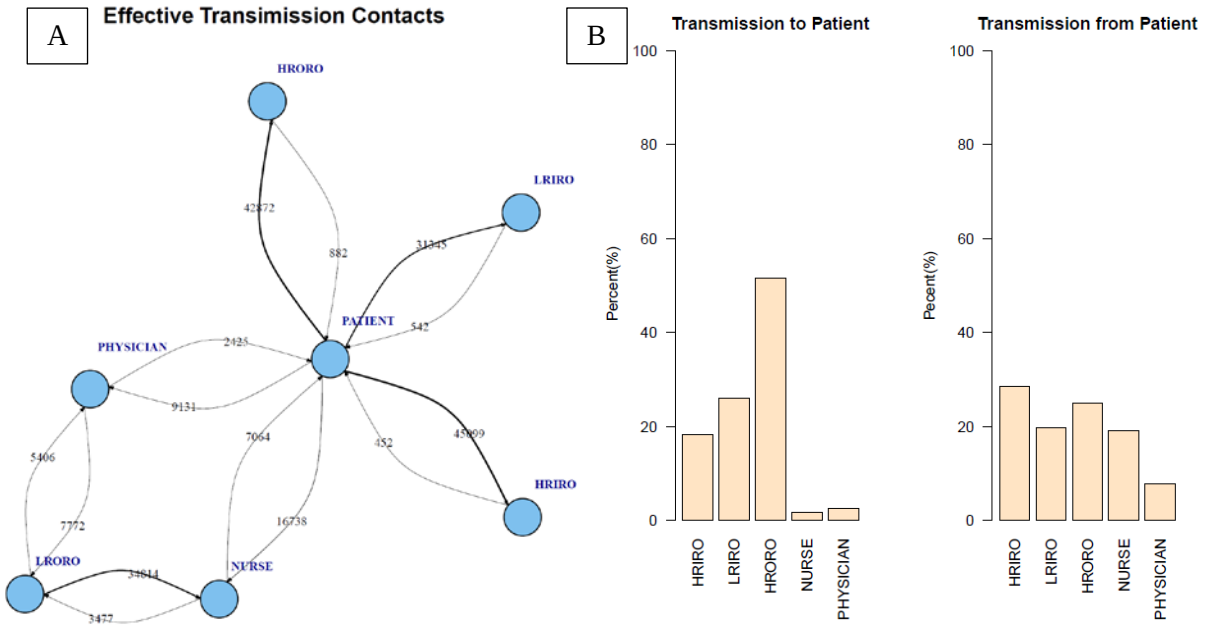


Figure 4.6 Effective transmission contacts among objects

4.5.1.4 Key performance measure summary

In summary, the results from the baseline model suggest that: 1) the majority of effective transmissions to patients are from objects other than from HCWs; and 2) based on the distribution of bacteria age at colonization, newly imported bacteria from asymptomatic carriers might be the main cause of HA CDI, although it is possible that some of the bacteria that circulating in the environment for a long time period might also be able to cause HA CDI. Therefore, admission screening might be an effective prevention method.

4.5.2 The effects of housekeeping

As *C. difficile* spores are hard to kill, special housekeeping procedures are needed to deal with them. Therefore, in practice, there are two types of housekeeping that can be observed: terminal housekeeping and ordinary housekeeping. Terminal housekeeping is more thorough and happens when CDI patients are clearly identified. Otherwise, ordinary housekeeping is performed. We evaluate the impact of the two parameters on the key performance measures based on the A-Test compared with the baseline model results. The following is the summary information obtained from these comparisons.

Incidence rate: It is suggested by the A-Test that the incidence rates are sensitive to the two parameters. However, the impact of ordinary-clean rate is much larger compared to the terminal-clean rate. To illustrate, in Figure 4.7, we plotted the detailed incidence rate responding to the change of ordinary-clean rate (Panel-A) and terminal-clean rate (Panel-B). The worst cases for both parameters happen at the parameter value at 0.1. However, the situation for ordinary-clean rate is much worse, as the incidence rate is much higher. Also, the figure shows that the effect of this parameter on the incidence rate is not linear. The improvement of ordinary-clean rate has much greater reduction of incidence rate in the lower range (i.e., 0 to 0.5) of the parameter values than that in the higher range (i.e., 0.5 to 1). This result suggests that

it is important to have a higher minimum standard for the ordinary-clean rate to control the *C. difficile* bacteria in the environment.

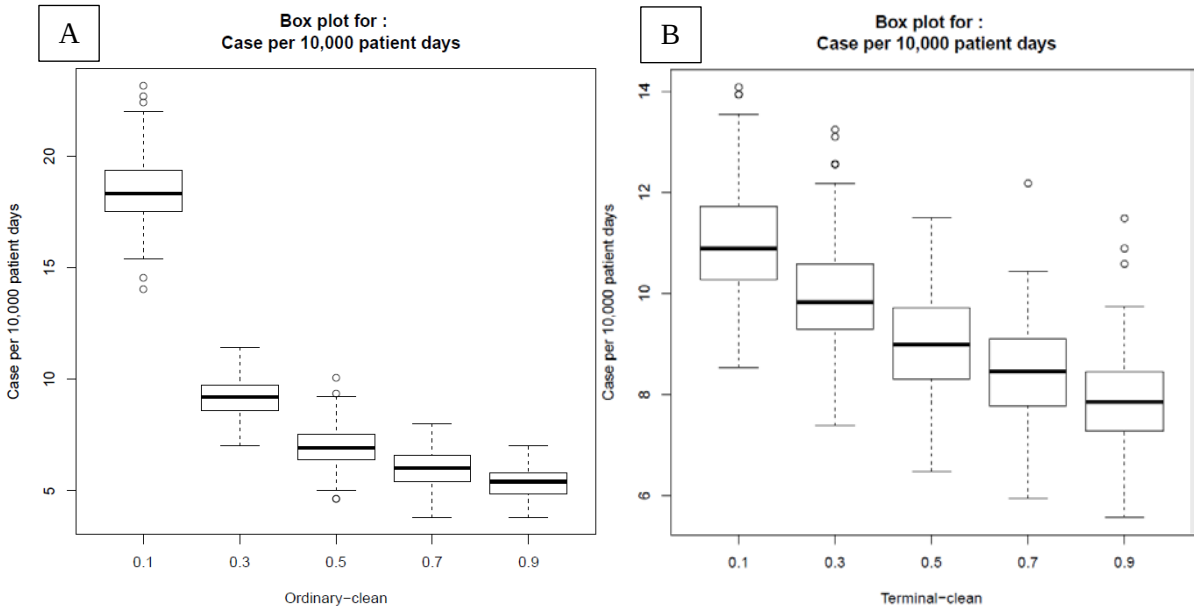


Figure 4.7 Distribution of incidence rate responding to housekeeping parameters' change

Sources of infection: It is suggested by the A-Test results that ordinary-clean has a larger impact than terminal-clean does: 1) at the lower parameter value range (0 to 0.5), ordinary-clean has impact on transmissions from both objects and HCWs; and 2) at the higher parameter value range (0.5 to 1), ordinary-clean still has impact on the transmissions from both objects and HCWs. However, terminal-clean only has impact on transmissions from HRIROs. These results provide further insights for why the ordinary-clean rate has a potentially larger impact on the incidence rate and the infection risk, as it affects multiple entities in the environment.

Age of bacteria: A-Test results also show that ordinary-clean rate has a similar pattern of impact on the max bacteria age as on the two measures examined above; while terminal-clean has very little impact on these measures. Combined with the results noted above, we can possibly reason that low ordinary-clean effectiveness may lead to the circulation of bacteria in the environment.

The results from the housekeeping effects analysis suggest that: 1) it is important to set a high minimum standard for the ordinary-clean effectiveness; otherwise, *C. difficile* bacteria might be able to circulate around in the environment for a long time; and 2) the ordinary-clean rate has a strong impact on the transmissions from all the entities including objects and HCWs, while terminal-clean rate impacts the transmissions from HROROs.

4.5.3 The effects of hand hygiene compliance

Hand hygiene compliance rate could be affected in many ways in practice. For instance, a high patient turnover rate can make the ward busy, and HCWs may tend to have lower hand hygiene compliance rate when the ward is busy.

The structure of the ward might also have impact on the effectiveness of hand hygiene. The soap and water based hand wash method is considered to be more effective to remove *C. difficile* spores from HCWs' hands than the alcohol-based hand rub method. Therefore, it is believed that installing more sinks in the ward can increase the effectiveness of hand hygiene for removing *C. difficile* spores.

In practice, hand hygiene is required for both before and after the visits of patients. We have two types of HCWs considered in the model: nurses and physicians. Therefore, there are four types of hand hygiene rates in this model. We examined the effects of the four parameters as follows.

Incidence rate: As indicated by the A-Test results, the two before-visit rates appear to have no significant impact on the incidence rate. In contrast, the two after-visit rates appear to present a significant impact on the incidence rate. In Figure 4.8, we plotted the detailed incidence rate responding to the change of nurse-after-visit hand hygiene rate (Panel-A) and physician-after-visit hand hygiene rate (Panel-B). As shown in the figure, the impact from the physician-after-visit hand hygiene rate is larger than the impact from the nurse-after-visit hand hygiene rate. These results imply that emphasis for hand hygiene compliance could be put on the after visit compliance of HCWs, especially of physicians.

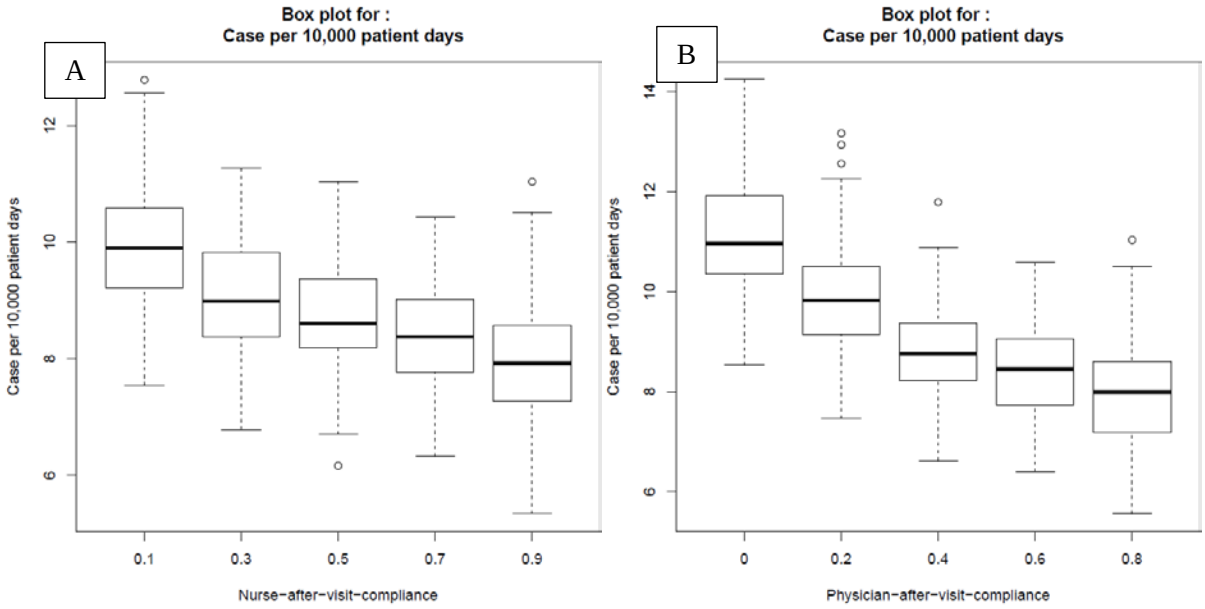


Figure 4.8 Distribution of incidence rate responding to the after-visit hand hygiene parameters' change

Sources of infection: The A-Test scores suggested that all four of the hand hygiene rates have significant but varying impact on the number of infections attributed sources. Figure 4.9 shows the distribution of the number of infections attributed to different sources by changing the values of the four hand hygiene parameters. Panel A and B are for the two before visit rates, and Panel C and D are for two after visit rates. As shown in Figure 4.9, the absolute values and changes in the two after visit cases (Panel C and D, especially in Panel D for physician-after-visit-rate) are much larger than that in the two before visit cases (Panel A and B). One possible explanation for this wide impact is that physicians have a wider contact network so that bacteria can easily spread through their hands if their hands were contaminated.

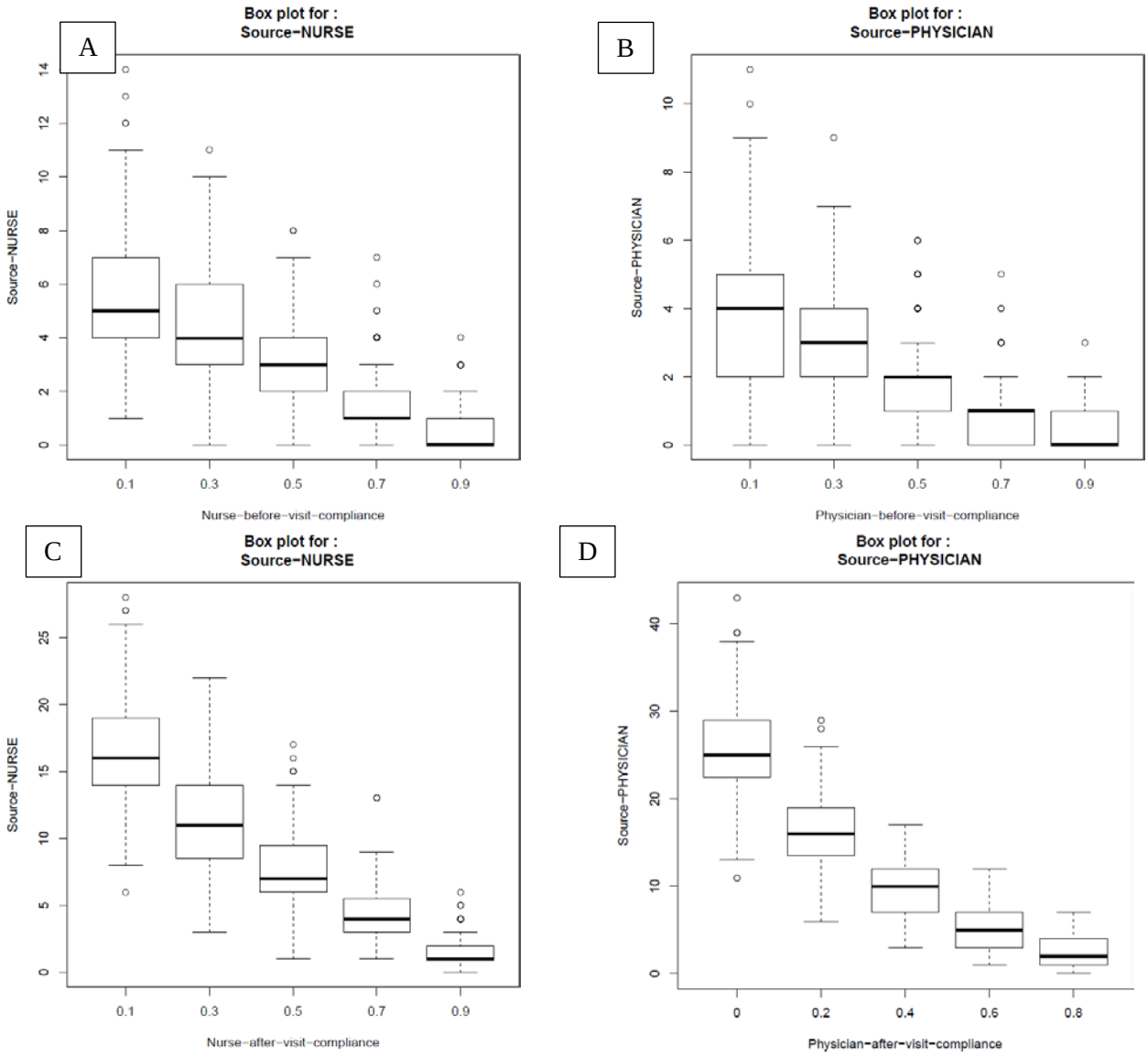


Figure 4.9 Distribution of incidence rate responding to hand hygiene parameters' change

These results further suggest that improving physician after visit compliance might be an effective policy to reduce HA CDI if the compliance rate is low.

Age of bacteria: The A-Test scores suggest that the two before-visit-hand-hygiene rates have almost no impact on the bacteria age measure. Whereas, the nurse-after-visit-compliance-rate appears to have a

significant impact on the bacteria age of both nurses and LROROs, as nurses have frequent contacts with LROROs. The physician-after-visit-rate also appears to have a significant impact on the bacteria age of multiple entities (e.g., physicians and LROROs). This is not unexpected as physicians have a wider contact network.

The analysis of hand hygiene compliance rates suggests that: 1) in general, the two after visit compliance rates can have a much larger impact on the transmission of the infections; 2) as the contact network of physician is wider, the physician after-visit-compliance rate may have the most significant impact on the spread of the HA CDI among the four parameters; and 3) hand hygiene policies could pay special attention to the compliance of these two rates, particularly when they are low.

4.5.4 The effects of patient turnover

In practice, high patient turnover might impact the transmission of HA CDI in two opposite directions. On one hand, high patient turnover rate means patients will stay in the hospital shorter, and hence patients have a smaller chance to get infected and serve as a transmission source. On the other hand, high patient turnover rate means that the hospital is busier, housekeeping might not be well performed, and HCWs might not have good hand hygiene compliance while dealing with the faster turnover rate. Since the impact of a higher patient turnover is rather complicated, we will limit the discussion in this chapter to the first impact, a shorter hospital stay. Specifically, we change the distribution of patient length of stay by shifting the average of the distribution and hold all the other parameters the same with the baseline model.

Incidence rate: The results of A-Test show that a small shorter or longer difference in the length of stay has a strong impact on the overall incidence rate, as shown in Figure 4.10. Further examination on infections risk reveal that will not have much impact on the max infection risk for patients. The possible explanation for this phenomenon is that, while the infection risk based on the length of stay does not change, the number of patients that stay longer in the hospital increases as patient turnover becomes slow.

Therefore, the number of patients that get infected becomes larger. However, as shown in Figure 4.10, the magnitude of the change is not large.

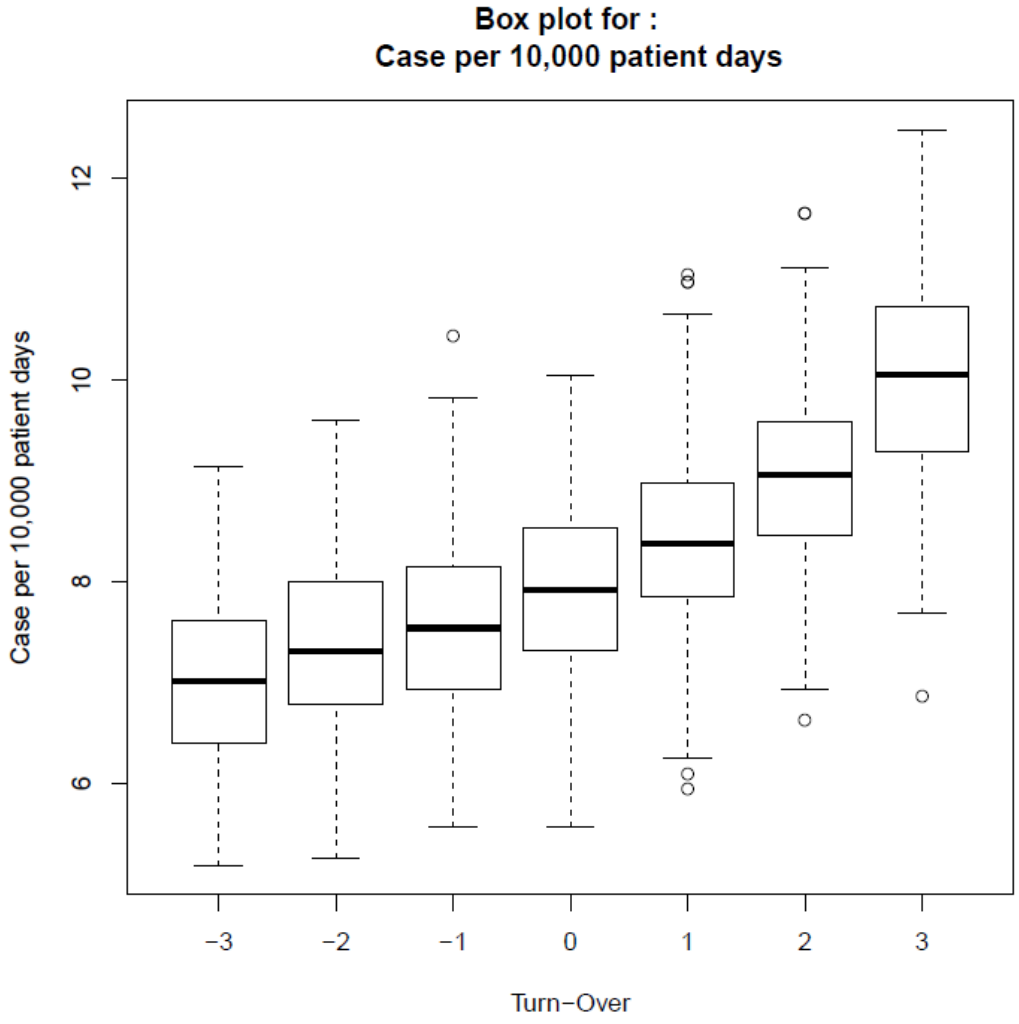


Figure 4.10 Distribution of incidence rate responding to patient turnover parameter's change

Sources of infection: The A-Test scores for the infection sources compared with the baseline model results shows that as the average length of stay become longer, the number of patients that get infected from the objects becomes larger, while the number of patients that get infected from the HCWs stays relatively stable. This result suggests that the increase is associated with the transmission from environment objects, which is consistent with the implication that the baseline model suggests.

Age of bacteria: The A-Test results for the max bacteria ages on various entities compared with the baseline model shows that patient turnover has distinct impacts on different entities, although the impacts are not statistically significant. Specifically, the max bacteria ages on the objects increase slightly and the bacteria age on HCWs has no change as patient turnover slows down.

The effects of patient turnover are complicated, as it also might have impact on other parameters such as hand hygiene and housekeeping. In this analysis, we did not consider the complicated version of the problem. All the other parameter values were held as the same as the ones in the baseline model. The results from this analysis suggest that reducing patient LOS might reduce (although not significantly) HA CDI incidence rate when high patient turnover rate puts no stress on housekeeping and HCWs.

4.5.5 The effects of antibiotic pressure

Antibiotic use is an important condition for the development of HA CDI. In this section, we look at the impacts of antibiotic use.

Incidence rate: The A-Test scores result shows that the increase of antibiotic pressure will significantly increase both the incidence rate, as well as the infection risk. To illustrate, Figure 4.11 is the boxplot of the incidence rate at different value points. As can be seen, the incidence rate is highly sensitive to the change of the antibiotic pressure as there is no overlap among the boxes in the plot.

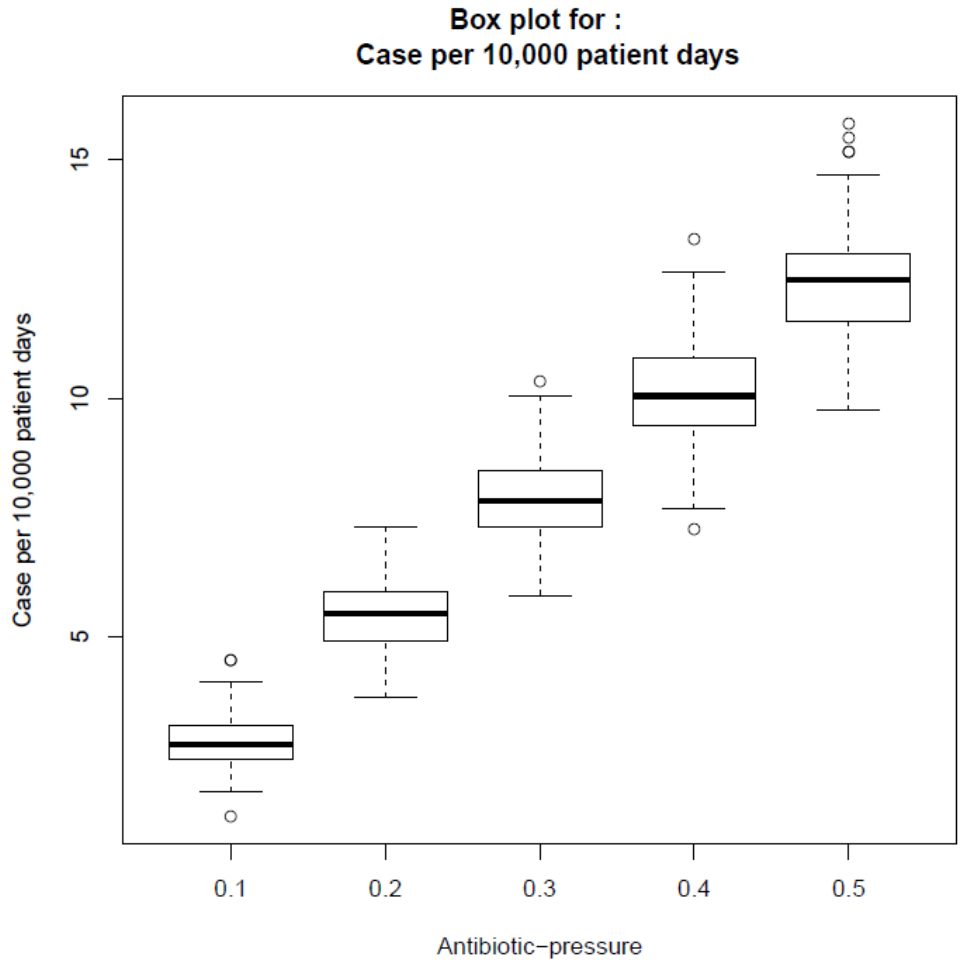


Figure 4.11 Distribution of incidence rate responding to antibiotic pressure parameter's change

Sources of infection: The A-Test scores for the infection sources suggest that the increase of antibiotic use leads to the increase of infections from all transmission sources, although the increase from the environment objects is significant higher.

Age of bacteria: The A-Test scores for the age of bacteria suggest that the max bacteria ages in various entities have almost no change.

In summary, both the infection risk and the incidence rate appear to be highly sensitive to the antibiotic use. The change of infection risk might be the underlying reason for the change of incidence rates.

Therefore, the result suggests that antibiotic stewardship program might be effective in containing HA CDI.

4.6 Discussion and conclusions

In this study, an agent based simulation model was built to evaluate the importance of the potential sources of the *C. difficile* spores and the impacts of various parameters involved in the transmission of HA CDI in a non-outbreak ward setting. The construction of the model went through multiple iterations and incorporated knowledge from a field study, published data, and consultation from infection control experts in a hospital. The model was then calibrated through a two stage procedure which utilized both LHS and GA to target five different patterns reported in the literature. The final calibrated model achieved high fidelity when compared to the literature. Both local and global sensitivity analysis were performed to examine the calibrated parameters. Based on the calibrated baseline model, several key measures such as bacteria age and source of infection were used to understand the roles played by different entities in the transmission of HA CDI. Several aspects of the model including housekeeping, hand hygiene compliance, patient turnover, and antibiotic pressure were further explored.

Assuming that the simulation model is a reasonable representation as suggested by the high fidelity, we have the following observations:

1) The distribution of bacterial age at colonization from the baseline model analysis suggests that newly imported bacteria might be the major cause of HA CDI in non-outbreak endemic setting, although it is possible that some of the bacteria that are circulating in the environment for a long time might also cause HA CDI. This result suggests that the genetic diversity of *C. difficile* bacteria noted by Eyre et al. (2013) might arise from the import of new *C. difficile* bacteria from outside of the ward. This baseline model result is also consistent with the results of Lanzas et al. (2011).

2) Objects (i.e., HROROs, HIROROs, LIROROs), especially HROROs, other than HCWs might be the major sources where patients obtain *C. difficile* from. Specifically, the analysis of the distribution of where these cases obtained the bacteria from suggests that more than 95% percent of HA CDI can be attributed to the various objects other than HCWs. Among these objects, HROROs appear to be the most important spreaders.

3) The incidence rate is highly sensitive to a large number of parameters, especially to the rates related to HROROs, the patient onset rate, and the transmission rates from HRIROs to patients.

The exploration of the model suggests that the prevention of HA CDI might benefit from the following policies:

1. The baseline model results suggest that HA CDIs in the non-outbreak setting are mainly caused by the bacteria imported from outside of the ward. This implies that admission screening might be an effective tool to further reduce the incidence in these setting. A recent study by Lanza and Dubberke (2014) demonstrated the effectiveness of admission screening for the reduction of HA CDI; our results give further evidence and explanation as to why admission screening might be effective.

2. The results from the housekeeping effects analysis suggest that under the condition of high terminal-clean effectiveness, the circulation of bacteria might be mainly caused by the ineffectiveness of ordinary-clean, as it has a wider impact range. The implication of these results is that we should set a high minimum standard for the ordinary-clean effectiveness.

3. The analysis of hand hygiene compliance rates suggests that the two after visit compliance rates (i.e., nurse and physician) have much larger impact on the transmission of the infections than the before visit compliance rates do. Moreover, as the contact network of physicians is wider, a low physician after-visit-compliance rate might potentially spread HA CDI widely. The research

suggests that hand hygiene policy should pay special attention to the compliance of these two after visit compliance rates.

4. The results of the patient turnover effect analysis are inconclusive, as the impact mechanism of the patient turnover is complicated. Nevertheless, the results suggest that reducing patient LOS might reduce the HA CDI incidence rate when high patient turnover rate does not put stress on housekeeping and HCWs.

5. Finally, the analysis of antibiotic pressure suggests that HA CDI incidence is highly sensitive to the antibiotic use. The result implies that antibiotic stewardship program might be an effective tool in containing HA CDI.

Chapter 5. Dynamic network analysis of hospital acquired *Clostridium difficile* infection transmission

5.1 Introduction

A widely used tool to better understand the dynamics of infectious disease outbreaks is network epidemiology. As shown in Chapter 2, since the introduction of this method to the analysis of the dynamics of HAIs by Myers et al. (2003), a few studies have attempted to further advance this field (Barnes, Golden, & Wasil, 2010; Curtis et al., 2009; Cusumano-Towner, Li, Tuo, Krishnan, & Maslove, 2013a; Herman et al., 2009; Ohst, Liljeros, Stenhem, & Holme, 2014; Ueno & Masuda, 2008). One of the major themes of these studies is to understand the outbreak dynamic on the HAI contact networks. The typical strategy of outbreak dynamic analysis in these studies is: 1) establish the network through either theoretic assumptions (Barnes et al., 2010; Meyers et al., 2003) or empirical data (Curtis et al., 2009; Cusumano-Towner et al., 2013a; Ueno & Masuda, 2008); 2) simulate disease-spread on the established network through epidemiology models (e.g., susceptible-infectious-recovery (SIR)); and 3) evaluate different parameter or structure settings of the simulation to find the best intervention actions.

One of the major assumptions associated with this analysis strategy subject to debate is that the established networks often do not have structure change and are static (Barnes et al., 2010; Curtis et al., 2009; Cusumano-Towner et al., 2013; Meyers et al., 2003; Ueno & Masuda, 2008). First, patients normally just stay in hospitals for a few days and the frequent patient admission and discharge are suspected to have significant impact on the spread of the HAIs. Static networks may not be able to capture the impact from the patient dynamics. Second, static networks are hard to adapt to the real time environment for the purpose of prediction and prevention, as the network structure often changes. Another issue associated with past network studies on HAI is that most of the networks have been constructed for one type of HAIs: MRSA. Many other types of HAIs exist and do not appear to have been

extensively studied. Infectious pathogens often have different transmission routes and it is possible that this can influence outbreak patterns. Hence, it is not clear whether the conclusion from these studies on MRSA might be applicable to other types of HAIs.

In this chapter, we explore the potential of studying HAI networks from a dynamic and predictive perspective for HA CDI with the objective of finding potential predictors for the outbreak of HA CDI based on network characteristics.

This research includes the analysis of two types of networks regarding the transmission of HA CDI. We first investigated the contact network for two wards which experienced outbreaks in 2011 and 2012 respectively. We compared the network characteristics in these two wards during three periods of time, which are before-outbreak period, in-outbreak period, and after-outbreak period. A traditional analysis of the network did not find any significant or particular pattern change in the statistics that are commonly used to characterize a network in the three periods for both of the outbreak wards. However, as part of the analysis, an animation of the dynamic network for one ward suggested a pattern: that the appearance of the HA CDI is correlated to the admission of *C. difficile* carriers (i.e., lab test antigen positive). Inspired by this observation, we investigated the CDI transmission for the whole hospital using an analysis approach that combined time series data mining and predictive classification models. The initial pattern we observed for one ward via the animation was confirmed for the wider context when using the quantitative analysis. We then extracted the characteristics of the transfer network and used them to predict whether an admission of *C. difficile* carrier will be followed by the appearance of HA CDI for different time windows. The predictive model achieved highest performance of 0.75 in terms of area under curve (AUC) (Hanley & McNeil, 1982) at the window size of six days.

Although the analysis and methodology can still be considered exploratory and preliminary, the analysis has demonstrated the potential of using network statistics as predictors for the transmission of HA CDI when they are used in a method that combines time series mining and predictive modelling. This study

suggests that the novel methodology developed to better understand HA CDI might be an innovative methodology that has a wider applicability to the field of dynamic network analysis.

The rest of the chapter is organized as following. In Section 5.2, we describe the construction and the analysis of the contact network for the two outbreak wards. In Section 5.3, we show the construction and the analysis of the hospital transfer network. The chapter then concludes with a discussion on the limitations of the study and the direction for future research.

5.2 Network analysis of the two outbreak wards

5.2.1 The definition of HA CDI outbreak and the two outbreak wards

The definition of HA CDI outbreak according to Provincial Infectious Diseases Advisory Committee (PIDAC) in Ontario includes three scenarios, which has been described in Chapter 1 and re-stated here:

1. For wards/units with ≥20 beds, three (3) new cases of nosocomial CDI identified on one ward/unit within a seven-day period OR five (5) new cases of nosocomial CDI within a four-week period;
2. For wards/units with <20 beds, two (2) new cases of nosocomial CDI identified on one ward/unit within a seven-day period OR four (4) new cases of nosocomial CDI within a four-week period;
3. Facilities that have a facility nosocomial CDI rate that exceeds their annual nosocomial baseline rate for a period of two consecutive months.

In the facility we studied, two outbreaks were recorded in the system after the introduction of the infection surveillance system.

The second and latest outbreak occurred in a ward with a bed size of 14 in October 2012. The outbreak had two patients, whose onset times were on the same day. The outbreak was a realization of Scenario 2. Both of the rooms have more than one bed, and they are physically close to each other.

The first outbreak happened in a ward with a bed size of 21 in June 2011. The outbreak also had two patients, whose onset times had a two day interval. The declaration of the outbreak was based on the third scenario. These two patients acquired the infection in two different rooms. These two patients also acquired the infection in two different but physically close rooms.

These outbreaks illustrate some of the difficulties with studying HAI outbreaks using empirical data. The sample sizes can be relatively small and the events can be infrequent. From a hospital perspective, these are preferred attributes, but they can create challenges for any quantitative analysis.

5.2.2 The construction of the networks and the data source

The core of network construction is to find the vertices and to establish the edges. Since our focus is on the transmission of CDI from patient to patient, the patients are the vertices of this network. The edges are the possible contacts that might lead to the transmission of infection. Based on the medical literature and discussion with an infectious disease specialist, three types of contacts were considered for the transmission of CDI: 1) room sharing, 2) physician sharing, and 3) nurse sharing. It is worthwhile to mention that we considered one specific type of nurse-patient contact for the establishment of nurse sharing edges. This type of nurse-patient contact was of the gastrointestinal type given the risk of nurse contamination with spores that this interaction will pose.

Four data sources in the hospital were used for the construction of this network. The first data source was the admission-transfer-discharge (ATD) system, which provides information about the movement and location of the patients in the hospital. This data source provided the information of the vertices of our network (patient and their times in the ward). The room sharing edges can be also derived from it. The second data source was the physician portal system, which provided the information about the physician sharing between patients. The third data source was the nurse charting system, which contains the

charting information performed by nurses. The last data source was the HAI surveillance system in the hospital which keeps record of the HAI patients.

The two contact networks for the two wards were constructed as dynamic networks. As patients are admitted or discharged dynamically, their contacts only existed for a short period of time.

5.2.3 The analysis

5.2.3.1 The extraction and comparison of network statistics

We computed the common summary statistics of the networks for the wards for the three periods (before-outbreak, in-outbreak, and after-outbreak). The statistics collected include number of vertices, number of edges, average degree, cluster coefficient, density, average path length, and diameter (Brandes & Erlebach, 2005). Each day has a network constructed for it and the key statistics are extracted from these daily networks. Then, these computed statistics are formed into time series groups. We then compared these statistics at different time periods. Depending on the window size chosen for the definition of these period, these statistics for each period can vary. Figure 5.1 shows the comparison of the distribution of the statistics through a boxplot for two windows sizes (two months and one month respectively) for the 2012-Outbreak network.

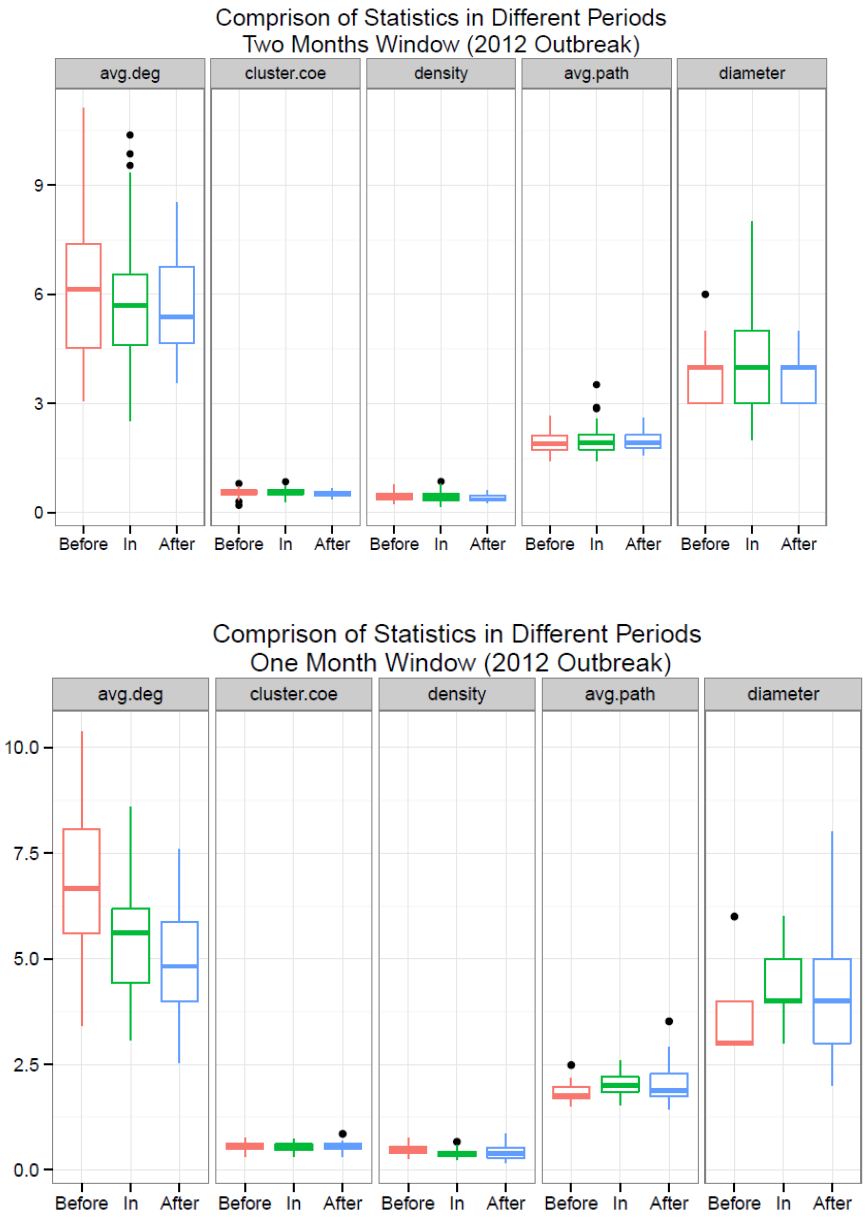


Figure 5.1 Comparisons of network statistics

As shown in Figure 5.1, there are significant overlaps of the boxes for the three periods for each of the extracted parameters especially when the window size is large. These overlaps suggest no significant statistical differences exist in the three periods. This observation was further confirmed by statistics

testing. We obtained similar results for the 2011-Outbreak. However, we have to be cautious to interpret these results, as our sample size is small.

5.2.3.2 The animation of the dynamic network and the pattern

Dynamic networks that change over time are often hard to understand by static analyses of numbers alone. The temporal changes, relationships, and sequences that define how the network changes can be hidden in aggregate statistics or statistics gathered via snap-shots. To help us understand how the networks being studied evolved over time, it was decided to create visual animations of the two outbreak dynamic networks.

One challenge for the animation of dynamic network is maintaining the "mental map" (Kolaczyk, 2014), as the vertices and edges may change in dynamic network when temporal information is added. This is especially true for the animation of the patient admission and discharge in hospital wards, as patients are constantly flowing in and out of the ward. In order to achieve the stability across visualizations, we included all of the patients in the animation period and fixed their position in an overall graph and only changed their status and edges with the updates of temporal information. Figure 5.2 is one of the frames from the animation for 2011-Outbreak on June 8th. In the figure, the vertices (dots) represent patients. The colors show the status of the patient on this particular time. Specifically, the white is for discharged patients; the grey is for future patients who will get admitted; the yellow is for recurrent CDI patients; the purple is for HA CDI patients from other facilities; the pink is for community acquired CDI; the red is for HA CDI patients; and the blue is for patients who are in ward now and are not related to any type of CDI. The lines represent the contacts among patients. The colors show the type of the contacts: the light blue is for nurse sharing contacts; the green is for physician sharing contacts; and the brown is for room sharing contacts. The full animation for the two outbreak wards can be found in *Outbreak2011.gif* and *Outbreak2012.gif*.

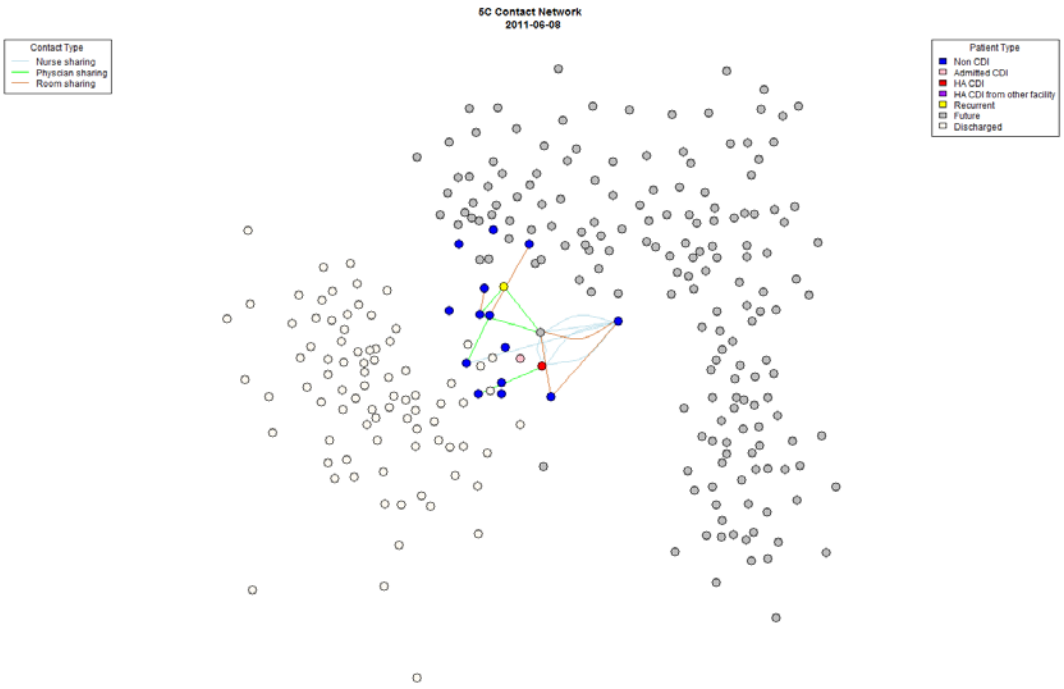


Figure 5.2 Animation frame for ward level dynamic contact network

It was observed through the animation that the occurrence of HA CDI appeared to be correlated with the presence with *C. difficile* carriers (i.e., community acquired CDI patients, recurrent CDI patients, HA CDI patients from other facility). Moreover, a few studies in literature have indicated that the occurrence of the HAI might be caused by the import of the bacteria from outside (Eyre, Cule, et al., 2013; Lanzas et al., 2011; Walker et al., 2012). But the studies often analyzed a single hospital site. It is not clear whether the observation from this study could be generalized into other places.

Therefore, we decided to examine the pattern (i.e., the admission of *C. difficile* carriers followed by the appearance of HA CDI) in a larger data set from the surveillance system in this hospital site. This data set included the admission and onset dates of both HA CDI patients and *C. difficile* carriers (non-HA CDI patients) from the beginning of 2010 to the end of 2013 at the hospital level. From the data, we derived two time series: CDI admission and HAI onset series, which were both indexed by date.

We evaluated the strength of the pattern through two statistics: confidence and significance that are commonly used in association rule analysis (Zhang & Zhang, 2002). Association rule analysis is primarily used in the mining of transaction data to discover the association of different items and has been successfully adopted to discover patterns in the time series data (G. Das, Lin, Mannila, Renganathan, & Smyth, 1998). In our context, confidence is defined as the ratio between the occurrence of the pattern and the occurrence of the admission of *C. difficile* carriers. Significance is defined as the ratio between the occurrence of the pattern and the occurrence HAI patients.

The strength of the pattern is evaluated with different time windows, from one day to thirty days. The choice of thirty days as the upper limit is determined by the definition of HA CDI outbreak, which uses four weeks as the longest threshold for the declaration of outbreaks.

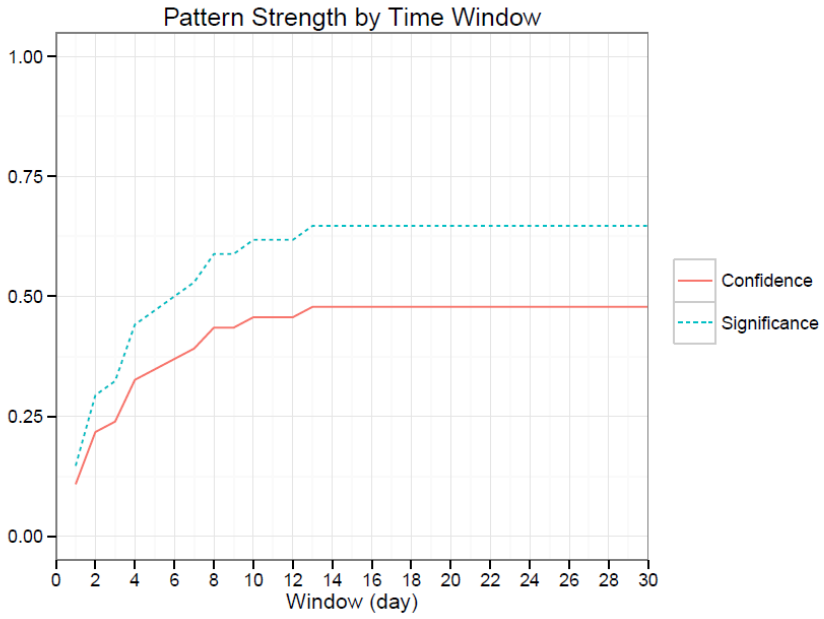


Figure 5.3 The strength of the pattern evaluated by association rule criteria

As shown in Figure 5.3, both confidence and significance initially increase within the time window, and then reach their maximum at seventeen days. The highest confidence level is 0.41. This number means we would expect 41% of the time that an admission of a *C. difficile* carrier will be followed by the

appearance of the HA CDI in the hospital within the window size. The highest significance level is 0.62. This number means 62% percent of the HA CDI occurrences are preceded by an admission of a *C. difficile* carrier within the window size. The detailed statistics is provided in Appendix F.

These numbers suggest that the pattern is potentially important. However, we have to be very cautious to interpret this importance as no epidemiology data is available to support the causal relationship between the admission cases and the following HAI cases. However, there is no evidence to reject this relationship neither.

Since this study was exploratory, we decided to assume the existence of this causal relationship and probe the situation further. This led to two additional research questions. Why some admissions lead to the appearance of HAI and some will not? Can we predict the occurrence? As our major focus of this study was to use network analysis to explore the relationship between the transmission of HAI and network statistics, we explored the two questions from a network statistics perspective.

5.3 Network analysis of inpatient transfer

5.3.1 The construction of the network and the data source

To understand the pattern for the admission of *C. difficile* carriers and the appearance of HA CDI, we obtained the inpatient transfer data during the period from the beginning of 2010 to the end of 2013 for the whole hospital site. We then constructed the transfer network from this data set. The vertices of this network are the wards in this hospital site. A directed edge is constructed between two wards whenever a patient is transferred from one ward to another. Therefore, more than one edge could exist between two wards. Whenever this happened, we simplified the network by just allowing one edge to exist but increasing the weight of the existing edge between the two wards. In addition to the normal vertices that are corresponding to the physical wards in the hospitals, we created two special vertices: admission

(ADM) and discharge (DIS) to deal with the admission and discharge of patients. The constructed transfer network is also a dynamic network. The edges are only “on” during the date when the transfers happen.

5.3.2 The predictive modelling

5.3.2.1 The extraction of network predictors

The statistics extracted from the network analysis fall into two groups. The first group reflects how busy the hospital is. The statistics collected in this group include the number of admissions, number of discharges, and number of transfers. The second group statistics are the common properties of networks, which may reflect the connectivity. These statistics include the average in/out degrees, degree density, size of giant component, diameter, average shortest path length, and cluster coefficient (Kolaczyk, 2014). Figure 5.4 shows the time series of the statistics generated from the dynamic network.

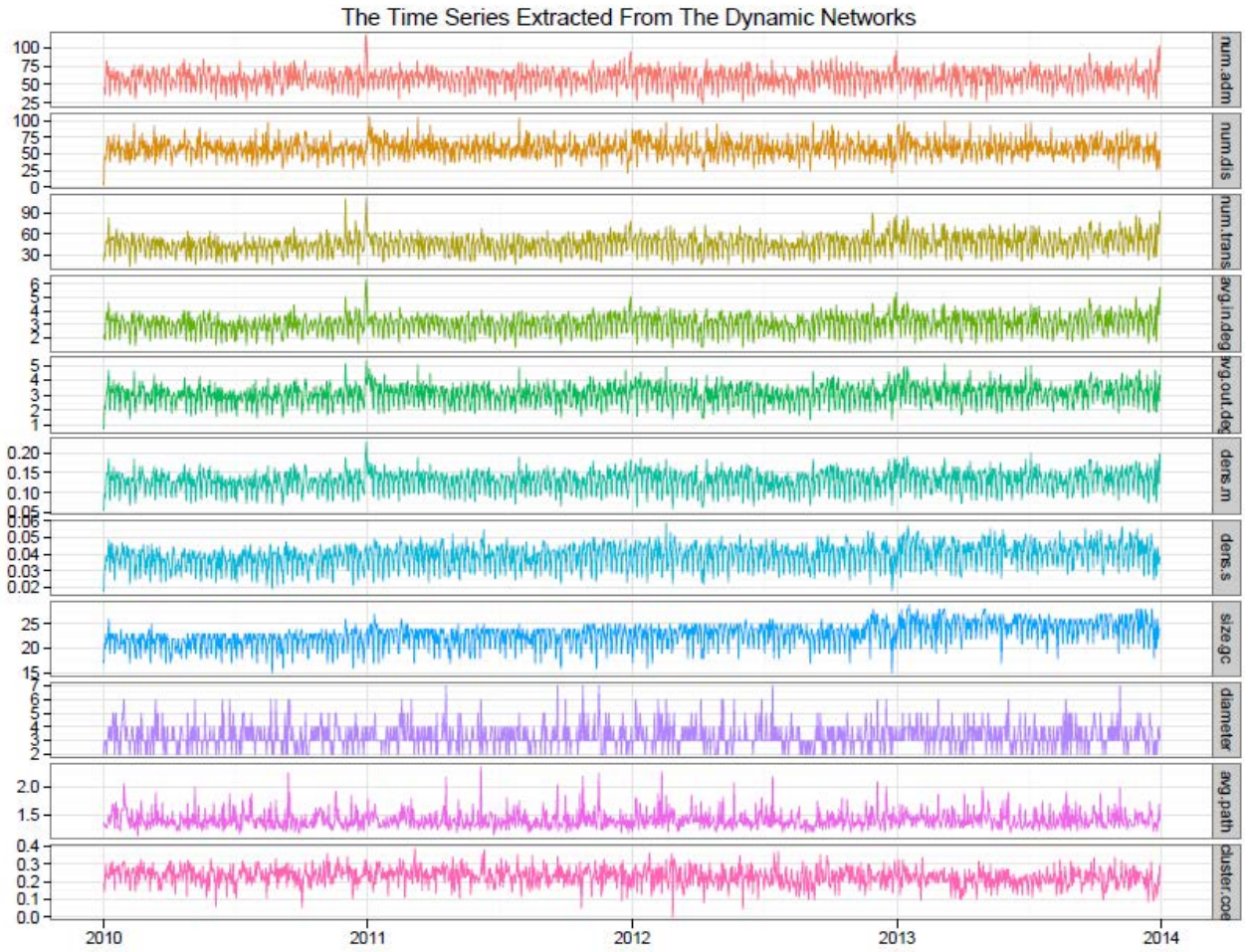


Figure 5.4 Time series of transfer network statistics

In order to answer the two questions proposed at the end of section 5.3, we identified three different time periods: 1) transmission (TRANS) period; 2) non-transmission (NOTRANS) period; and 3) no-admission (NOADM) period. The TRANS period is when the admission of a *C. difficile* carrier is followed by the appearance of HA CDI within the time window. The NOTRANS period is when the admission of a *C. difficile* carrier is not followed by the appearance of HA CDI within the time window. The NOADM period is the period when there is no admission of a *C. difficile* carrier. The classification of the period is determined by the window size used. As a demonstration, Figure 7 shows the change of cluster

coefficient during different periods in the year of 2013. The window size used for the definition of periods is 17 days.

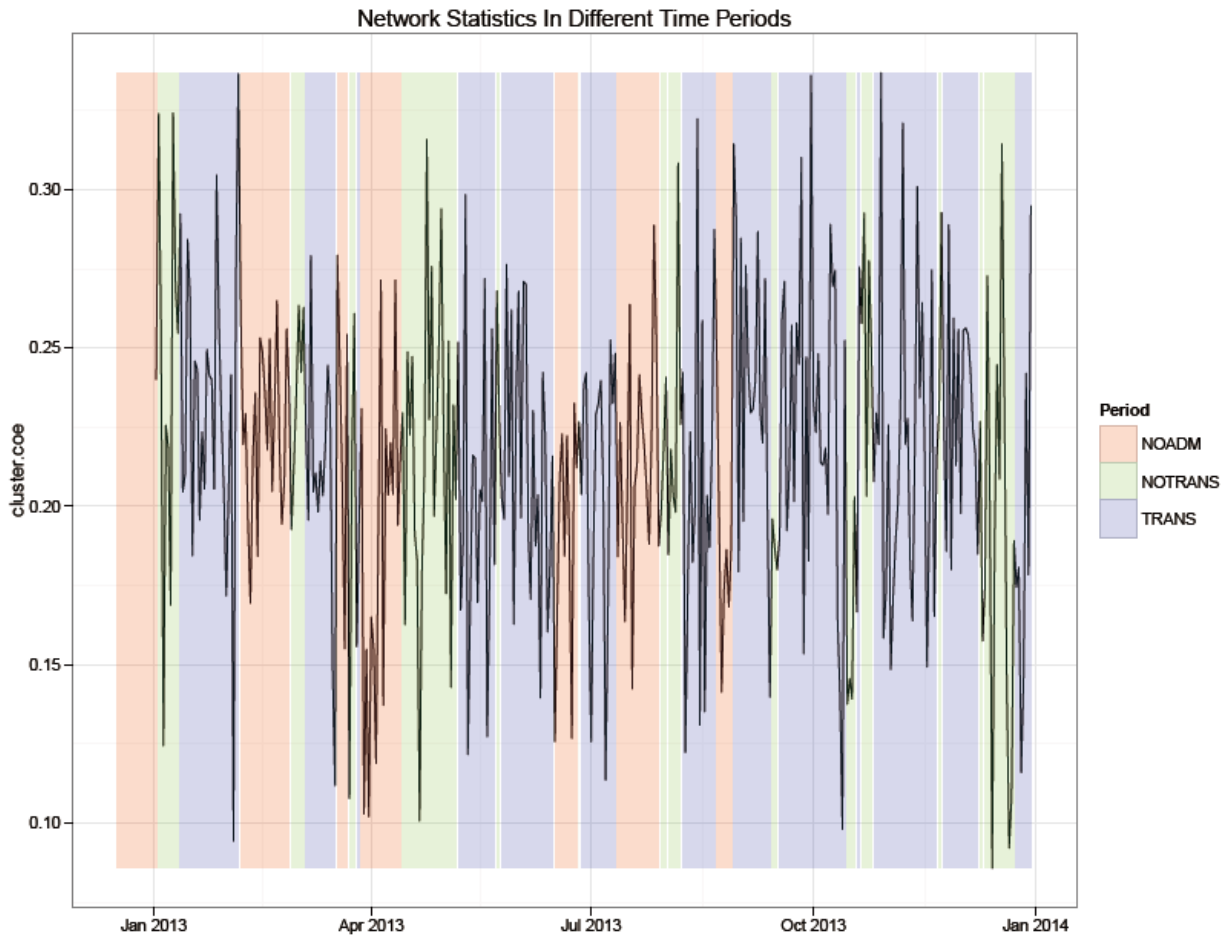


Figure 5.5 Cluster Coefficient change with time period background

5.3.2.2 The predictive models and their performance

We then built classification models to predict the type of the periods based on the summary statistics extracted from the networks corresponding to these periods. We tested the predictive performance by two classification algorithms (one linear and one non-linear) on this data set, including logistic regression (LG) and support vector machine (SVM) (Kuhn & Johnson, 2013). We split the data into a training set

(70%) and a testing set (30%) for each window size. Data were resampled thirty times to build the model and to test the performance for each of the data set generated by different window size. Figure 5.4 is the performance achieved by the predictive models on the testing set and training set by the two algorithms. As we can see from the figure, SVM does consistently better than logistic regression for each window size on the training sets. However, it suffers from over-fitting, as its performance is not much better on the testing data set than logistic regression. The best performance achieve on average on the testing data in term of AUC is when the window size is six days and the AUC is 0.75. For small and large window size, the class distribution (i.e. TRANS vs NOTRANS) are extremely imbalanced. This imbalance might be main reason for the overfitting for SVM model at these window sizes.

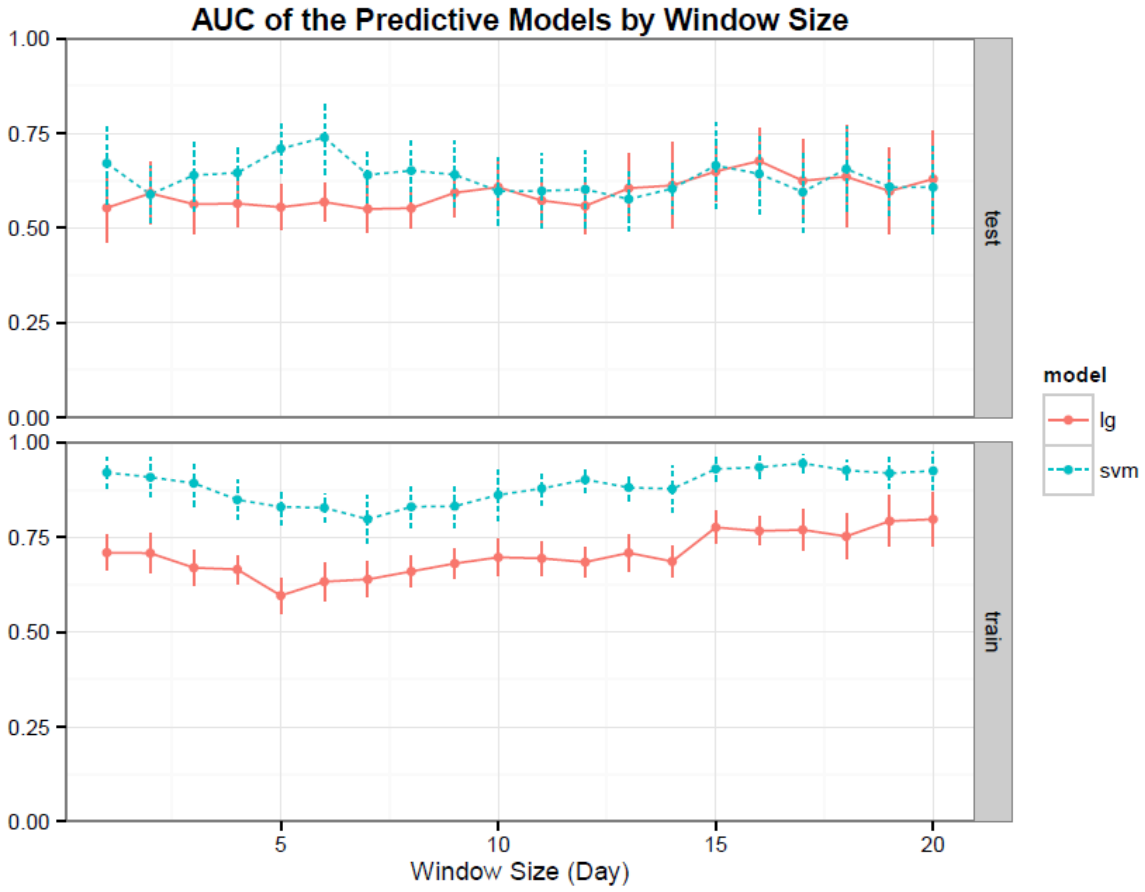


Figure 5.6 Performance of the predictive models

5.3.2.3 Variable importance

In an exploratory analysis, it is important to know what variables have a strong relationship with the outcome. To investigate this, we computed the variable importance based on the two fitted models for a window size of six days. The definition of model based variable importance often varies with the model used (Kuhn & Johnson, 2013). Therefore, relative importance was computed for comparison. Figure 5 shows the results from the SVM and LG models. The top two variables agreed to by both models are ***cluster.coe*** (cluster coefficient) and ***num.trans*** (number of transfers). It is reasonable as both of them indicate higher chance of transmission of the bacteria/spores.

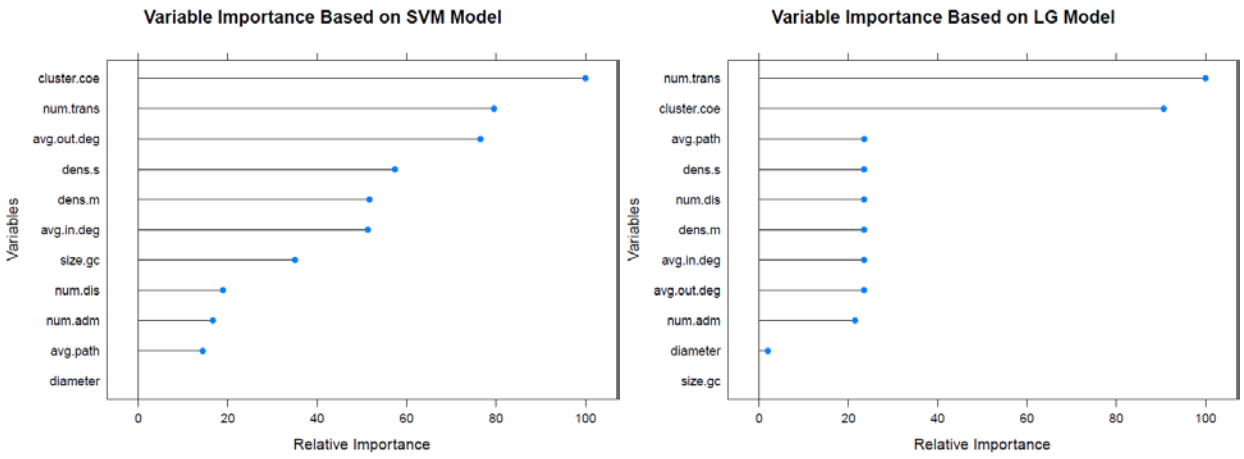


Figure 5.7 Variable importance computed from the predictive models

Cluster coefficient measures the tendency for the nodes to cluster together. In our context, a higher cluster coefficient means the higher chance of the wards in the hospital being connected through the transfer of patients. Therefore, it indicates a higher chance of the transmission.

The number of transfers metric measures how frequently the transfer of patients occurs in the hospital. The larger this number is, the more frequently patients are transferred. There are two indications from a larger transfer number. First, it indicates the hospital is busier, which might cause stress for healthcare workers. The stress might lead to less compliance of certain conducts, such as hand hygiene and

environmental cleaning standards, which might increase the chance of transmission of the bacteria (Cimiotti, Aiken, Sloane, & Wu, 2012). Second, a larger transfer number also implies the increased chance of contact with other entities in the hospital, which also increases the chance of transmission of the bacteria. Thirdly, increased transfers commonly occur in winter months when antibiotic use for respiratory infections is relatively increased, resulting in more "at risk" patients for CDI.

The implication suggested by the two important variables is that limiting patient transfer in both space and frequency dimension might be a good strategy for containing the transmission of HA CDI infection. We might consider two possible ways to achieve this goal suggested by the two variables: 1) improve patient flow design so that inter-ward transfer (i.e., reduce cluster coefficient) could be reduced (Hendrich & Lee, 2004); and 2) increase bed capacity so that patient transfer might be reduced because of bed blocking resulted from high occupancy rate (Krall, O'Connor, & Maercks, 2009).

5.4 Discussion and conclusion

Our analysis on the networks of outbreak wards showed that the common network statistics have almost no correlation with the transmission of HA CDI. However, the insights gained via the animation of these two networks, allowed us to discover that the occurrence of HA CDI is correlated with the admission of *C. difficile* carriers in the ward. This was despite having evidence that was not very strong due to the small sample size. The extended analysis on the hospital wide transfer network showed that the observed pattern in the hospital wide context is relatively strong. The predictive models based on the statistics extracted from the transfer network appear to offer good performance when predicting such a pattern. The variable importance analysis of the predictive models suggests that the cluster coefficient of the transfer network and transfer frequency were the two most important variables for the prediction.

The significance and contribution of this chapter's study is suggested to be threefold. First, we found that our occurrence pattern suggesting that HA CDI is correlated with the admission of *C. difficile* carriers

from the two types of network supports precious studies (Eyre, Cule, et al., 2013; Lanzas et al., 2011; Walker et al., 2012) on the transmission of HA CDI. These studies had suggested that HA CDI might be caused by admission of *C. difficile* carriers other than the bacteria circulating in the ward. Our study adds new evidence to support this suggestion. The practical meaning of this strong indication is that patient screening might be an effective way to prevent HA CDI outbreaks. Second, the use of time series mining methods and classification models on the analysis of dynamic network for the successful prediction of HA CDI transmissions provides a potentially new way of looking at HAI transmission and analyzing dynamic networks. The methodology could possibly be extended for the analysis of similar situations in other fields, where dynamic network modelling could be applied. Third, the two important variables obtained from the variable importance analysis of the predictive models also suggested new insights for the control of HA CDI transmission. Limiting patient transfer might be a fruitful strategy to control HA CDI transmission.

As a preliminary and exploratory study on HA CDI, there are several limitations. First, the analysis of outbreak ward network has a small sample size. The conclusion derived from this analysis requires further testing with more data and a variety of hospital sites. Second, although the analysis of the transfer network suggests that the network statistics have good performance on the prediction of HA CDI transmission, the prediction itself does not imply any causal relationship. We must be very careful about the interpretation of this result. With the available data we have, we are not able to neither reject nor accept this relationship. It would be interesting to examine this relationship through some biologic data such as strain typing results. Third, the modelling and resulting prediction is not at the ward level. As the prediction does not tell which ward will have HA CDI, further analysis is needed to make inference at the detailed locations.

In conclusion, our exploratory analysis of the two types of networks regarding HA CDI transmission demonstrates that network statistics have the potential to be good predictors for the transmission of HA CDI. Also, the constructed predictive model should be relatively easy to implement in practice and might

be useful for practitioners in the HAI prevention field. Finally, our analysis of HAI dynamic networks that combines time series mining and predictive modeling enriches the dynamic network analysis literature with an innovative methodology.

Chapter 6. Predicting hospital acquired *Clostridium difficile* infection using machine learning

6.1 Introduction

In this chapter, we explore the potential of using multiple predictive models to identify HA CDI in the hospital. This specific study was motivated by three observations as suggested by the review in Chapter 3, which include:1) there has been a changing epidemiology of HA CDI in recent years; 2) the number of studies in the literature on predictive modeling of HA CDI is small (i.e., six) ; and 3) these studies often uses one single method.

Although it is possible that one specific model could have superior performance under the condition that the intrinsic structure of the modeling subject is well understood and captured by the specific model, often multiple methods modelling approach has a more robust and better performance. Literature in machine learning has shown that using different methods together produce much better results (Dietterich, 2000; Pirracchio et al., 2015), as different models have different assumptions and capture different structures. Therefore, in this chapter, a range of machine learning predictive models were utilized to perform the prediction on the data set collected from the hospital under the guidance from the hospital`s infectious disease control unit. We analyzed eight algorithms in three categories of models. Each of the methods has strengths and weaknesses with respect to exploiting data patterns and information in the dataset. Using the insights obtained from the eight algorithms and their analysis, we constructed a ‘super learner’ model that combined the eight models to provide better coverage of the data characteristics.

6.2 Methods

6.2.1 Data collection and feature engineering

A case control study (Breslow, 1996) was designed for this analysis. We first identified the HA CDI patients from the infection control database during the period from January, 2010 to December, 2013. For each HA CDI patient, the control cases included all the patients who were admitted to the same ward as the HA CDI patient during a two month period starting from one month before the admission date of the HA CDI patient and ending at one month later of the admission date of the HA CDI patient. The idea behind the selecting of controls was that we want to make sure that all the selected patients were exposed to a similar hospital environment in terms of transmission, so that the impact of environment could be ruled out (or reduced), since we were studying individual risk factors. According to the definition of CDI outbreak provided by provincial infectious disease advisor committee, the longest time period considered is 4 weeks (close to one month) and the space considered is at ward level. Therefore, we decided to include all the patients admitted one month before and one month after admission of CDI case as controls. In total we have 6827 patients in our data set, among which 252 are HA CDI patients. Based on the review of risk factors and guidance from the hospital's infectious disease control unit, we collected four groups of data from the EMR systems. The four groups of data were: 1) basic profile, 2) comorbidity, 3) surgical intervention, and 4) medication.

Basic profile data included patient demographics (age and gender), patient type, admission and discharge time, and recent admission history.

Comorbidity data included the diagnosis type and diagnosis code for the patients. There were ten different diagnosis types in this hospital. Possible values included: most responsible, pre-admit comorbidity, post-admit comorbidity, secondary diagnosis, service transfer diagnosis, morphology, admitting diagnosis, proxy most responsible diagnosis, external cause of injury, and newborn. The detailed explanation is

provided in Appendix G. The diagnosis coding system used by this hospital is the ICD-10 system (WHO, 2010). There are more than 20,000 different codes in this system. We created binary variables based on the first character of the code for four diagnosis types: most responsible, pre-admit comorbidity, post-admit comorbidity, and secondary diagnosis. The rest of the diagnosis information was aggregated to a numeric variable. Surgical intervention data had the information about what type of surgical intervention patients received. The coding system used by the hospital for surgical intervention is the Canadian Classification of Health Interventions (CCI), developed by the Canadian Institute for Health information (CIHI). We created binary variables based on the first two characters (the first digital number and the first alphabetic character).

We identified 181 drugs (by generic name) in the medication data set which was collected under the guidance of medical experts. The 181 drugs were grouped into 32 subgroups, including 26 groups of antibiotics, antiviral, antifungal, PPI, Immunosuppressive, Corticosteroids, and Others. The 26 groups of antibiotics were: Penicillins, Cephalosporins (first to sixth generation), Monobactams, Carbapenems, Macrolide, Lincosamides, Stretogramins, Aminoglycoside, Quinolone (first to fourth generation), Sulfonamides, Tetracyclines, Glycopeptides, Lipogylycopeptides, Oxazolidinones, Rifamycins, Polypeptides, Tuberactinomycis, and other antibiotics. All the drugs that did not fall in either the 26 groups of antibiotics or antiviral or antifungal or PPI or Immunosuppressive or Corticosteroids were classified as “Others”. We created two numeric variables for each of the drug groups to record the duration and the aggregated dose of the drug use for each patient. When computing these two numeric variables for each drug group, we only considered the amount that was taken through the oral route. The detailed coding information is provided in Appendix H.

A total of 179 predictors were created.

6.2.2 Predictive models

Different predictive algorithms make different assumptions and learn different structures about the data set (Dietterich, 2000; Polley & van der Laan, 2010). To deal with this, we used a range of predictive algorithms were employed to explore the data set. These algorithms can be grouped into three categories (Hastie et al., 2009; Kuhn & Johnson, 2013): 1) linear models; 2) nonlinear models; and 3) ensemble models.

In the linear model category, we explored two algorithms, including: 1) penalized logistic regression (LR); and 2) penalized linear discrimination analysis (LDA). In the nonlinear model group, we tested four algorithms, including: 1) neural network (NNT); 2) support vector machine (SVM); 3) k nearest neighborhood (KNN); and 4) decision tree (i.e., CART (Lewis, 2000)). In the ensemble category, we explored two algorithms: 1) random forest, and 2) boosted tree.

The whole data set is divided into training set (70%) and testing set (30%). Each model is trained with the same training data set and finally tested by the same test data set. All the 179 predictors were fitted into the models, with the hope that this procedure could create an even playing field to compare the models.

All of the models except random forest were trained through a 10-fold cross validation procedure (Kohavi, 1995). In this procedure, the training set is randomly partitioned into 10 sets of equal size. For each parameter set, the 10-fold cross validation procedure takes turn to fit models on 9 of 10 subsets, and tests on the left one (fold error). Then, the cross validation error is computed for each parameter set based on the 10 sets of fold error. At the conclusion of the procedure, the parameter set having the lowest cross validation error will be selected for the fitting of the final model. The final model is then fitted on the whole training set. There is no need for the random forest algorithm to go through the cross-validation procedure as the error of random forest is based on the out-of-bag error rate for each tree.

6.2.3 Class imbalance

In our data set, two classes (CDI patients and non-CDI patients) are in an extreme imbalance situation. Only 3.6 % patients are CDI patients. Class imbalance can have significant impacts on the performance of models (Longadge & Dongre, 2013). The literature has a few methods existing to deal with class imbalance, such as model tuning, alternating cut-offs, adjusting prior probabilities for certain models, changing case weight (in cost function), and sampling methods. To deal with the class imbalance in this study, we used the sampling method approach as the sampling methods are not limited by the models we use. This approach is also aligned with the model tuning method that is applied in the 10-fold cross-validation to select final models. There are two types of sampling methods in the literature: up sampling and down sampling. Up sampling tries to increase the number of infrequent class through simulation or imputation to match the number of frequent class, while down sampling reduces the number of frequent class to match the number of infrequent class. In order to counter the low sensitivity problem, we retrained our model by using up sampling as down sampling leads to the number of case in training set decreasing significantly. As we have many near-zero-variance variables (Kuhn, 2008) in the data set and we use 10-fold cross validation procedure to train our model, the significant decrease of size of training set will create trouble for many models, such as linear discrimination analysis.

6.3 Results

6.3.1 Prediction performance comparison

Table 2 summarizes the performance (AUC) of the 10 models on both the training data set and the testing data set. All of the algorithms had high performance on the training set (i.e., the lowest is 0.91). However, all of the algorithms suffered from overfitting. The four nonlinear algorithms were the ones that overfitted the data the most. The highest performance on the test data set was achieved by the linear discrimination

model method at 0.89. Compared with performance of the models in literature, this result is better than average. However, we must be cautious about this comparison, as the study settings are different.

Table 6.1 Performance measures of the predictive models

Model		**Description**	**AUC**	
			Training set	**Testing set**
Liner models	LR	logistic regression	0.91	0.82
	LDA	linear discrimination analysis	0.93	0.89
Nonlinear models	NNT	Neural network model	1	0.76
	SVM	Support vector machine	0.99	0.69
	KNN	K-nearest neighbor	1	0.57
	CART	Decision tree (CART algorithm)	0.95	0.75
Ensemble models	RF	Random forest	1	0.81
	BT	Boosted tree	1	0.80

Figure 6.1 is the comparison of the receiver operating characteristic curves (Hanley & McNeil, 1982) from the predictive models.

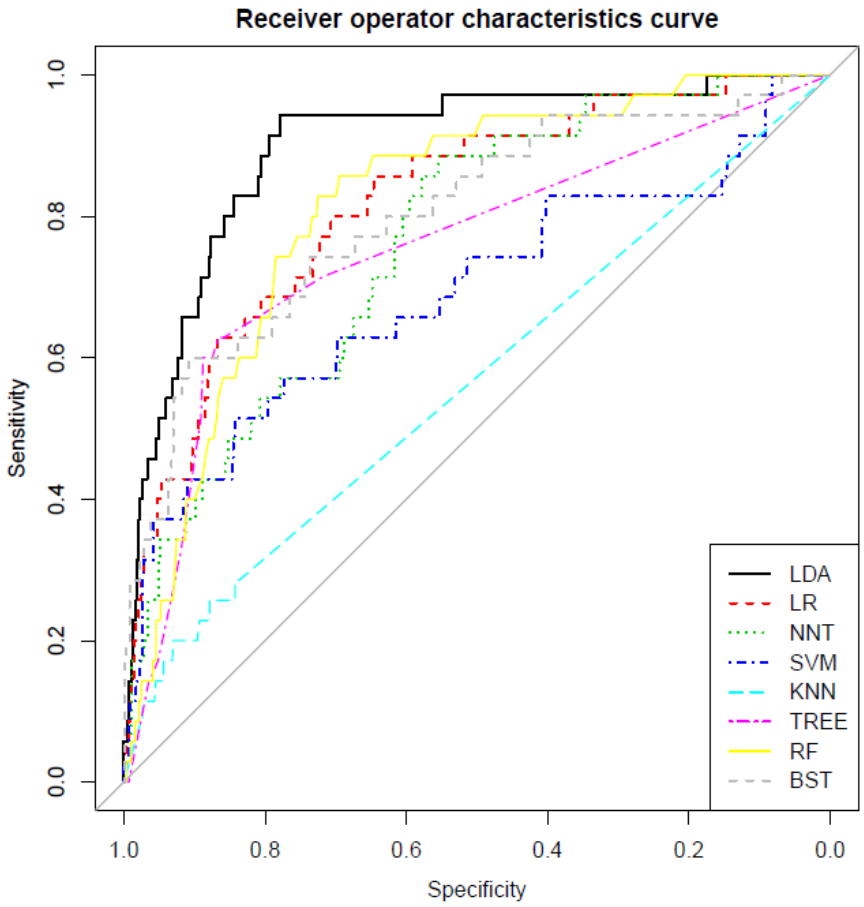


Figure 6.1 Comparison of the ROC converses of the predictive models

As shown in the figure, no curve fully dominates all of the others, although the majority of the frontier is formed by the curve of the linear discrimination model. This indicates that a super learner (Polley & van der Laan, 2010) that combines all the models might be able to improve the performance. We constructed a super learner based on the eight models simply through voting. The AUC based on the test data was improved to be 0.91.

6.3.2 Predictor importance

It is often a goal when building models to keep them simple, as simple models can focus on the key factors and are also likely to be more interpretable. Another reason for preferring simple models is that, in

some situations the collection of data is costly and simple models can be more efficient to work with. Therefore, it is important to evaluate the importance of predictors and identify the factors to focus upon.

Predictor importance can be measured in two ways: model-specific and model non-specific. The advantage of using model-specific approach is that the importance of the predictors is tied to the model's performance (Kuhn, 2008). Intuitively, the important predictors suggested by high performance models will have higher impact.

Therefore, we computed the predictor importance from the linear discrimination model (Kuhn, 2012), as this model has the highest performance, and its ROC curves almost dominates other models' as suggest in Figure 6.1.

Table 6.2 is a list of the most important 20 variables suggested by the linear discrimination model ordered by their relative importance. The 20 variables include five variables related to drug exposure, seven diagnosis codes, three therapeutic intervention codes, three admission times, comorbidity (DIAF_ALL), and age. The top three most important three variables are comorbidity (DIAF_ALL), exposure to PPI drug, and age, and they have larger distance from the other variables in the table. These three variables are widely mentioned in the literature as suggested by Table 3.1.

Table 6.2 Top 20 important variables ranked by linear discrimination model

Rank	Variable	Description	Relative Importance
1	DIAG_ALL	Number of diagnosis Code	100
2	DRUG_PPI_DURATIO N	Exposure duration to PPI drugs	74
3	AGE	Patient age	64
4	DRUG_Cephalosporins 1_DOSE	Exposure dose to first generation Cephalosporins antibiotics	55
5	DRUG_Cephalosporins 1_DURATION	Exposure duration to first generation Cephalosporins antibiotics	53
6	M_ICD10_MR	Diseases of the musculoskeletal system and connective tissue (most responsible)	53
7	CCI_1V	Therapeutic interventions on the musculoskeletal system	51
8	K_ICD10_MR	Diseases of the digestive system (most responsible)	50
9	C_ICD10_MR	Neoplasms (most responsible)	46
10	CCI_1R	Therapeutic interventions on the genitourinary system	46
11	ADMIT_JULY	Admit time in July	45
12	S_ICD10_MR	Injury, poisoning and certain other consequences of external causes (most responsible)	44
13	ABD_INTVN	Abdominal intervention	43
14	ADMIT_MAY	Admit time in May	43
15	ADMIT_NOV	Admit time in November	43
16	R_ICD10_MR	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified (most responsible)	42
17	N_ICD10_MR	Diseases of the genitourinary system (most responsible)	42
18	F_ICD10_MR	Mental and behavioural disorders (most responsible)	41
19	DRUG_Lincosamides_ DURATION	Exposure duration to first generation Lincosamides antibiotics	40
20	DRUG_Lincosamides_ DOSE	Exposure dose to first generation Lincosamides antibiotics	40

The most important variable is the number of diagnosis codes (DIAG_ALL). A large number of diagnosis codes associated with a patient implies that the patient might have many comorbidity factors. This result suggests that sicker patients are more vulnerable to CDI.

The second most important variable suggested by the model is the exposure time to PPI. Although PPI has been reported as a risk factor for CDI, a definite association between PPI and CDI is still not

confirmed (Biswal, 2014). Moreover, PPIs are a chronic and over prescribed medication and tied to LOS. Therefore, it is possible that PPIs are just a surrogate for something else.

The specific antibiotic drugs suggested by the model are first generation Cephalosporins and Lincosamides. These two classes of antibiotics are often mentioned in the literature (Owens et al., 2008). However, Cephalosporins are the workhorse drugs used in this hospital. Therefore, its high rank might be due to the frequency of use other than strong intrinsic risk.

Half of the variables in the top 20 list related to the diagnosis codes and therapeutic intervention codes. As suggested by the literature, it is very understandable that diseases of digestive system (K_ICD10_MR), neoplasms (C_ICD10_MR), abdominal intervention (ABD_INTVN) are in the list. It is also understandable that diseases and interventions related to the musculoskeletal system (K_ICD10_MR and CCI_1V) and diseases and interventions of genitourinary systems (N_ICD10_MR and CCI_1R) are in the list. For musculoskeletal system diseases and interventions, there are two possible reasons. First, infections of the joint/bone or auto-immune situations may need to use steroids (oral) drugs, which might increases the risk for CDI. Second, in the hospital, first generation Cephalosporins are used as prophylaxis for musculoskeletal system surgeries. As discussed in previous paragraph, Cephalosporins might be high risk antibiotics for CDI. For diseases and interventions of genitourinary systems, it might be due to the fact that in this hospital first generation Cephalosporins are used to treat the common infections (i.e., urinary) in these systems. However, it is not clear why diseases of mental and behavioural disorders (F_ICD10_MR) are also in the list.

There are also three variables related to the admission time of the patients. This result suggests that there might be seasonality associated with the occurrence of HA CDIs in the hospital.

6.4 Discussion and conclusions

Our analysis on the prediction of HA CDI through eight different modeling methods showed that the linear discrimination model had the strongest predictive power for the prediction on our data set. The comparison of ROC curves generated by different models reveals that the models as a group can capture different structure in the data set while individual models fail to fully capture the structure. The super learner that combined all the models together exploited this situation and improved the prediction performance.

The important variable given by the linear discrimination model had a reasonably good agreement with the literature. However, it also pointed out some issues to look at, such as the association of CDI with mental and behavioral disorders diseases and seasonality.

The potential significance and contribution of this chapter's research is twofold. First, the relative high performance of our predictive models suggests that the models could be possibly refined further and deployed in practice to help the prevention of CDI in the hospital. Second, logistic regression is the traditional approach in medical literature of studying risk factors and constructing predictive models. Our analysis suggests that it is beneficial to introduce additional predictive modeling methods to improve prediction performance and to reveal new insights in the data set. The study showed that while most of the eight methods provided a relatively high performance when used alone, the super learner model was capable of improving the performance.

The major limitation of this research is that this study is a retrospective analysis and we are able to collect all the data at the same time. In reality, data in healthcare system is generated piece by piece. For real time use, the model would have to be modified and further research is required on this topic. Another major limitation is that our analysis is based on available data (and how they were recorded) from one hospital and the results are affected by the characteristics of patients from this hospital. Replication and additional hospital data sets should be considered in future research. In conclusion, our exploratory

research using multiple predictive modeling methods for the prediction of CDI shows the potential benefit of introducing and combining multiple methods to improve prediction performance and to reveal new insights in the risk factors. It is recommended that this multiple modeling methods approach might be considered for future studies on the predictive model construction and risk factor analysis.

Chapter 7. Thesis conclusions

7.1 Overview

The thesis set out to investigate predictive models for HA CDI and its outbreaks in hospital settings as a strategy to prevent the spread of the disease. The fundamental issues for the development of a predictive model include the understanding of the subject, the selection of appropriate modeling methods and predictors, and the collection of relevant data. We consider our research as being exploratory and preliminary because the understanding of HA CDI is still developing and a solid foundation of theory and empirical results has yet to be created. While it is possible to initiate the research on modeling methods, predictors, and data, there are associated risks and limitations when the field is relatively young. This view of HA CDI research is supported by the relatively small number of robust descriptive, normative or predictive research findings found in the literature. The situation is further complicated by the two levels of modeling required – population and individual level.

The thesis explored a number of fundamental issues, including:

1. methodologies for prediction at the population level,
2. methodologies for prediction at the individual level,
3. transmission mechanisms of HA CDI at the ward level,
4. predictors and related data at the population level,
5. predictors and related data at the individual level.

These explorations have generated insights about the modeling of HA CDI.

7.2 Summary of findings and contributions

The findings and contributions related to the research are:

1. **Methodologies for prediction at the population level**

    a. Chapter 2, the literature review of the modeling at population level, provided an overview of the modeling approaches available in the literature dealing with the outbreak prediction for HAIs.

        i. The reviewed literature showed that the outbreak prediction methods for HAIs appear to have been strongly influenced by the methods for community acquired infections (CAIs) in both “what-if” and “anomies-detection” frameworks.

        ii. We found that agent based models and network models appear to be increasing in popularity for the modeling of HAIs due to the growing availability of electronic data in hospitals.

        iii. These observations led to the studies in Chapter 4 and Chapter 5.

    b. Chapter 4, the use of the agent based simulation methodology resulted in what can be possibly considered a reasonable prototype for the prediction of HA CDI transmission at the ward level. With suitable modifications, such as real-time feeding of parameter values, it might be possible to provide predictions in hospital settings.

    c. Chapter 5, the analysis of dynamic hospital transfer network, appears to offer a new way of analyzing and predicting the transmission of HA CDI.

2. **Methodologies for prediction at the individual level**

a. Chapter 3 and Appendix A (literature review of modeling at the individual level), summarized the current state of predictive modeling of HA CDI and HAI in general at individual level.

    i. The review revealed that the knowledge base pertaining to the risk factors for the development of HA CDI is evolving and is not well understood.

    ii. The review also revealed that logistic regression appears to be the prevalent methodology used and that most of literature uses one algorithm for the risk factor analysis and the identification of patient at risk.

    iii. These observations inspired the study in Chapter 6.

b. Chapter 6, the multiple algorithms modelling approach for the study of HA CDI at individual level, appears to improve the prediction performance.

3. Transmission mechanisms of HA CDI at the ward level

a. Chapter 4, the agent based simulation model, clarified some of the issues regarding the transmission of HA CDI at ward level.

    i. The model suggests that the newly imported *C. difficile* bacteria are possibly responsible for the genetic diversity in the endemic setting reported in the literature.

    ii. The model also suggests that environment objects other than HCWs might be the major sources for *C. difficile.*

    iii. These findings, together with the results from the model exploration, suggest several potential prevention policies for HA CDI (see the details in Chapter 4).

4. Predictors and related data at the population level

a. Chapter 4, with the construction of the simulation model, collected and tested many relevant parameters regarding transmission of HA CDI. The sensitivity analysis and model exploration result helped identify the potentially important parameters for the transmission. These results could be possibly used to guide the collection of data in practice.

b. Chapter 5, with the construction of the dynamic networks, demonstrated how network statistics could be possibly extracted for the prediction of HA CDI transmission. These statistics could be potentially used to guide the collection of data in practice, and could also be used to build other predictive models or systems.

5. Predictors and related data at the individual level

a. Chapter 3 reviewed the potentially important risk factors for HA CDI found in the literature.

b. Chapter 6 demonstrated the transformation of data available from various information systems in a real hospital to features used by machine learning algorithms for prediction.

c. These results could be possibly used to guide the construction of other predictive models.

7.3 Recommendations

Constrained by the lack of knowledge around CDI transmission and the changing epidemiology of CDI, we developed three types of models trying to decode some of the mysteries. Although the modeling approach is exploratory and the results are preliminary, many lessons have been learned. This section summarizes these lessons as recommendations for both infection control practitioners and modellers.

7.3.1 Recommendations for infection control practitioners

The motivation behind this thesis is to help infection control practitioners to prevent HA CDI in practice. The Recommendations we informed from this thesis for infection control practitioners in practice include HA CDI prevention policy recommendations and model usage recommendations.

7.3.1.1 Policy recommendations

The exploration of the simulation parameters suggests that the prevention of HA CDI might benefit from the following policies:

1. Admission screening, as HA CDI are mainly caused by the bacteria imported from outside of the ward;
2. Setting high standard for the cleaning effectiveness of ordinary-clean rate, as it affects multiple environment objects and therefore has a wider impact;
3. Improving after-visit hand hygiene compliance rate for HCWs, especially for physicians as they have a wide contact network;
4. Prompting antibiotic stewardship, as the risk of infection is highly sensitive to antibiotic use.

Network analysis has suggested cluster coefficient and number of transfers are the two most important variables for the possible transmission suggested. The implication suggested by the two important variables is that limiting patient transfer in both space and frequency dimension might be a good strategy for containing the transmission of HA CDI infection. We might consider two possible ways to achieve this goal suggested by the two variables:

5. Improving patient flow design so that inter-ward transfer (i.e., reduce cluster coefficient) could be reduced;

6. Increasing bed capacity so that patient transfer might be reduced because of bed blocking resulted from high occupancy rate.

The machine learning study has suggested a list of important variables for the identification of HA CDI victims (Table 6.2), which plausibly match the characteristics of the patients and the medical practices in the studied organization. Therefore, we recommend:

7. Monitoring patients presented with characteristics that may increase their infection risks;

8. Monitoring medical intervention that may lead to high risk for HA CDI.

7.3.1.2 Model usage recommendations

Simulation model

1. We believe the simulation model constructed in this study has a relatively comprehensive capture of the important parameters involved the HA CDI transmission. In our model, approximately 30 parameters are involved. This relative comprehensive capture allows us to investigate a system at a fine degree of granularity. Therefore, the simulation model provides Infection control practitioners with a platform to look at different effects of the model easily.

2. If supplied with real time data about the parameter, the simulation model can be easily turned into a prediction tool to be used in practice foretell the probability of a transmission or outbreak at ward level.

Network model

3. The network model expands the modeling scope from ward level to hospital level. It provides a new way of looking at hospital CDI transmission. With further validation, the model might be used in practice for the prediction CDI transmission at the hospital level.

Machine learning models

4. Compared with the models in the literature, the machine learning models constructed in thesis have relative good performance. If the model further augmented with quality data such as lab test data, the models can be easily deployed as a complementary tool in practice to assist infection control practitioners for identification of HA CDI patient.

7.3.2 Recommendations for modellers

Simulation model

1. Although agent based simulation is an effective tool for the modeling with complex systems, such as the transmission of HA CDI. The bottom up modeling demands a great amount of data at a very detailed level. The computation cost of agent based simulation is very high, especially when the calibration of the model involves many parameters unknown or with limited knowledge. This usually means that the calibration process will require many runs. Therefore, it is not recommended to use agent based simulation in the situation where data is scarce and where time is a constrain for the modeling.

2. Agent based simulation is still a relative new modeling approach. There are not many software tools or platforms to support the full life cycle of modeling and analysis. We had to write a substantial amount of code for the model and the analysis, and assemble various platforms together to perform the work. It is recommended that software engineering approach should be used to manage the modelling and development activities when agent based simulation is considered.

Network model

3. Through the innovative dynamic network analysis approach, we are able to discover and predict some interesting pattern regarding HA CDI transmission. It is recommended that modellers should further explore this field.

4. The construction of the network demands a lot of a data. The data availability might be constrained the IT systems. The extraction of the data from the IT systems may add extra load for the system which may have impact on the real time operations. Therefore, it is recommended that the modellers should plan carefully with the IT people when similar study is considered. This recommendation also applies to the machine learning study, as it has similar data issue in practice.

Machine learning model

5. As there are many software packages that cover a wide range of machine learning algorithms, a modeller can be equipped with very sophisticated algorithms in a few lines of code. Therefore, it is recommended that modellers spend more time understand of the risk factors and obtain quality data to improve the performance of the models.

7.4 Future research

First, this study is based on one single site. Further validation of the results from this thesis are needed, especially the results from network analysis and machine learning study.

Second, it is possible to address some of the limitations associated with the three major studies in the thesis.

1. The agent based simulation model has provided insights into the sources of HA CDI and the research has suggested ways to improve infection control. However, extensions can be made to the model to make it more accurate and practical in collaboration with infection control experts.

One possible extension is to expand the setting to the whole hospital to further verify and validate the results. Another extension is to model the entities at a higher granularity such that prevention policies can achieve precise targeting. For example, there are only two types of HCWs modelled in the current research: physicians and nurse. In reality, there are many types of HCWs, and physicians and nurses can be further classified into sub-categories. In an extended model, each of the categories could have a different contact pattern with the patients. The same type of extension could be made for the environment objects; creating sub-classifications. However, it might be hard to obtain the detailed data at these higher levels of granularity. Further research can proceed in this direction if the data is available.

2. The network analysis suggests two potential directions for future research. First, it is possible to consider the extraction of new network statistics for the analysis. In this study, the network statistics used for the analysis are relatively simple statistics. There might be better statistics that can relate network dynamics to the transmission of infectious diseases, as indicated by the predictive power of the statistics on the prediction of the HA CDI occurrence pattern in this study. Second, research into other methods for the analysis of HAI transmission dynamic network can be undertaken. The analysis of dynamic networks is a developing field and is enjoying increasing popularity (Kolaczyk & Csárdi, 2014). However, not many analysis methods are available yet to support the analysis of this type of network. In this study, we combined time series data mining methods and predictive modeling methods for the analysis of the two types of dynamic networks. It could be beneficial to determine if other methods from different fields could be introduced for the analysis of dynamic networks.

3. In the machine learning study, we built the models retrospectively. In this case, all of the data is available at the time of modeling. For real-time use in a live setting, the models would have to be modified and further research is required to understand the type and extent of the modifications.

Third, with the increase accessibility of “next-generation” whole genome sequencing technology and other technology, it is likely that hospitals can have high quality data regarding the transmission of HA CDI. Therefore, Exploring different methods on these high quality data to obtain the insights of HA CDI transmission might lead to breakthroughs for the prevention of HA CDI.

Finally, the topics related to the fundamental issues can be explored further. In addition to the ongoing efforts to better understand HA CDI, the design of an information system that systematically manages the data related to HA CDI could be explored. The information strategy could also be applied to other HAIs. Another activity coupled with enhanced data management would be to create a platform supporting multiple predictive methods to build a real-time monitoring and prediction system.

Specific ideas are:

1. **Design an information system that collects data related to HAIs.** For the three studies we conducted, we encountered difficulty in extracting relevant data from the hospital information systems. Research on how to design information systems that can manage data related to the surveillance of HAI transmission is required. We believe the systems have to overcome several difficulties to take a full advantage of the information infrastructures in modern hospitals. While each of the difficulties has been researched individually, the challenge is how to combine them into a coherent framework.
    a. *System fragmentation*. Hospitals often use secondary data for the surveillance of HAIs (Sorensen, Sabroe, & OLSEN, 1996). However, data in hospitals is highly fragmented and possibly scattered in dozens of various departments, laboratories, and administrative systems that have their own legacy (Fernando & Dawson, 2009; Rada, 2007). Active research in bringing together this data has been carried on for many years and it has been shown to be very challenging.

b. *Data synthesis*. Simply bringing the data together might not be sufficient. We have to synthesize it for HAI prediction purposes. For example, it is believed in practice that the absence or reduction of staff combined with the increase of admissions will increase pressure on the staff (Cimiotti et al., 2012). Therefore, some infection prevention practices, such as hand hygiene, might be not well followed under these stressful situations and there might be an increased risk of HAI for patients. Both the staffing and the admission data are recorded in systems, but without synthesizing the data, they cannot be directly used for the predictive modeling purpose.

c. *Unstructured data*. Most of the data, especially that generated through various clinical entry systems, is presented as free-form text. Converting unstructured free-form text into structured data is still an issue for many areas of clinical decision support modeling. This includes the topic of HAIs, as the data can contain a great deal of "noise" (Ehrentraut, Tanushi, Dalianis, & Tiedemann, 2012).

d. *Data temporality*. As patients transverse different stages of their hospital stay, data values are generated piece by piece. Predictive modeling has to align the modeling purpose with the data availability at different stages of patient care. Also, appropriate mechanisms of incorporating the newly generated data have to be designed in order to make the models functional in practice.

2. **Introduce additional models to build a real time prediction system**. The literature and practice in machine learning and other field has shown that ensemble methods (i.e., combine different methods together) often have a more stable and better performance than a single method approach. Therefore, it is likely that a library of algorithms in the prediction system is desirable. The introduction of additional modeling methods can happen in two dimensions: horizontal and vertical.

a. *Horizontally*. There are relatively many algorithms for the individual level modeling. However, more algorithms are needed for the population level modeling.

b. *Vertically*. It is possible to consider combining the models at the population and individual levels to develop a two stage prediction system for the prevention of HA CDI and other HAIs. For example, it is possible to consider using the agent based and network models to evaluate the risks for outbreaks at the first stage. The second stage, machine learning model can be then used to identify the patients at risk. This approach could potentially lead to more precise targeting of individuals and environmental objects for the prevention of HAI.

7.5 Final remarks

In conclusion, it is suggested that this thesis contributes to the body of knowledge for the development of predictive models for the prediction of HA CDI and its outbreaks from five perspectives. In addition to observations about prevention methods and the potential value of network statistics, the multiple modeling methodology developed as part of the research appears to have benefits over simpler approaches. However, this thesis is at the exploratory and preliminary level and in some ways it raises more questions than it has answered. It is hoped that future research can clarify some of these uncertainties and that a predictive system can be developed for the prevention of HA CDI, as well as other HAIs.

Appendix A. Review of predictive modeling of hospital acquired infections

Abstract

Prevention of hospital acquired infections is a major concern regarding patient safety in modern hospitals. Recent work in the literature has highlighted the potential of predictive modeling in addressing this concern. However, no review paper discussing the predictive modeling of HAIs was able to be found in the literature. This Appendix offers an attempt to fill this perceived gap.

A.1 Introduction

Recent work has highlighted the potential for using predictive modeling to better understand patients' risk factors and to guide surveillance and other preventative actions (Horvitz, 2010). However, we were unable to locate a review of this emerging field. This paper attempts to summarize how predictive modeling is being applied in the field of HAI prevention. Specifically, we try to answer the following questions:

1) What are the modeling objectives in the HAI predictive modeling context?

2) What are the methods used for the predictive modeling?

3) How good is the performance of the models?

4) What are the data used for the modeling?

5) What are the modeling difficulties, if any?

A.2 Methods

A.2.1 Search strategy and information sources

We searched peer-reviewed English literature indexed in Web of Science, Scopus, and PubMed systematically between January 1st, 1990 and December 31st, 2013. Pilot search was first conducted to identify key wards. Then, a query was synthesised and used for the search in the three databases. The following search terms were used in the three databases.

TITLE: (Predict* OR detect* OR automat* OR comput*) AND TITLE: (Method* OR model* OR algorithm* OR system*) AND TOPIC: (“Hospital acquired infections” OR “Healthcare associated infections” OR “nosocomial infections”).

Only articles that have the full description of the modeling purpose, the methods or models, the performance, and the data about the predictive modeling of HAIs were selected for the statistical analysis.

A.3 Results

340 articles were identified from the three databases through the query (Figure A-1). After screening the title and abstract, we downloaded 95 full-text articles to have detailed analysis and data extraction. 45 articles were excluded based on the eligibility criteria. 50 articles were left for the statistical analysis (Table A-1).

Figure A-1 Review process flow chart

Table A-1 Publications of HAI predictive modeling from 1990 to 2013

Year	Country /Region	HAI Type	Study Setting	Reference
2012	USA	BSI*	950-bed tertiary hospital; 160 patients.	Hirsch et al., 2012
2008	Switzerland	MRSA*	365-bed university hospitals; 13,262 patients	Harbarth et al., 2008
2011	Taiwan	BSI	2400-bed university hospital; 558 patients	Su et al., 2011
2004	USA	BSI	600-bed teaching hospital; 120-bed community hospital	Trick et al., 2004
2000	France	HAI	676 ICU patients	Escolano,Golmard, Korinek, & Mallet, 2000
2013	Germany	MRSA	Tertiary center; 3091 patients	Elias et al., 2013
2003	USA	MRSA	279-bed teaching trauma center; 494 patients	Lodise, McKinnon, & Rybak, 2003
2008	Taiwan	MRSA	Two hospitals	Hsu, Lin, Chen, Liu, & Muder, 2008
2009	Taiwan	UTI*	733-bed teaching hospital; 5533 patients	Chung, Lo, Lee, Hsu, & Liu, 2009
1998	USA	CDI*	Two hospitals	Hornbuckle et al., 1998
2013	Taiwan	UTI	730-bed tertiary teaching hospital, 11251	Lo, Lee, & Liu, 2013
2011	Mexico	BSI	Two hospitals in Brazil and Mexico	Graves, Barnett, & Rosenthal, 2011
2009	France	Sepsis	12 ICUs, 2268 patients	Adrie et al., 2009
2013	USA	BSI	Two hospitals with 1950 beds in total	Al-Hasan, Lahr, Eckel-Passow, & Baddour, 2013
2005	USA	TB*	Two hospitals, 516 patients	Wisnivesky et al., 2005
2012	Netherland	Meningitis	University teaching hospital	van Mourik et al., 2012
2011	USA	CDI	Tertiary care medical center; 35,350 patients	Dubberke et al., 2011
1993	USA	HAI	Tertiary teaching hospital	Kahn, Steib, Fraser, & Dunagan, 1993
2013	Swiss	BSI	2200-bed academic medical center	Stewardson et al., 2013
2005	Brazil	SSI*	Two teaching hospital; 609 patients	de Oliveira, Ciosak, Ferraz, & Grinbaum, 2006
1999	USA	HAI	113 ICU patients	Hurr, Hawley, Czachor, Markert, & McCarthy, 1999
2013	France	SSI	University hospital, surgery unit; 168 patients	Hautemanière, Florentin, Hunter, Bresler, & Hartemann, 2013
2013	Korea	Pneumonia	Community hospital; 580 patients	Park et al., 2013
2013	USA	CDI	University hospital in Rochester	Stevens, Concannon, van Wijngaarden, & McGregor, 2013
2007	USA	CDI	900-bed tertiary care medical center	Peled et al., 2007
2010	Colombia	SSI	A trauma centre; 614 patients	Morales, Escobar, Villegas, Castaño, & Trujillo, 2011

2009	Canada	HAI	Sources from CIHI; 469349 patients	Daneman, Simor, & Redelmeier, 2009
2011	USA	UTI	413-bed university teaching hospital	Choudhuri et al., 2011
1990	USA	HAI	900-bed university teaching hospital	Broderick, Mori, Nettleman, Streed, & Wenzel, 1990
2011	Taiwan	Pneumonia	Six medical centers; 444 patients	Fang et al., 2011
2012	USA	CDI	A 360-bed community hospital	Chandra et al., 2012
2009	USA	Pneumonia	Four hospitals; 178 patients	Mirsaeidi, Peyrani, & Ramirez, 2009
2010	USA	Pneumonia	32 hospitals; 17,048 patients	Kinlin, Kirchner, Zhang, Daley, & Fisman, 2010
2013	USA	CDI	University teaching hospital;159 patients	Lee et al., 2013
2008	Italy	HAI	1850 bed tertiary care teaching hospital	Tacconelli et al., 2008
2008	Netherland	Pneumonia	One hospital, 153 patients	Visscher, Kruisheer, Schurink, Lucas, & Bonten, 2008
2011	Taiwan	HAI	800-bed university hospital; 1,367 patients;	Fang et al., 2011
2009	USA	SSI	Multiple community hospitals	Olsen et al., 2009
2013	Australia	Pneumonia	Nationally; 23,247 patients	Sanagou, Wolfe, Leder, & Reid, 2013
2012	USA	Pneumonia	National database ;1,438,035 cases	Pearl & Bar-Or, 2012
2008	Brazil	HAI	Tertiary pediatric referral hospital; 754 patients	Lopes et al., 2009
1992	USA	Pneumonia	350-bed tertiary hospital	Joshi, Localio, & Hamory, 1992
2013	France	ESBL*	650-bed teaching hospital;671 patients	Goulenok et al., 2013
2012	USA	TB	Community hospital; 315 patients	Aguiar et al., 2012
2013	France	CDI	860-bed university teaching hospital; 40 patients	Khanafer et al., 2013
2006	Demark	HAI	A Danish hospital	Leth & Møller, 2006
2008	Germany	HAI	1182-bed university teaching hospital;	Steinmann et al., 2008
2013	Canada	SSI	National data; 18,1894 cases	van Walraven & Musselman, 2013
2011	France	HAI	Three university hospitals	Proux et al., 2011
2009	USA	CDI	700-bed tertiary teaching hospital; 605 patients	Yadav et al., 2009

*_BSI - Bloodstream Infection; CDI - *Clostridium difficile* Infection; ESBL- Extended Spectrum Beta Lactamase; MRSA - Methicillin Resistant *Staphylococcus aureus*; SSI - Surgical Site Infection ; TB - Tuberculosis ; UTI - Urinary Tract Infection.

Within the 50 articles, 16 of them explicitly stated the pathogens that the paper was concerned about. Among these pathogens, *Clostridium difficile* was the most frequently studied one, flowed by MRSA. 11 (22%) of 50 articles studied HAI in general, which is an ambiguous definition. The distribution of 50 papers by the HAI type is shown in Figure A-2.

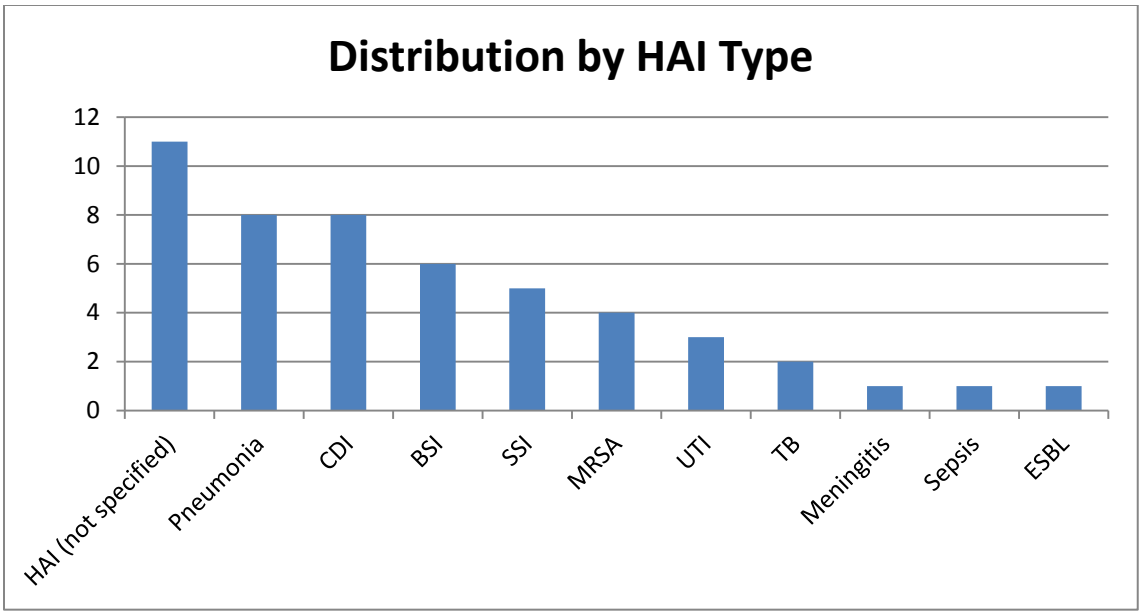


Figure A-2 Number of predictive modeling publications by HAI types

Most of studies were published after 2007. The distribution of the articles by year is shown in Figure A-3. A trend of increasing interest about this subject is suggested by this figure.

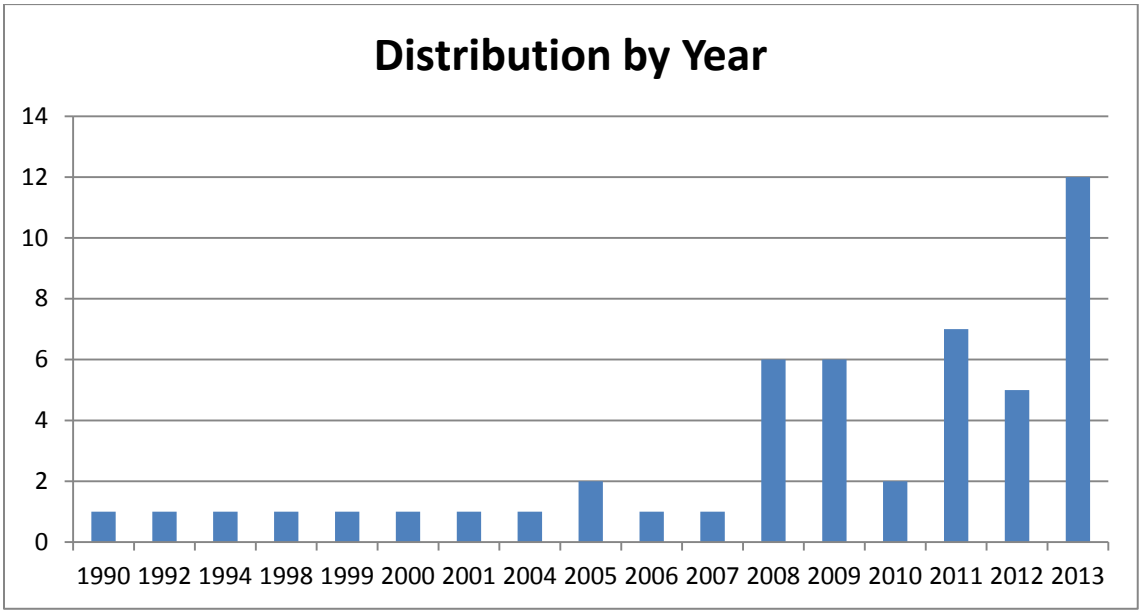


Figure A-3 Number of publications over time (1990-2013)

Most of the studies (92%) were from developed countries or regions. Studies from USA contributed to almost half (44%) of this literature, followed by France (12%) and Taiwan (12%). Figure 4 shows the distribution of the studies by country or region where they came from.

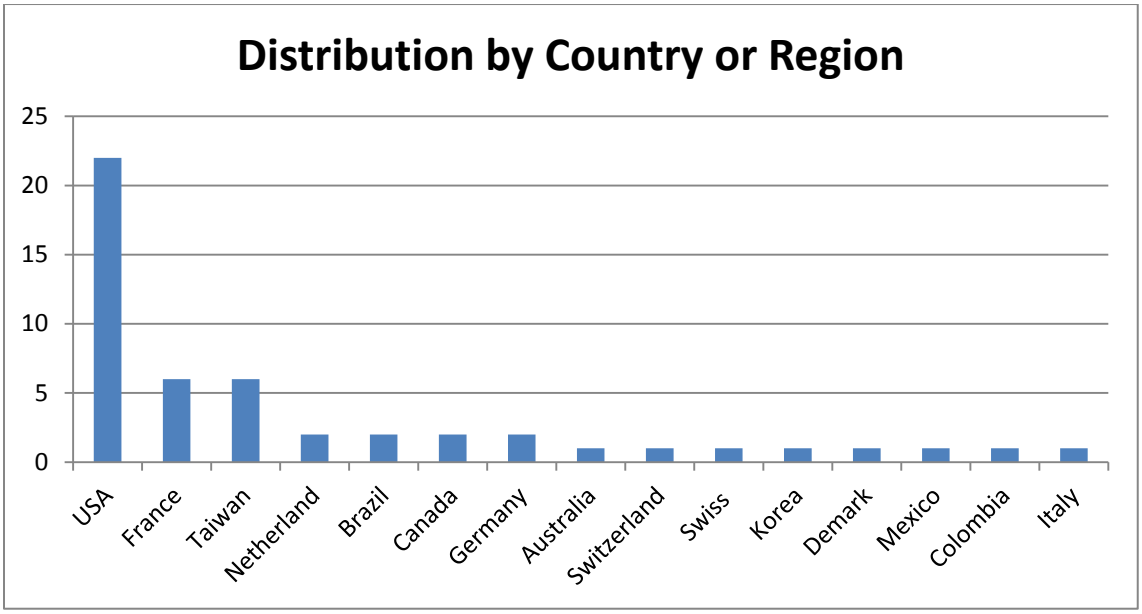


Figure A-4 Number of publication by country or region

Majority of the studies (66%) were conducted in one single institution. Most of the institutions were big university hospitals or tertiary medical centers. The studies involved two or more than two institutions were six (14%) and ten (20%) respectively. Figure A-5 is the distribution of the articles by their study setting. These numbers suggest that overall, results and observations in the literature must be carefully considered because 80% of the research is based on a very small sample size and this limits the power or strength of any statistical analysis.

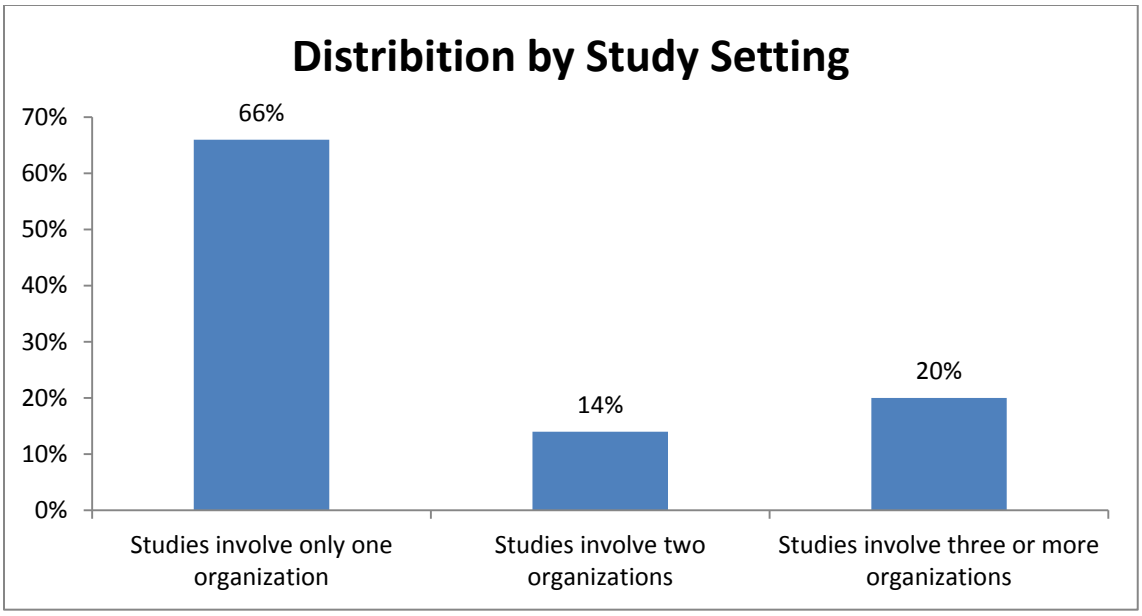


Figure A-5 Percentage of publications by study setting

A.3.1 Objectives of the predictive models

A number of objectives of the predictive modeling were identified in the literature. We summarized them in Table A-2. The following subsections are the elaborations of these objectives.

Table A-2 Objectives of HAI predictive modeling

Index	Modeling objectives	References	Count
1	Identify the patients with infections retrospectively	Broderick et al., 1990; Choudhuri et al., 2011; Daneman et al., 2009; Fang et al., 2011; Hautemanière et al., 2013; Kahn et al., 1993; Leth & Møller, 2006; Lo et al., 2013; Park et al., 2013; Sanagou et al., 2013; Steinmann et al., 2008; Trick et al., 2004; van Mourik et al., 2012	13
2	Identify patients with high risk of contracting HAIs in real time	Chandra et al., 2012; Chung et al., 2009; Dubberke et al., 2011; Graves et al., 2011; Hornbuckle et al., 1998; Joshi et al., 1992; Kinlin et al., 2010; Lee et al., 2013; Pearl & Bar-Or, 2012; Peled et al., 2007; Proux et al., 2011; Su et al., 2011; Tacconelli et al., 2008; van Walraven & Musselman, 2013; Wisnivesky et al., 2005	15
3	Predict the outcome of patients with infections	Adrie et al., 2009; Al-Hasan et al., 2013; Fang et al., 2011; Hirsch et al., 2012	4
4	Predict patient pathogen carriage at admission	Aguiar et al., 2012; Elias et al., 2013; Harbarth et al., 2008; Hsu et al., 2008	4
5	Assess the usefulness of certain data or predictors for the prediction	de Oliveira et al., 2006; Goulenok et al., 2013; Hurr et al., 1999; Khanafer et al., 2013; Lopes et al., 2009; Mirsaeidi et al., 2009; Morales et al., 2011; Olsen et al., 2009; Stevens et al., 2013; Yadav et al., 2009	10
6	Predict patient care path	Escolano et al., 2000	1
7	Predict the probability of drug resistance	Lodise et al., 2003	1
8	Predict length of stay and other economic burdens	Stewardson et al., 2013	1
9	Predict the pathogens that cause the infection	Visscher et al., 2008	1

A.3.1.1 Identify the patients with infections retrospectively

One of the most common objectives was to identify the patients with certain infections retrospectively. The rationale behind this objective had multiple points. As the problem of HAIs becomes a public concern, the governments of many countries have created policies that ask hospitals to report the rate of HAIs in their organizations (Thomas& Viner-Brown, 2010). Traditionally, this work is done through the manual review of patient charts by trained infection preventionists under infection surveillance programs (Halpin, Shortell, Milstein, & Vanneman, 2011). As the definitions of HAIs are very complicated, it is very time consuming for the preventionists to perform this manual review. Moreover, the infection review process also creates inconsistency of the identification results, as human judgement is involved because of

the complicated definition of HAIs, even though these definitions are standards acknowledged by the community (Burke, 2003; van Mourik et al., 2012). Furthermore, because of the inconsistency of the results, it is also difficult to compare the results generated by different people, sometime even within the same organization (Trick et al., 2004). Consequently, it might make the report and surveillance program not as effective as it is intended to be.

A.3.1.2 Identify patients with high risk of contracting HAIs in real time

Another common objective was to identify patients who have high risk of suffering from certain HAIs in real time. Predictive modeling of HAIs to identify high risk patients in real time can be beneficial in several ways in the context of HAI prevention. First, potential outbreaks might be prevented by putting high risk patient in the watching list (Chung et al., 2009). Secondly, from the individual patient perspective, potential harm might be mitigated for patients by helping physicians with the diagnosis as the physicians sometimes have to make some tough decisions without full information. For example, when a patient is suffering from diarrhea and the information about the bacteria that caused the diarrhea is not available (e.g., growing the bacteria culture may require multiple days). In this case, physicians have to choose antibiotics to deal with the unknown bacteria without full knowledge. If the diarrhea is not caused by bacteria such as *Salmonella*, *Shigella*, and *Campylobacter* species, empirical antibiotic therapy against those bacteria might be harmful for patients with *Clostridium difficile* associated diarrhea (Peled et al., 2007). Thirdly, it is also suggested that appropriate antibiotic therapy might be achieved for HAIs by simply providing the risk factors (Goulenok et al., 2013). Finally, by identifying high risk patients for certain HAIs, preventive interventions can be applied more effectively and efficiently. For example, Kinlin et al. (2010) derived a model to stratify risk of contracting nosocomial pneumonia for patients after Coronary Artery Bypass Graft (CABG) surgery such that the patients can be prioritized to receive the arguably beneficial preventive interventions (i.e. particularly silver-coated endotracheal tubes and selective digestive decontamination), as these preventive interventions are potentially associated with risks and substantial economic costs.

A.3.1.3 Predict the outcomes of patient with infections

Another similar objective was to predict the outcome (e.g. mortality) of patients with certain infections, as the prediction on the patient outcome may also have impact on the physicians' treatment strategies. For example, the prediction of outcome or severity has impact on the physician's decision on the site of care (e.g. discharge home, or long term care center) (Fang et al., 2011). The prediction also has impact on choice of drugs, especially for infectious diseases. As antibiotic resistance has become a challenge issue for the whole population (Levy & Marshall, 2004), hospitals are encouraged to implemented antibiotic stewardship. However, when patients' conditions are critical, certain restricted drugs might be the only effective choice. Predicting the outcome (mortality) becomes an important part for the antibiotic stewardship program and the result can have great impact on the physicians' treatment strategies for patients (Hirsch et al., 2012).

A.3.1.4 Predict patient pathogen carriage at admission

A fourth objective was to predict patient pathogen carriage at admission. Patients with asymptomatic carriage of certain pathogens might be big threats for other patients, as well as for themselves when they were not identified at their admission point. However, it is also costly and often not practical to screen every patient at the admission. For instance, the screening of MRSA carriage on admission is recommended by current guidelines for infection control. However the evidence for benefit remains controversial (Dancer, 2008). Therefore, predictive models are often used to reduce the number of patients that needed to be screened, especially in the low disease prevalence situation (Elias et al., 2013; Harbarth et al., 2008).

A.3.1.5 Assess the usefulness of certain data or predictors for the prediction

In the healthcare setting, many data values are collected for various reasons. It is logical to use already-existing data or indices to build the predictive models. Therefore, many of the papers were interested in evaluating the usefulness of a particular data set or predictors on the prediction of the occurrence of

certain HAIs. For example, for the National Nosocomial Infection Surveillance (NNIS) index alone, there were two studies (de Oliveira et al., 2006; Morales et al., 2011) that evaluated the applicability of this index for the prediction of SSI. There was also another study (Mirsaeidi et al., 2009) that evaluated the applicability of The Acute Physiology and Chronic Health Evaluation II (APACHE II) index for the prediction of ventilator-associated pneumonia.

A.3.1.6 Others objectives

A few other objectives of the predictive modeling in the HAI context were also found in the literature. Escolano et al. (2000) predicted the state path of patient care for ICU patients in terms of severity and complexity of the patients' nosocomial infection. Lodise et al. (2003) predicted the probability of methicillin resistance patients with *Staphylococcus aureus* bacteremia in order to give suggestions for the antibiotic therapy. Stewardson et al. (2013) built a predictive model to predict the length of stay and other economic burdens of patients with bloodstream infections caused by one type of *Enterobacteriaceae* to justify the prioritization of infection control.

A.3.2 Predictive modeling methods and their performance

Despite there being many modeling methods found in the HAI predictive modeling literature, the dominant methods were classification algorithms. This is consistent with the result from the previous section that most of the objectives in the HAI context are related to patient classification. Table A-3 is a summary of the methods.

Table A-3 Methods of HAI predictive modeling

Index	Modeling objectives	References	Count
1	Logistic Regression	Chandra et al., 2012; Chung et al., 2009; de Oliveira et al., 2006; Fang et al., 2011; Goulenok et al., 2013; Harbarth et al., 2008; Hsu et al., 2008; Hurr et al., 1999; Joshi et al., 1992; Khanafer et al., 2013; Kinlin et al., 2010; Morales et al., 2011; Olsen et al., 2009; Peled et al., 2007; Sanagou et al., 2013; Stevens et al., 2013; Su et al., 2011; van Mourik et al., 2012; van Walraven & Musselman, 2013; Yadav et al., 2009	20
2	Rule based method	Al-Hasan et al., 2013; Choudhuri et al., 2011; Daneman et al., 2009; Dubberke et al., 2011; Hautemanière et al., 2013; Hornbuckle et al., 1998; Leth & Møller, 2006; Lo et al., 2013; Mirsaeidi et al., 2009; Park et al., 2013; Steinmann et al., 2008; Trick et al., 2004; Wisnivesky et al., 2005	13
3	Statistical test	Adrie et al., 2009; Elias et al., 2013; Fang et al., 2011; Lee et al., 2013; Lodise et al., 2003; Tacconelli et al., 2008	6
4	Survival analysis	Escolano et al., 2000; Graves et al., 2011; Stewardson et al., 2013	4
5	Expert system	Kahn et al., 1993	2
6	Regression tree	Aguiar et al., 2012; Hirsch et al., 2012	2
7	Artificial neural network	Chung et al., 2009; Fang et al., 2011; Pearl & Bar-Or, 2012	3
8	Natural language processing	Proux et al., 2011	1
9	Bayesian network model	Visscher et al., 2008	1
10	Decision tree	Lopes et al., 2009	1

A.3.2.1 Rule based method

Thirteen (26%) papers used rule based methods in this literature. This number makes it the second most commonly adopted method. It appears common to use rule based method especially for the retrospective case finding, as HAIs usually have standard case definitions (Garner, Jarvis, Emori, Horan, & Hughes, 1988; Horan, Gaynes, Martone, Jarvis, & Grace Emori, 1992). Indeed, many automated HAI detection systems use rule based methods (Freeman, Moore, García Álvarez, Charlett, & Holmes, 2013). Intuitively, rules generated through this approach should be accurate enough to separate HAIs and non-HAIs. However, the performance varies greatly in different settings. The area under curve (AUC) metric for those studies that reported AUC had an average of 0.76, a median of 0.76, and a range from 0.59 to 0.88.

One explanation for this variability might be due to the fact that most definitions of HAIs are complicated and the rules generated through these definitions and the availability of the data related to these

definitions vary greatly in different study settings. For example, for most of the HAIs, the standard definitions for them provided by CDC always have two types: cultured-based definitions and clinical-based definitions (Kahn et al., 1993). The issue with definitions is compounded by the hospital data. The data related to these definitions are often found in different hospital systems. This is important since studies in the literature rarely have full access to the integrated data from various systems (Leal & Laupland, 2008). Moreover, the resulting rules generated from the different definitions might not agree with each other. For example, Apte et al. (2011) compared the results from two rules which were based on clinical medication diagnosis codes and laboratory culture data respectively to identify SSIs. The two rules only had 81.3% positive agreement and 50% negative agreement of their results. Bouzbid et al. (2011) compared the performance of seven HAI detection strategies that were based on four data sources (i.e. microbiology data, drug prescriptions, medico-administrative, and hospital discharge summaries) in an intensive care unit during the period between 2000 and 2006. The study also showed that the classification result from different rules that were based on different data sources exhibited great variability in terms of sensitivity and specificity.

A.3.2.2 Logistic regression

The most commonly used modeling method in this literature was logistic regression (40%). It was used for two purposes: 1) to build the classification algorithm, and 2) to identify the important predictors.

The main benefit of using logistic regression as opposed to rule based methods is that the modellers are not confined by the definition of HAIs. They can add any suspicious predictors into the model. However, the performance of logistic regression also varies. The AUC of these studies that reported this metric had an average of 0.80, a median of 0.80, and a range from 0.67 to 0.97. Based on these numbers, logistic regression did only slightly better than rule based methods.

Logistic regression was also used to obtain the importance of the predictors, which is often represented by the regression coefficient (odd ratio). In most of cases, logistic regression models were converted to score

models to be used in practice by giving weight to predictors according to their regression coefficient (Harbarth et al., 2008; Su et al., 2011).

A.3.2.3 Machine learning classification algorithms

Many machine learning classification algorithms were also found in the literature. Compared with rule based methods and logistic regression, classification using machine learning approaches was relatively rare (Ehrentraut, Tanushi, Dalianis, & Tiedemann, 2012). Within the machine learning methods, the studies can be further divided into the ones that use unstructured clinical data and the ones that do not.

Handling unstructured clinical narratives requires the use of Natural Language Processing (NLP) algorithms to encode the data. Although progress of applying the NLP methodology to clinical data has been achieved in few areas, such as automatic assignment of ICD-9-CM codes to clinical free text (Meystre, Savova, Kipper-Schuler, & Hurdle, 2008), very few implementations have been created for the detection of HAIs. We found one study (Proux et al., 2011) in the literature reviewed that described an architecture and system to monitor HAIs in real time by using NLP to process unstructured patients records. The preliminary experiment achieved a sensitivity of 87.6% and a specificity of 97.4% performance.

The studies which did not involve unstructured clinical narratives were relatively easier for the modellers. However, the machine learning methods were not widely used in the HAI context. There were two articles using regression tree methods, three Artificial Neural Network models, one decision tree model, and one Bayesian Network model (details can be seen in Table A-2). The AUC of these studies that reported this metric had an average of 0.79, a median of 0.80, and a range from 0.72 to 0.87.

A.3.2.4 Other methods

Two other large groups of methods found in the literature were statistical tests and survival analysis methods. Statistical tests were often used for filtering important predictors from a large number of variables. For example, Chi-square test or Fisher Exact Test was often used to test qualitative variables

(Adrie et al., 2009; Elias et al., 2013; Fang et al., 2011; Lee et al., 2013; Lodise et al., 2003), while Wilcoxon test or Student's t-test or Kruskal-Wallis test was used for continuous variables (Adrie et al., 2009; Fang et al., 2011; Lee et al., 2013; Tacconelli et al., 2008).

Survival analysis methods were often used for the prediction of time-related metrics, such as length of stay. For example, three papers (Escolano et al., 2000; Graves et al., 2011; Stewardson et al., 2013) in the literature used multiple state models for the patients' state change with HAIs in their hospital stay.

A.3.3 Data for the predictive modeling of HAIs

Modeling cannot be done without data. As already shown in the previous sections, data has a great impact on the choice of the methods and their performance for different modeling purposes.

A.3.3.1 The data sources

Data values regarding HAIs can be found scattered in four types of systems (de Bruin, Seeling, & Schuh, 2014; Halpin et al., 2011):

1. Administrative systems (admissions-transfer-discharge system, for patient demographics, location, and diagnosis codes);
2. Laboratory systems (including microbiology culture data, biochemical test data, and radiology/imaging data);
3. Pharmacy systems (ordering and administration of antibiotics and other medications);
4. Clinical entries (or electronic medical records, including physician and nurse entries).

Most of the papers did not explicitly describe where and how they extracted the data from different databases. Figure A-6 is the distribution of using these data sources summarized from the 50 articles.

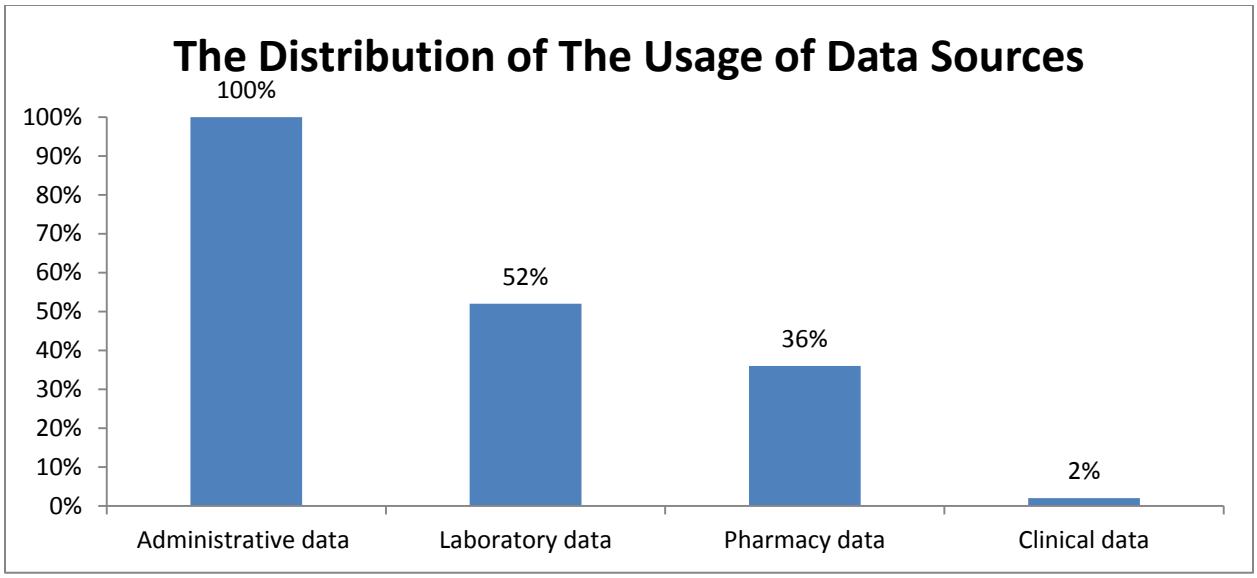


Figure A-6 Percentage of publication using different data sources

As shown in Figure 6, administrative data were used by all the studies. However, administrative data might be not very effective for the predictive modeling of HAIs. One study (Sherman et al., 2006) showed that the method of identifying HAIs by reviewing administrative data alone only had a sensitivity of 61% and a poor positive predictive value (20%).

Laboratory data were the second (52%) most common source for the predictive modeling of HAIs in terms of the number of studies. However, in some sense, laboratory data can be considered the most important source for the predictive modeling of HAIs, as laboratory tests, either microbiologic tests or chemical tests, provide the highest quality evidence for the confirmation of infections. The importance of the laboratory results is demonstrated by the case definitions of HAIs; the definitions often include the presence of the pathogens of the infections. Unfortunately, there are still accuracy issues with the laboratory tests, (i.e. sensitivity and specificity) and must be used with other information.

More than one third of the studies used pharmacy data. One important component of pharmacy data for the prediction of HAIs are the records of antibiotic exposure for patients. Antibiotic exposure is noted as an important risk factor for certain HAIs, such as CDI and SSI. Other medication data, such as the use of

an immunosuppression drug, has also been considered important for the prediction of HAIs, as this type of drug often increases patients' susceptibility to infections.

Only one (2%) study used clinical entries, as it appears to be very difficult to process the free-text format data (Proux et al, 2011). However, the clinical entries are widely used by infection preventionists in practice.

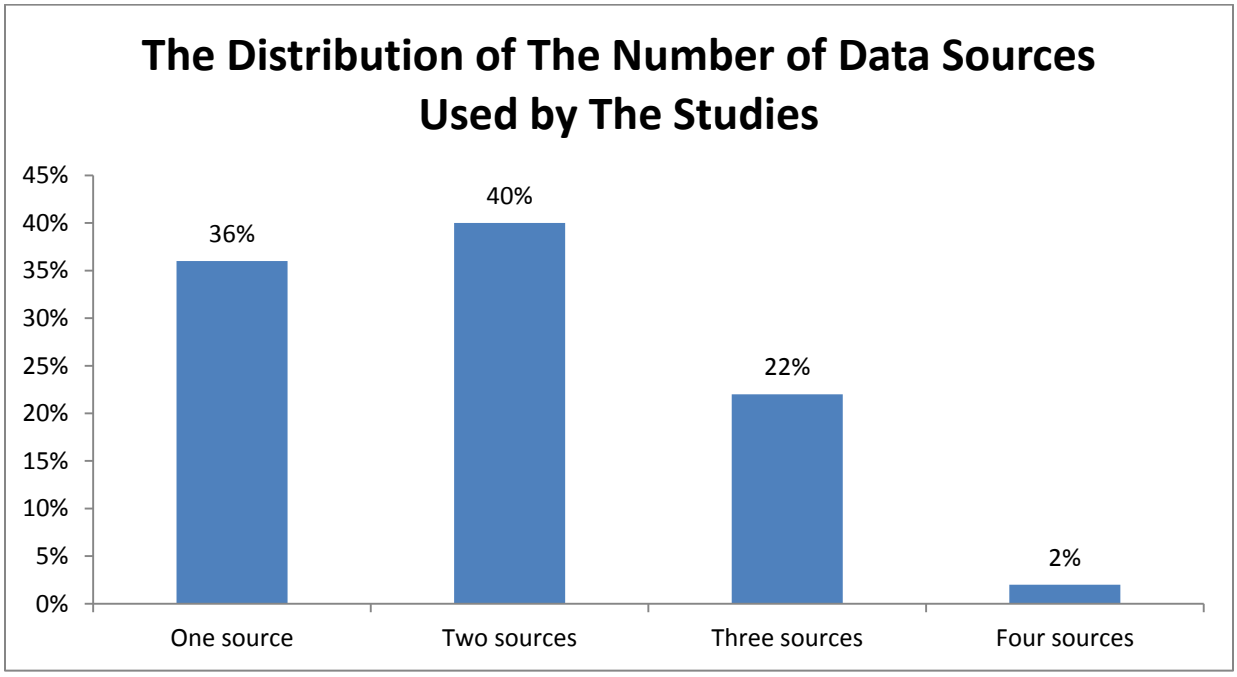


Figure A-7 Percentage of publications by number of data sources used in the study

Figure A-7 is the distribution of the number of the data sources used by the publications. Eighteen (36%) of the articles just had one data source, which was administrative data exclusively. For the classification algorithms, the AUC of these studies that reported this metric (5) had an average of 0.76, a median of 0.75, and a range from 0.59 to 0.95.

Twenty (40%) of the articles had two data sources. Fourteen (70%) of the twenty studies used the laboratory data besides the administrative data. Six (30%) of them used the pharmacy data. For the

classification algorithms, the AUC of these studies that reported this metric (6) had an average of 0.80, a median of 0.83, and a range from 0.76 to 0.87.

Eleven (22%) of the articles had three data sources. All of them have the same three types of data sources: administrative data, laboratory data, and pharmacy data. For the classification algorithms, the AUC of these studies that reported this metric (6) had an average of 0.82, a median of 0.82, and a range from 0.70 to 0.97.

Only one (2%) study (Proux et al., 2011) used all four types of data sources. The reported performance of the preliminary experiment seems superior compared with the other studies that use less data sources, as the overall sensitivity and the specificity were 87.6% and 97.4% respectively.

The result from this small sample size seems to suggest that more data does not guarantee better accuracy, but it does suggest that on average more data might be better.

A.3.3.2 The data temporality

Another dimension of the "data" issue is the temporality property of the data. Typically, data are often not produced en masse at a single time, but are generated piece by piece along the care pathway of the patients, as shown in Figure A-8.

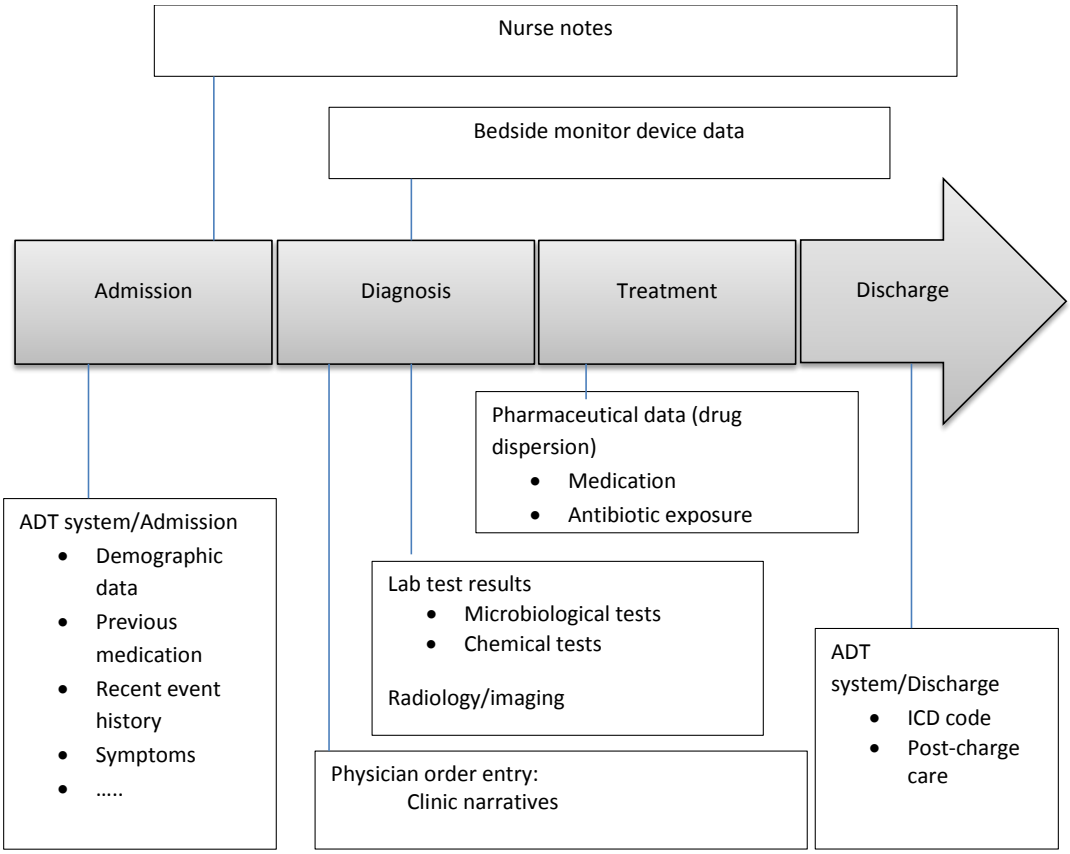


Figure A-8 Temporality of hospital data

The availability, format, and the difficulties of dealing with the data have great impact on the predictive modeling of HAIs in practice, especially for the real time prediction (van Mourik, Troelstra, van Solinge, Moons, & Bonten, 2013). However, most of the studies reviewed avoided the discussion of this issue. Only Proux et al. (2011) presented an architecture and system for monitoring HAI in real time based on Natural Language Processing. However, the full evaluation of the effectiveness of their system in a real hospital environment is still pending.

A.4 Discussion

This review analysed literature regarding predictive modeling of HAIs between 1990 and 2013 in three data bases: Web of Science, Scopus, and PubMed. Fifty full-text articles were identified for the detailed analysis; to understand the modeling purpose, modeling methods, and modeling data in the HAI predictive modeling context.

- Although predictive modeling of HAIs has existed for a relatively long time (Halpin et al., 2011), increasing interest about this topic emerged after year 2007. Most of the research reviewed in this study was from developed economies, despite the fact that other economies have a higher prevalence of HAIs (Allegranzi & Pittet, 2007).
- In terms of a research setting, the majority of the studies involved a single institution, and most of the institutions were large teaching hospitals or medical centers. This characteristic suggests that it would be difficult to make a fair comparison between studies. It also implies that external validation or replication of the findings in the studies is troublesome or even infeasible.
- The modeling purpose behind the reviewed research is very diverse partially because prediction is a loosely defined term in the literature. Eight different modeling purposes were synthesised from the 50 articles. Generally speaking, case classification at various points in time (i.e. on admission, before culture result, at discharge, retrospective case finding) is the dominant objective of the predictive modeling of HAIs, although other minor modeling purposes (e.g. predicting the length of stay of HAI patient) also exist. While the efforts for retrospective case finding modeling is a response to the reporting demanded by public policies, the intention of the classification modeling at other points in time is to assist clinicians with simple tools (e.g. risk score). Although models have been developed for these purposes, the validation and adoption of these models in practice are still challenging because of the study setting mentioned above.

- Aligning with the modeling purposes, the major modeling methods are classification algorithms. The most common methods are rule based and logistic regression, although other more complicated machine learning methods are also employed in this literature. In general, none of the methods had significance in terms of AUC (i.e. average is less than 0.8), although logistic regression appears to do slightly better than the rule based methods. Despite the low performance issue, there are two benefits of predictive modeling especially in the retrospective case finding context. First, predictive modeling can make the case finding result consistent and comparable at least within the same organization. Second, it can save human labor by reducing manual review. These two benefits are very valuable for the surveillance of HAIs. However, low performance is a big issue for the application of predictive modeling in real time prediction situation. Poor performance of the predictive model is a common problem in predictive modeling in the medical field (Berner, 2007). It is one of the important reasons in this field that the modeling strategy shifted from giving "Greek Oracle" predictions to generating "Reminders". Predictive modeling of HAIs is also in the same situation. For reminder systems, "alarm fatigue" related to the poor performance of the predictive models is a big concern for the adoption of the models in practice (Sendelbach, 2012). Indeed, ironically, clinical alarms are on the list of top ten patient safety hazards (Cvach, 2012; Keller, Diefes, Graham, Meyers, & Pelczarski, 2011).
- The processing of data has a great impact on the modeling itself and the implementation of the developed models. However, very few of the papers in this literature talk about the related issues. Nevertheless, several challenges can still be summarized from the literature, especially for the real time application perspective.
    - One challenge is that of system fragmentation. Data in hospitals are highly fragmented and scattered in dozens of various departments, laboratories, and related administrative systems that have their own legacy (Fernando & Dawson, 2009; Rada, 2007). It is very challenging to bring all of them together. Moreover, simply putting all the data together

does not help to build and use a model; much of the data might not be valid or usable for certain purposes.

- Another major challenge is to find the relevant data and synthesize it for HAI prediction purposes, as we know that many relevant data elements are usually not collected for the HAI prediction purpose. For example, it is believed in practice that the absence or reduction of staff combined with the increase of admissions will increase pressure on the staff (Cimiotti, Aiken, Sloane, & Wu, 2012). Therefore, some infection prevention practices, such as hand hygiene, might be not well followed under these stressful situations and there might be an increased risk of HAI for patients. Both the staffing and the admission data are recorded in systems, but without synthesizing the data, they cannot be directly used for the predictive modeling purpose.
- We also have to face the challenge of dealing with massive amounts of unstructured data. Most of the data, especially those generated through various clinical entries systems, are presented as free text. Converting unstructured free text into structured data is still a challenge for many areas in clinical decision support modeling including HAIs, as the data can contain a great deal of "noise" (Ehrentraut et al., 2012).
- Finally, we have to deal with the temporality challenge associated with the data. As patients go through different stages of their hospital stay, data values are generated piece by piece. Predictive modeling has to align the modeling purpose with the data availability at different stages of patient care. Also, appropriate mechanisms of incorporating the newly generated data have to be designed in order to make the models functional in practice.

This study also has a few limitations. First, the key words in the search are not for specific types of HAIs. We might have missed publications that modelled one specific HAI. Second, we searched three databases,

and the number of the articles fully reviewed is relatively small. This small number of studies might or might not be the reality. Some of the findings might need further validation.

In conclusion, the review suggests predictive modeling of HAIs could be useful for many purposes in hospitals. However, many challenges have to be overcome to make the predictive models effective in practice.

A.5 Acknowledgement

I would like to acknowledge Dr. William Cicottelli and Josh Mores for reviewing this appendix.

Appendix B. Simulation model design

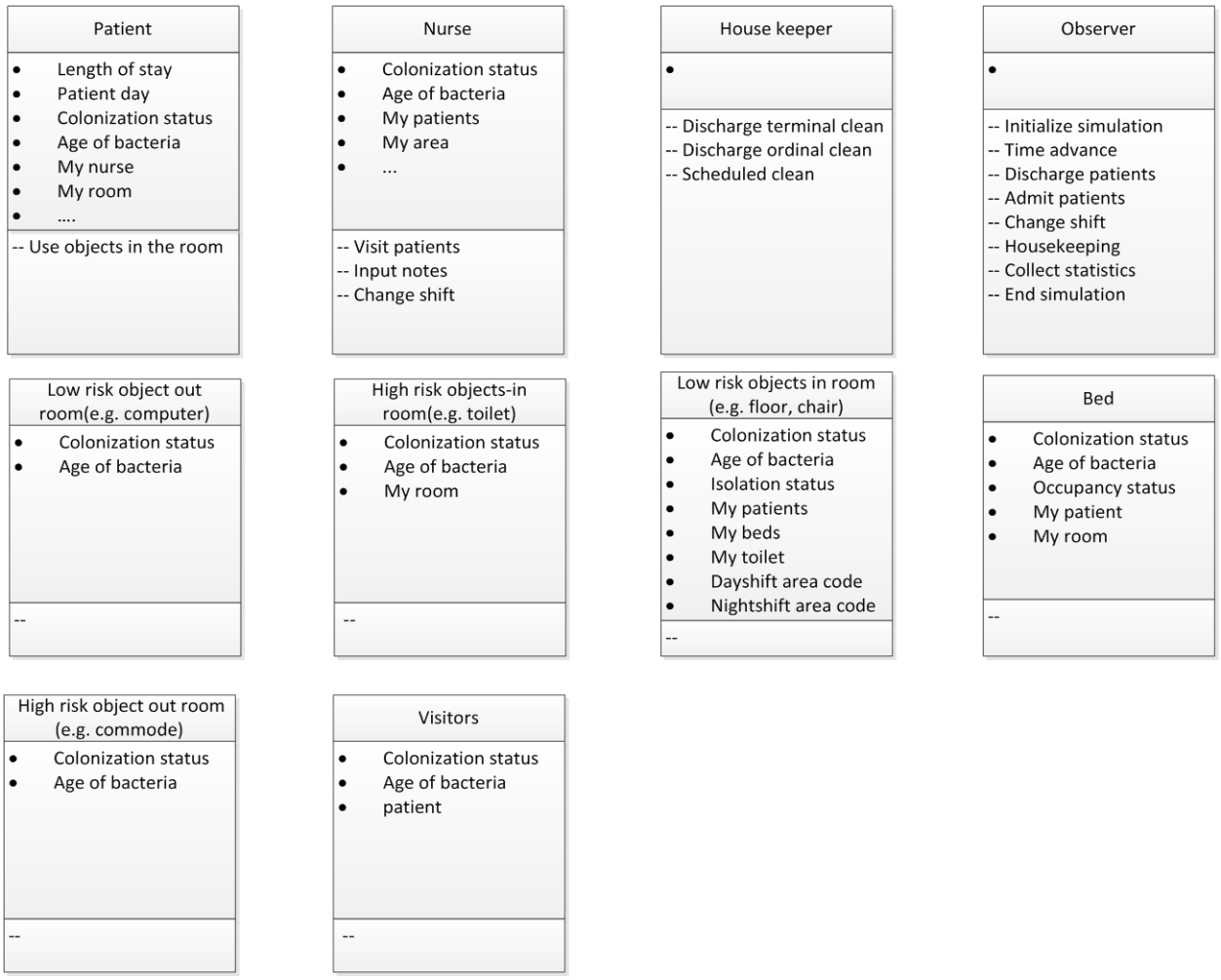


Figure B-1 Simulation model entities and attributes

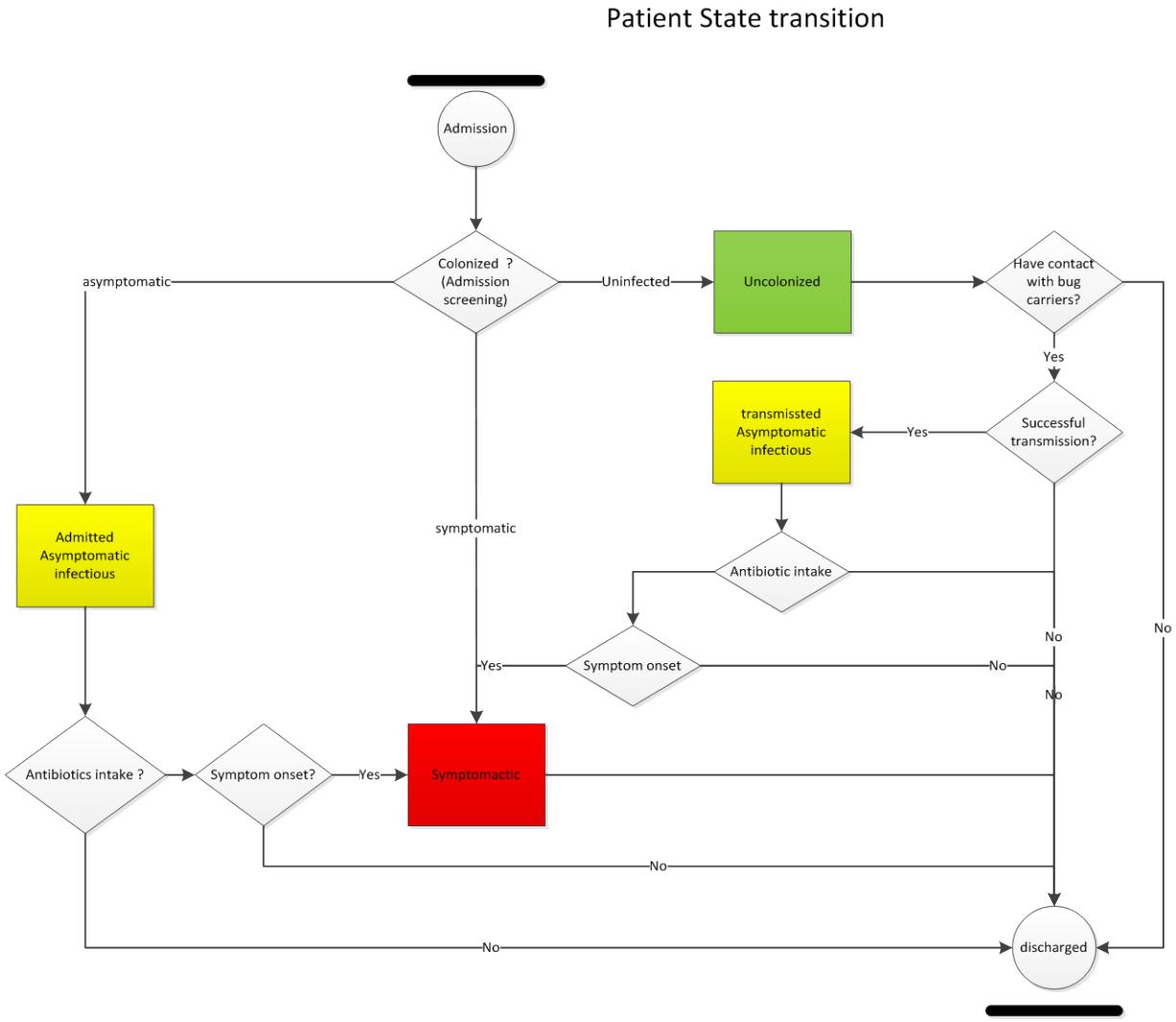


Figure B-2 Patient state transition

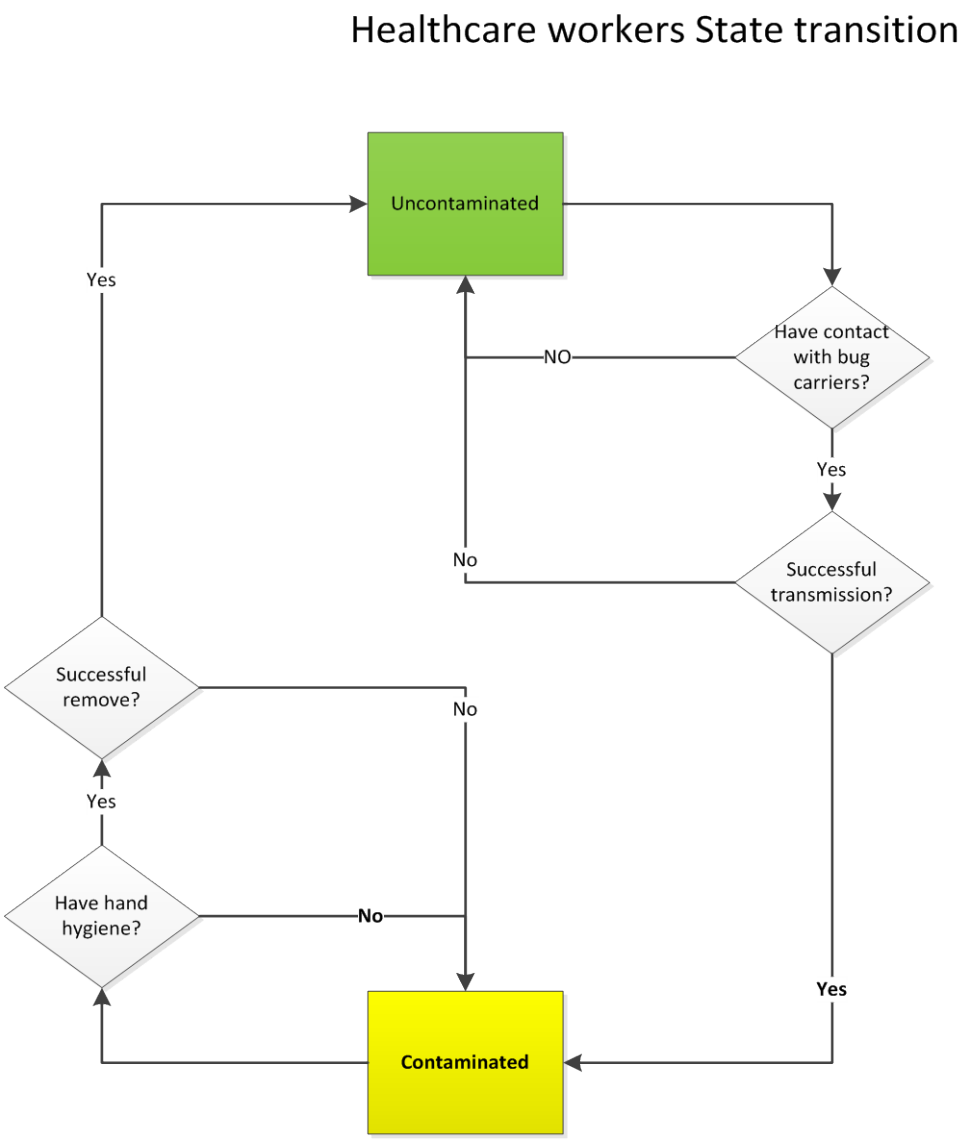


Figure B-3 Healthcare worker state transition

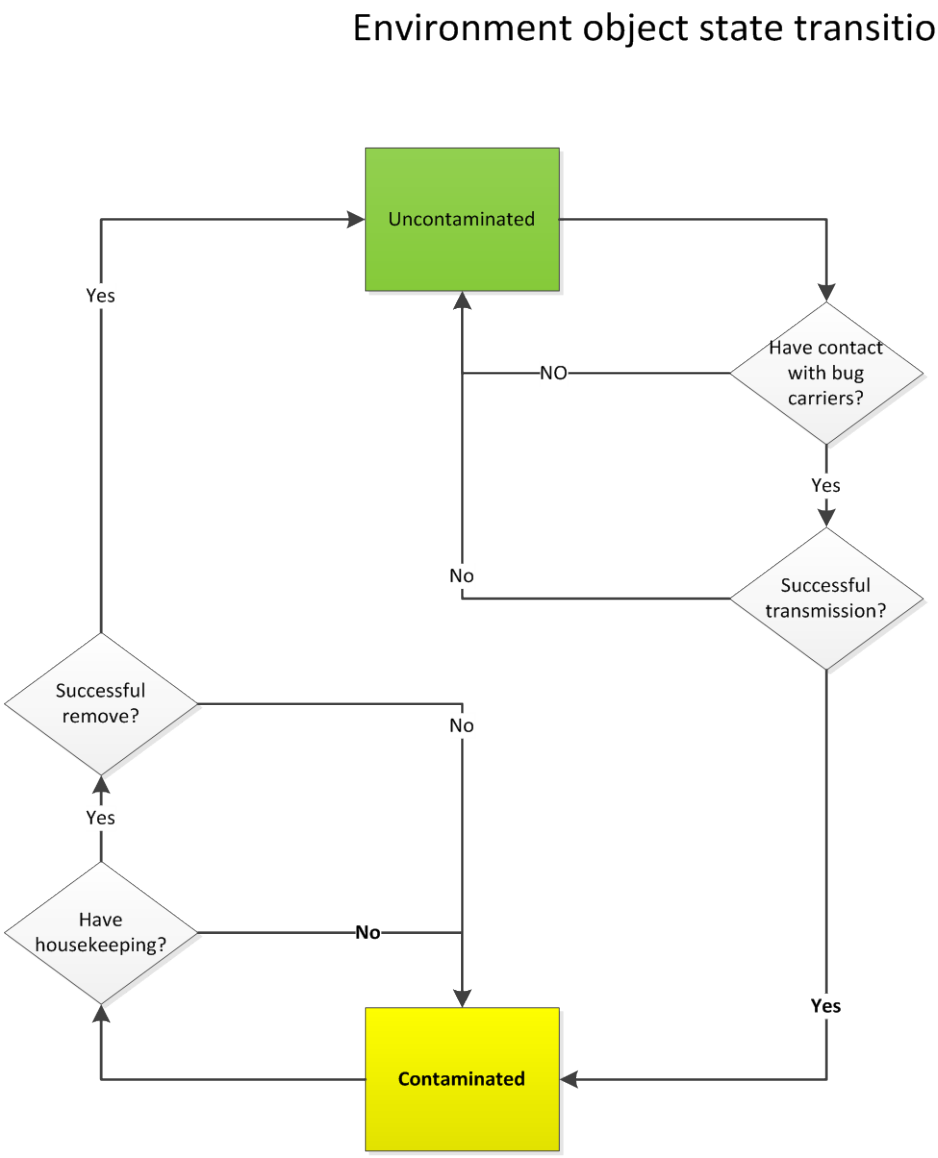


Figure B-4 Environment object state transition

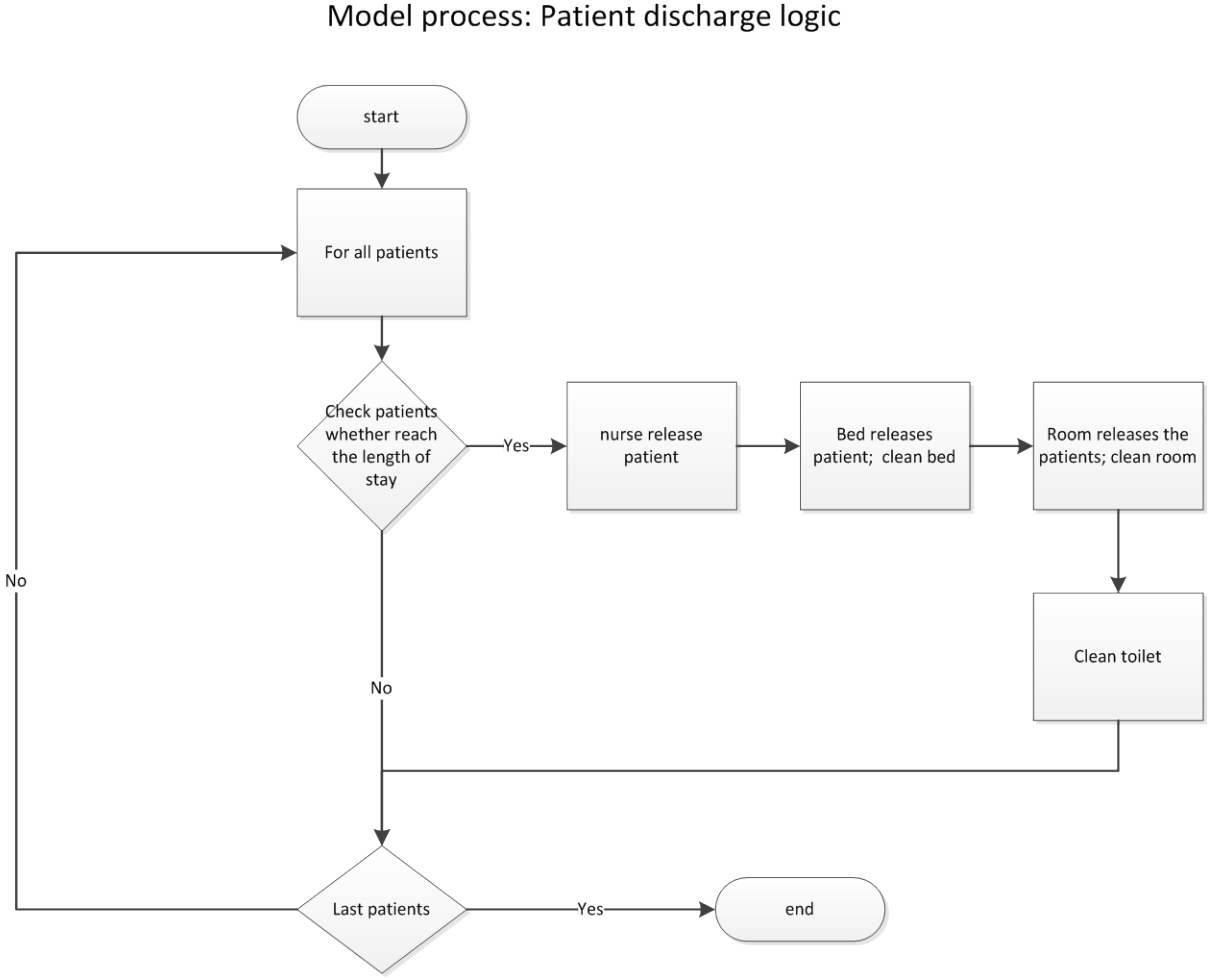


Figure B-5 Patient discharge logic

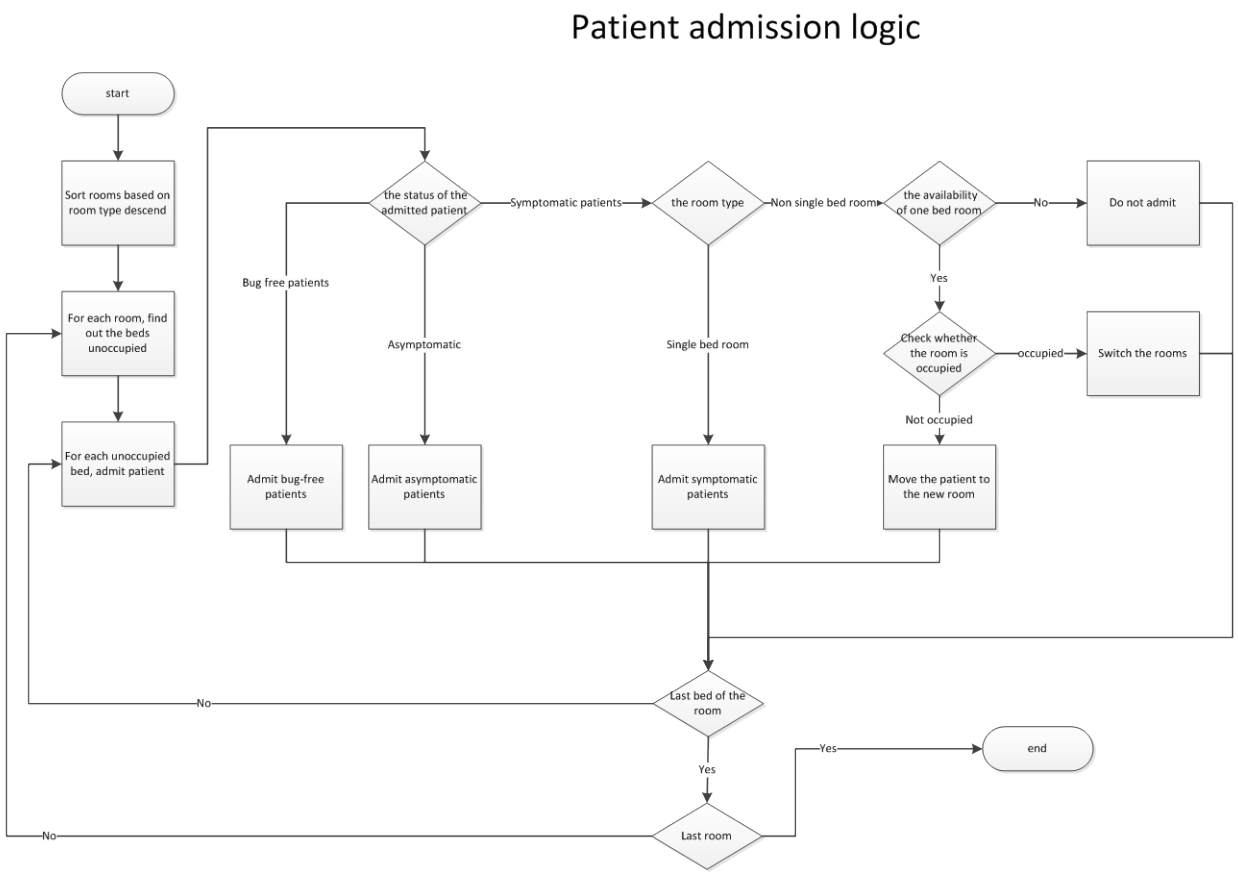


Figure B-6 Patient admission logic

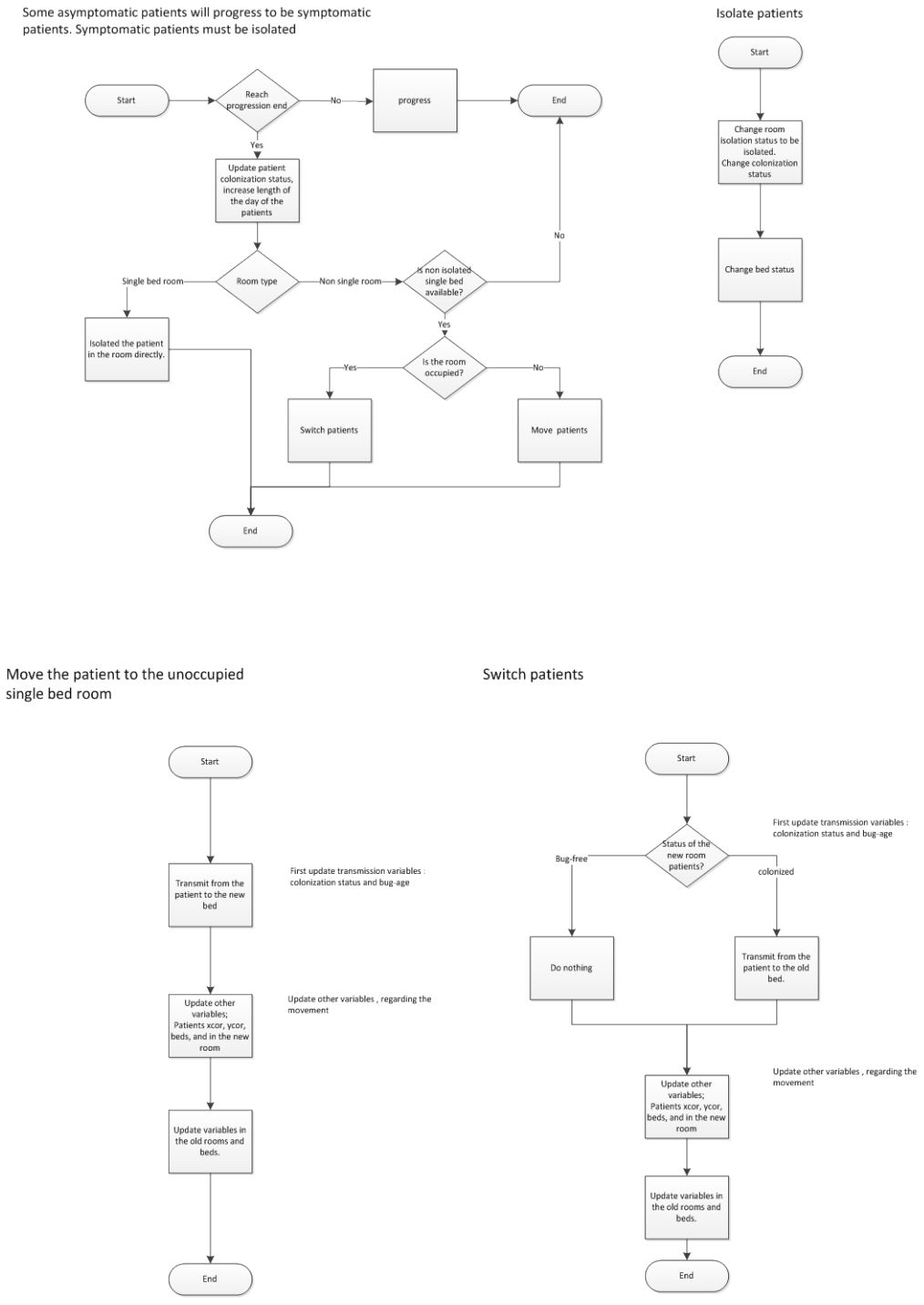


Figure B-7 Natural progress logic

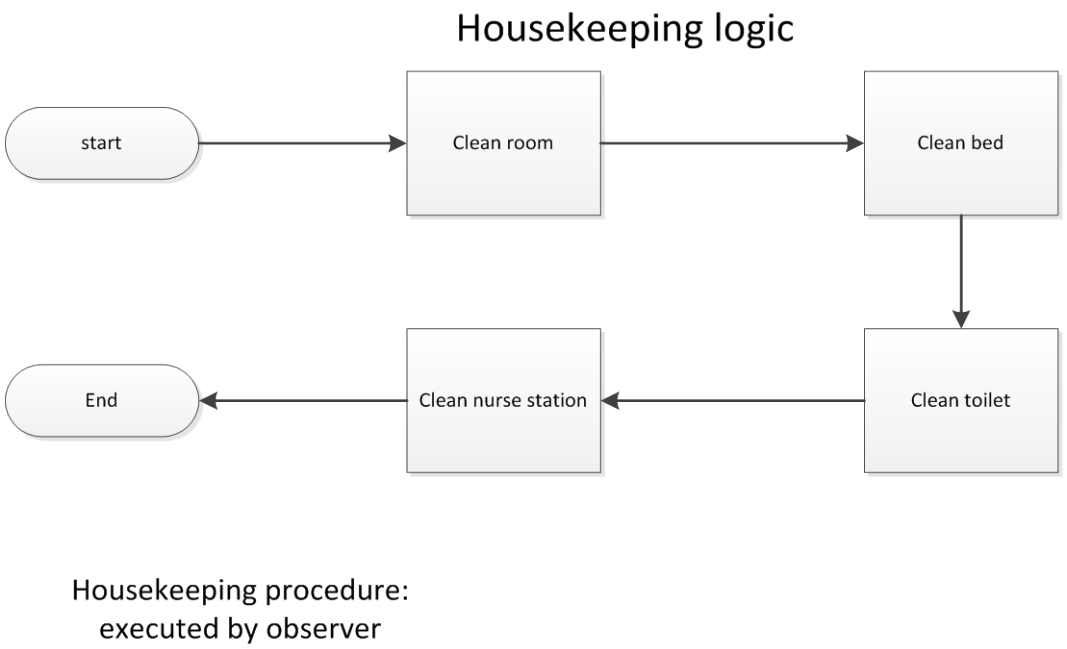


Figure B-8 Housekeeping logic

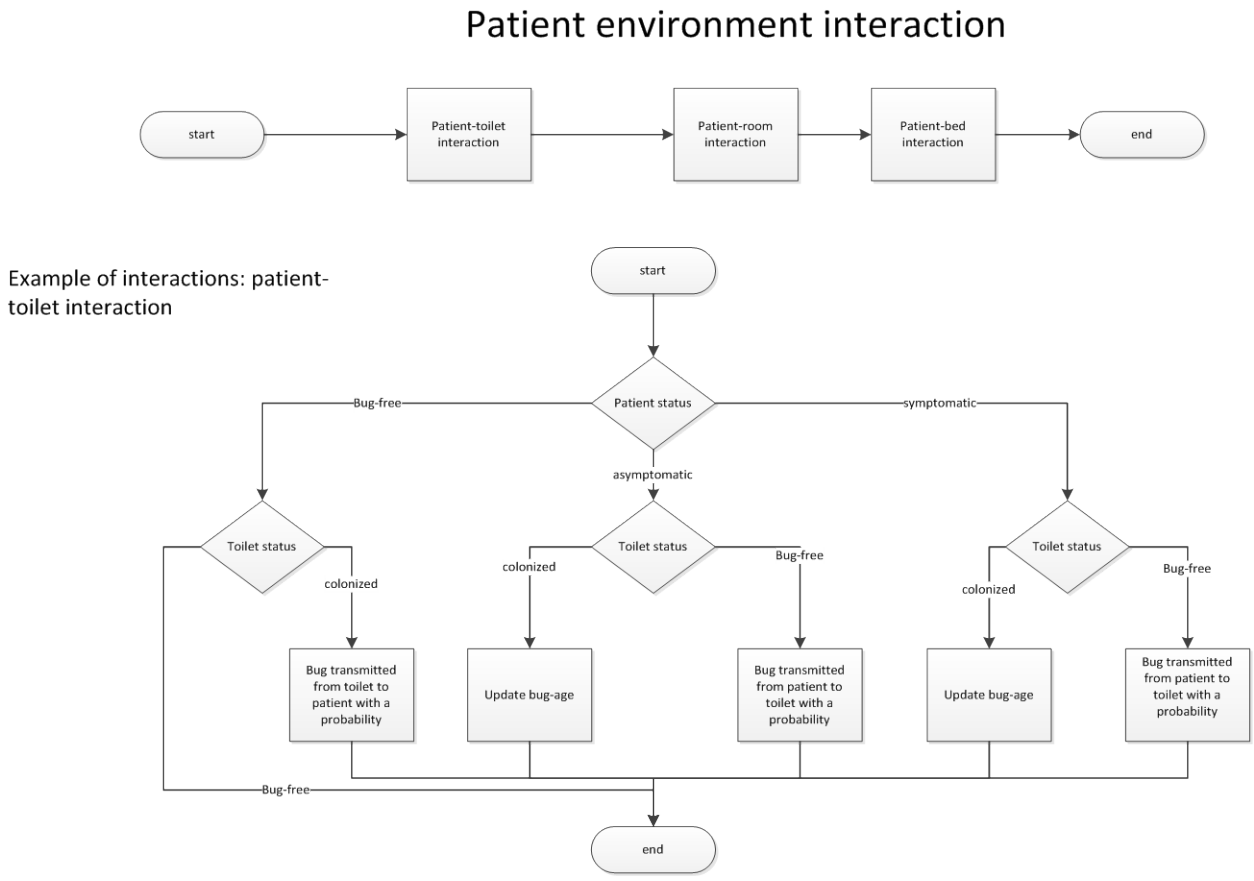


Figure B-9 Patient environment object interaction

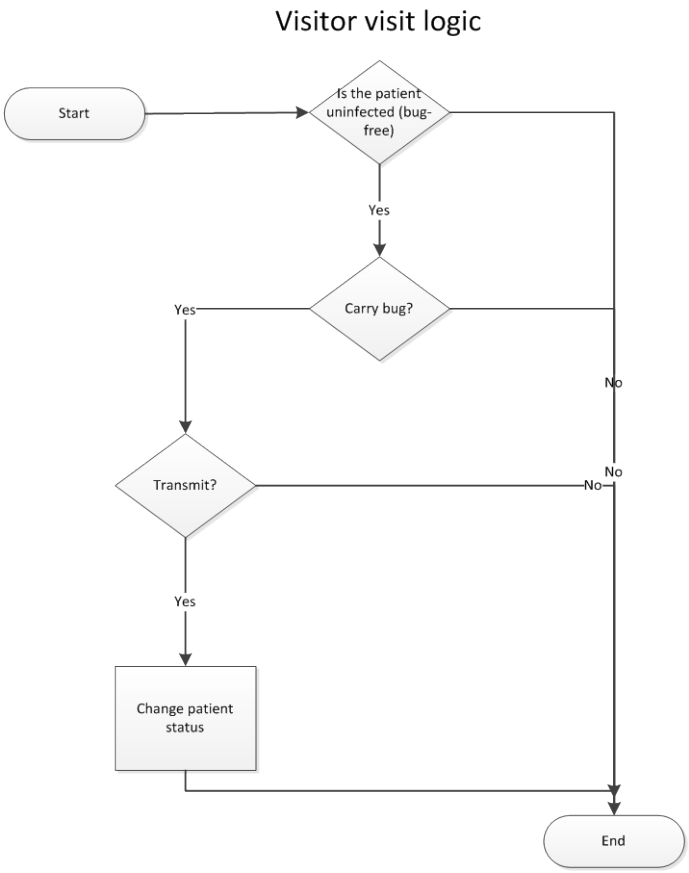


Figure B-10 Visitor visit logic

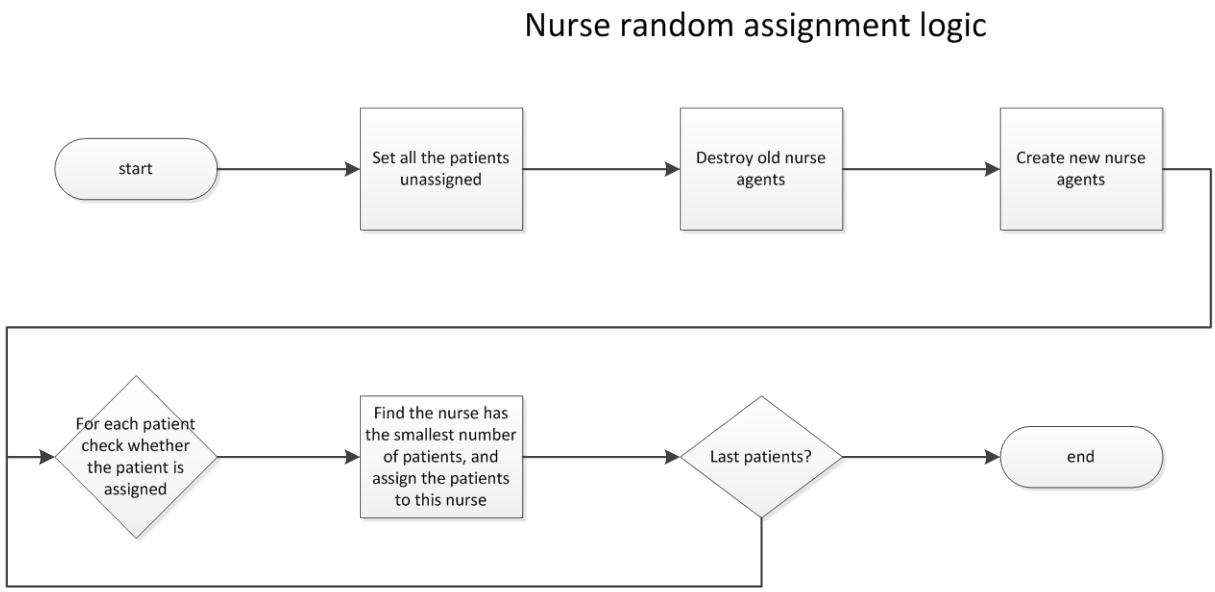


Figure B-11 Nurse random assignment logic

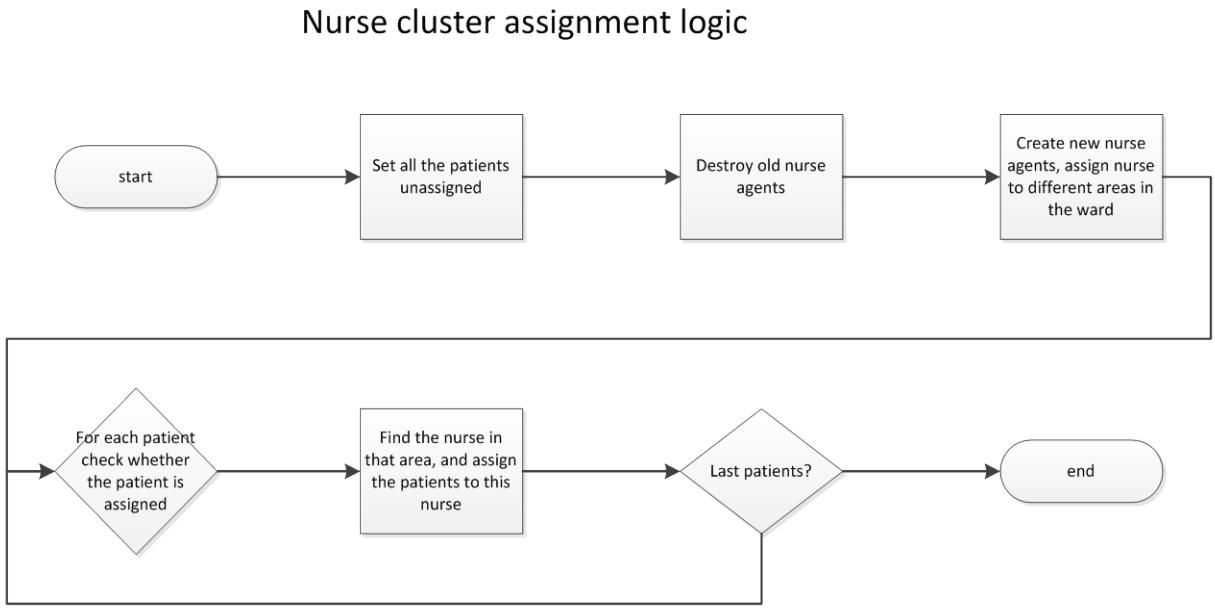


Figure B-12 Nurse Cluster assignment logic

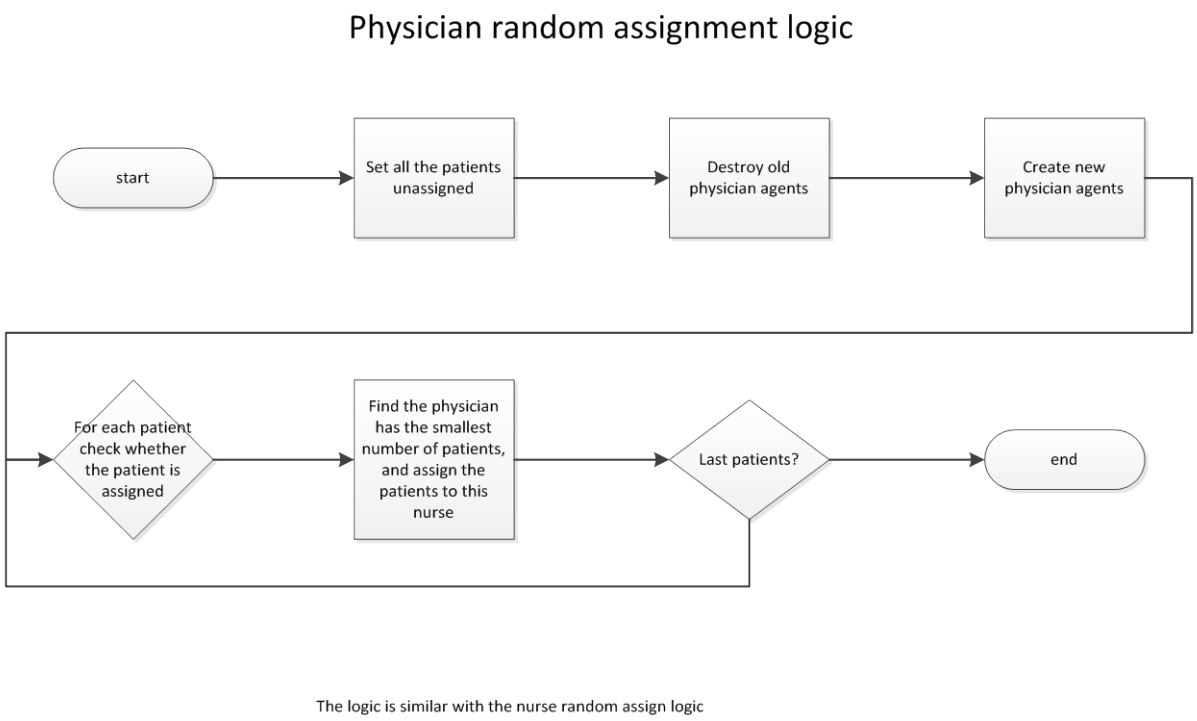


Figure B-13 Physician assignment logic

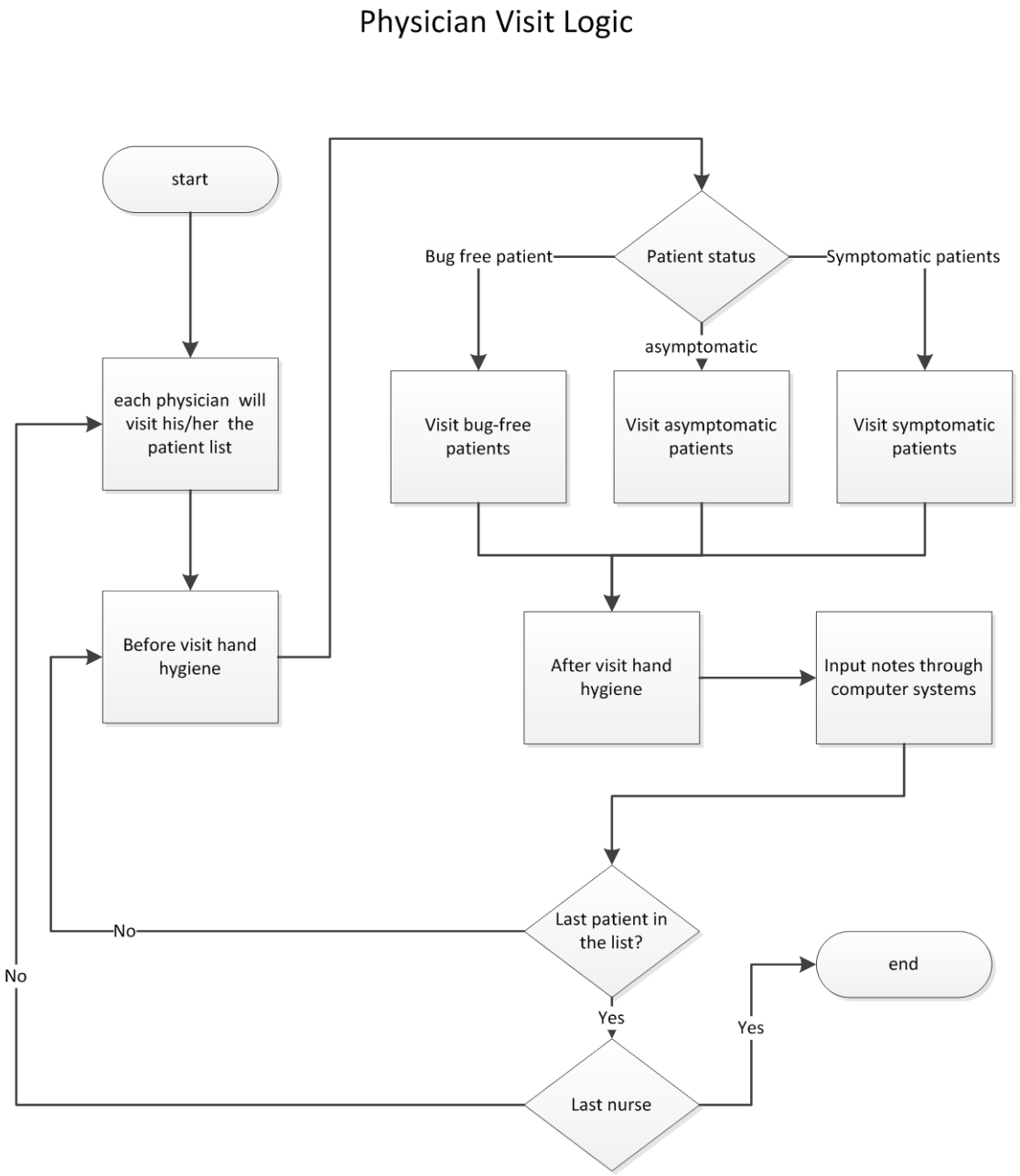


Figure B-14 Physician visit logic

Nurse visit logic

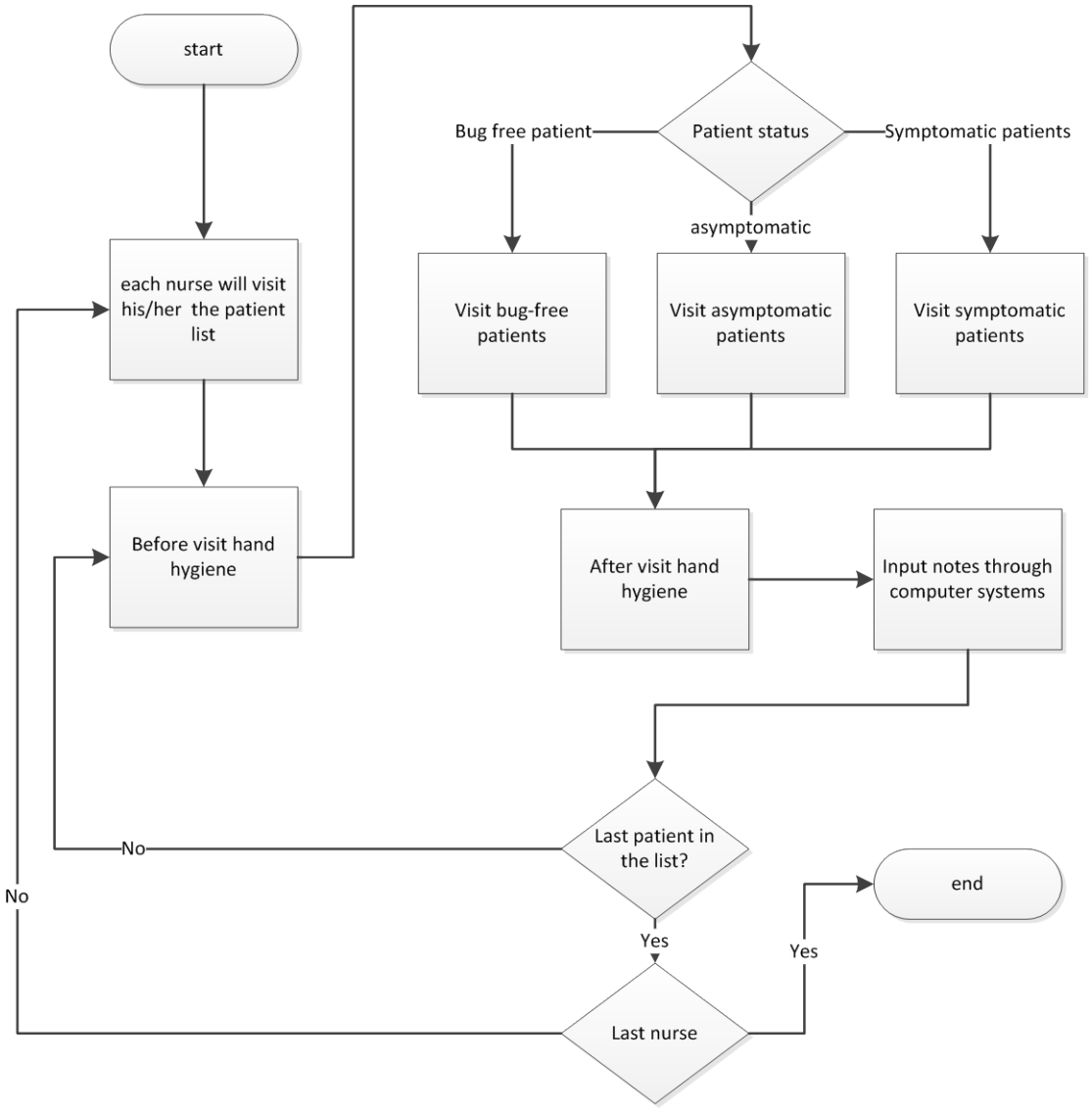


Figure B-15 Nurse visit logic

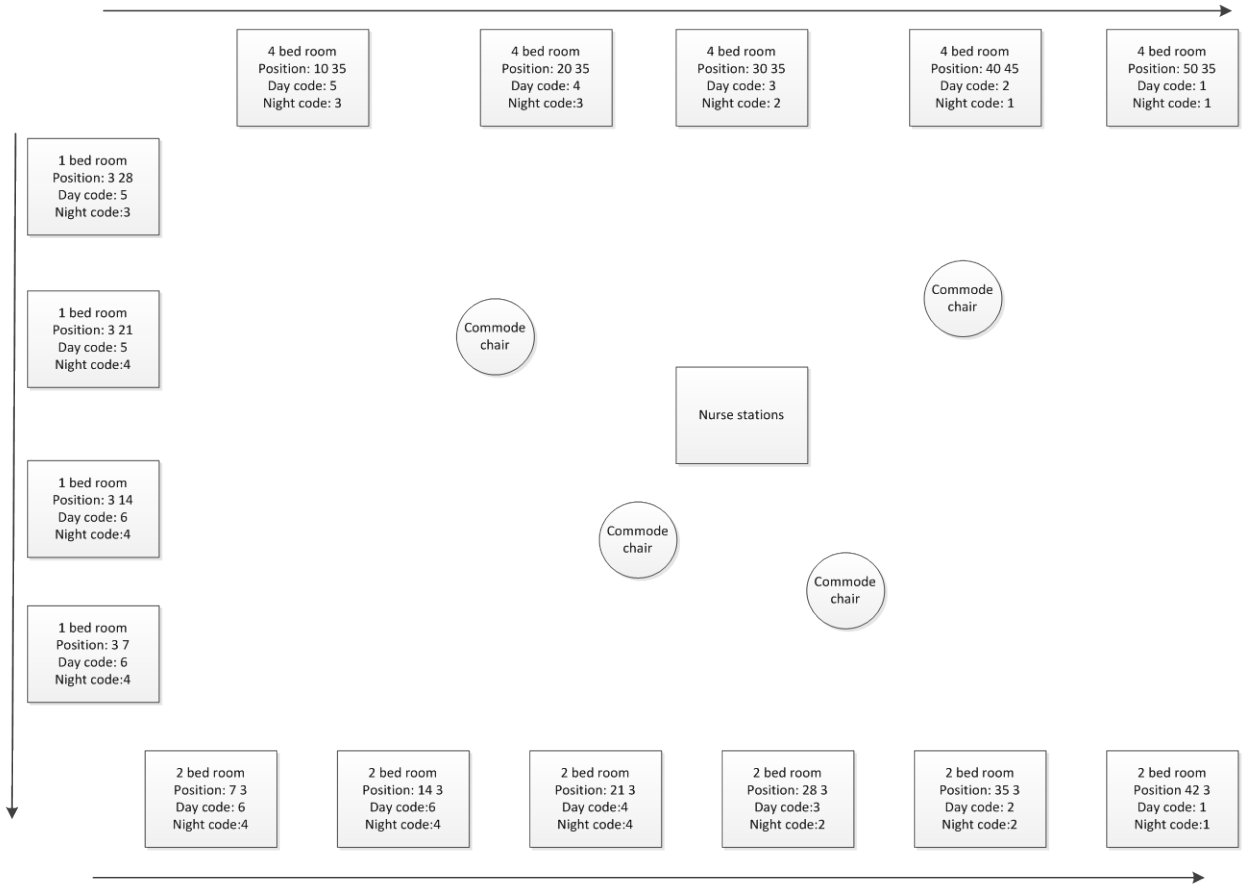


Figure B-16 Ward layout

Appendix C. Simulation model calibration process

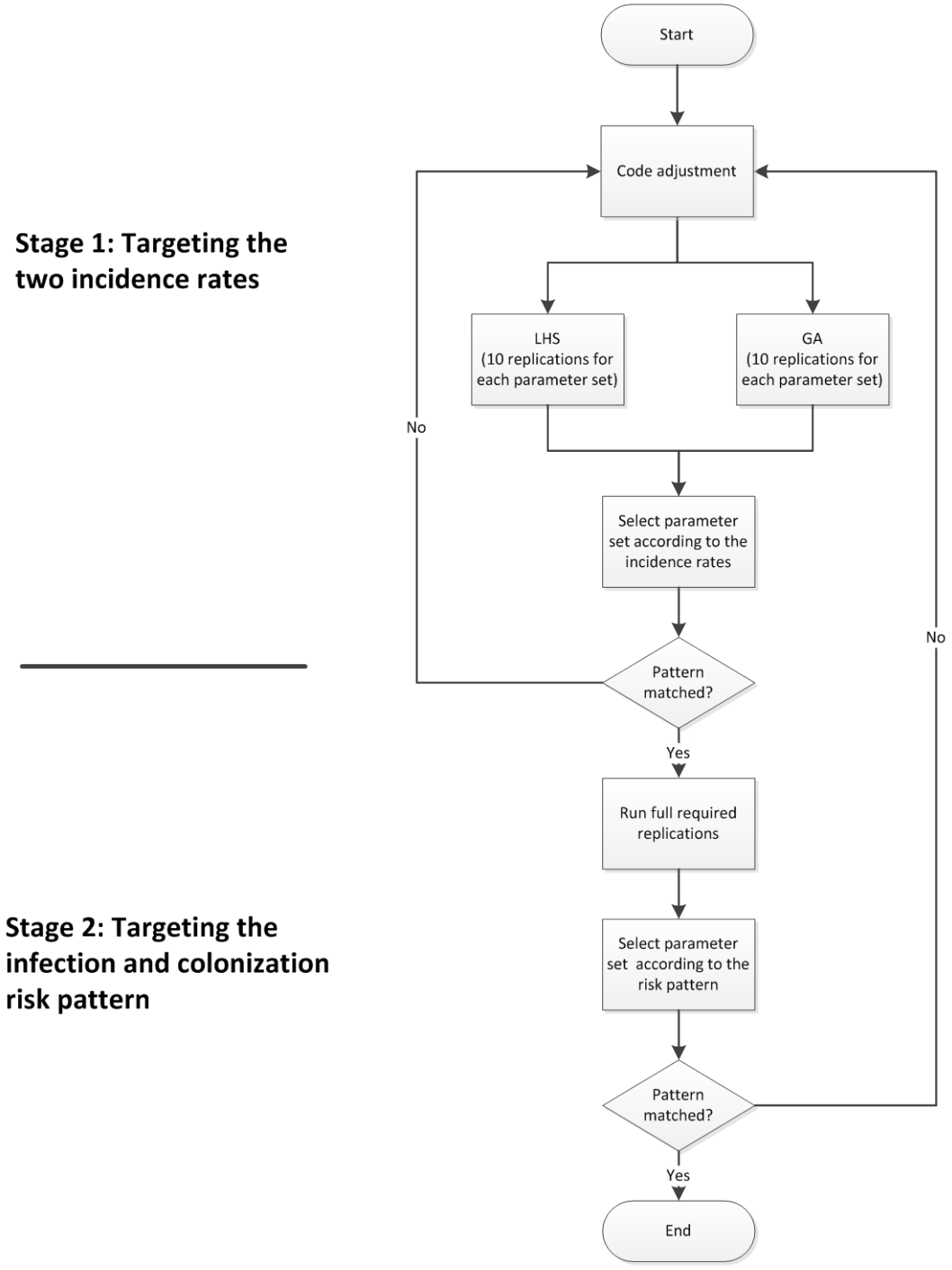


Figure C-1 Simulation model calibration process

Appendix D. Simulation model uncertainty and sensitivity analysis process and results

We exploit the NetLogo BehaviorSpace experiments' "headless" feature to run multiple experiments from the command line automatically (http://ccl.northwestern.edu/netlogo/docs/behaviorspace.html). R scripts were written to generate the experiment parameter sets and provide ".xml" experiment files for the NetLogo model. R scripts were also written to produce corresponding command lines for the execution of the experiments in windows CMD environment. The whole procedure is demonstrated in Figure 1.

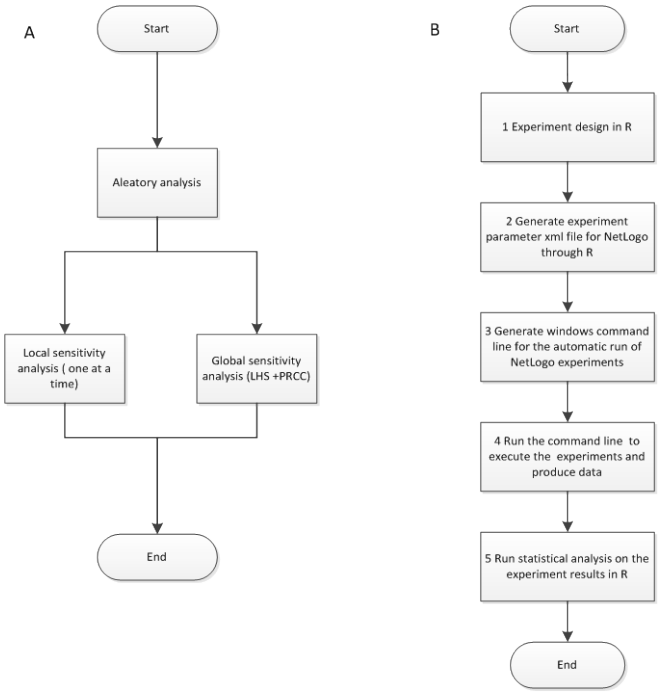


Figure D-1 Flow chart of the uncertainty and sensitivity analysis process

Appendix E. Simulation model explorations A-Test results

E.1 The effects of housekeeping

E.1.1 Incidence rate and infection risk

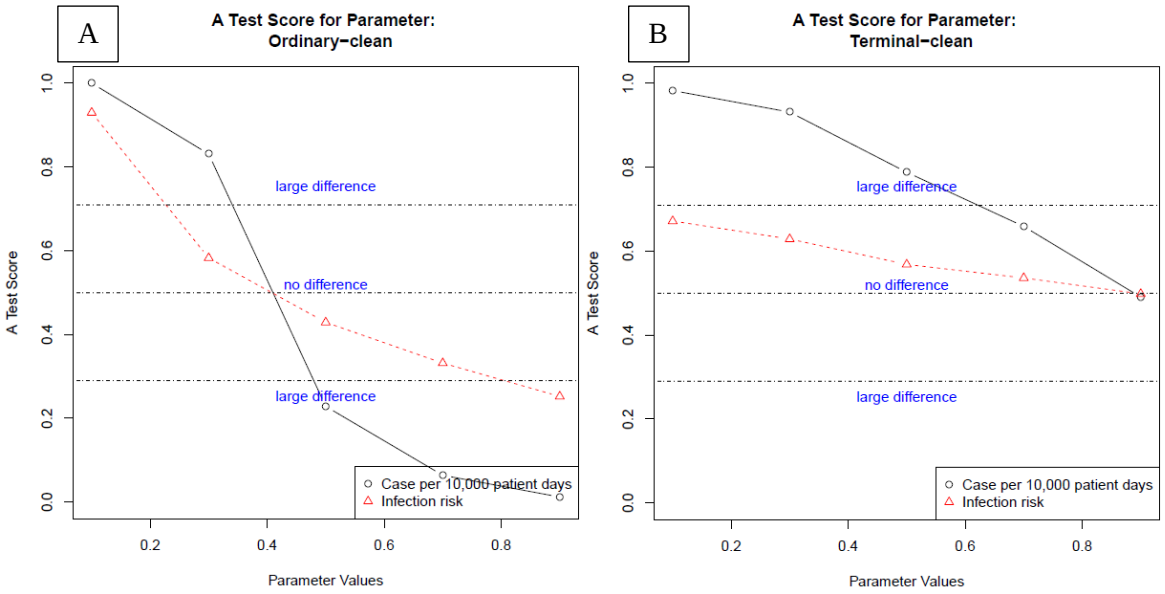


Figure E-1 A-Test scores of incidence rate and infection risk for housekeeping parameters

E.1.2 Sources of infection

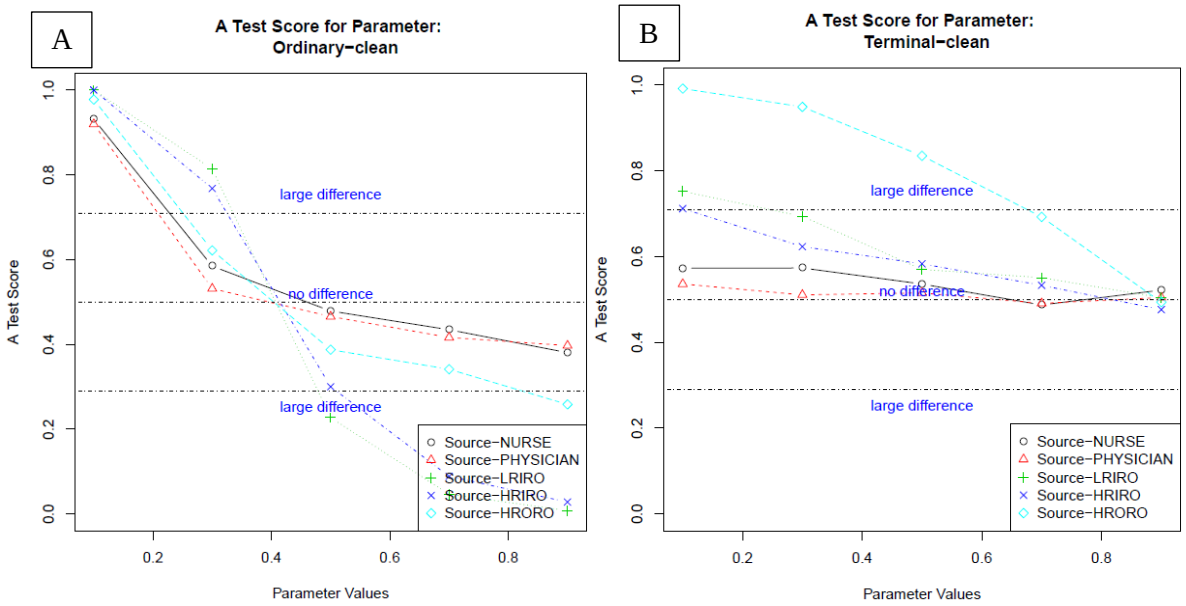


Figure E-2 A-Test scores of incidence rate and infection risk for housekeeping parameters

E.1.3 Age of bacteria

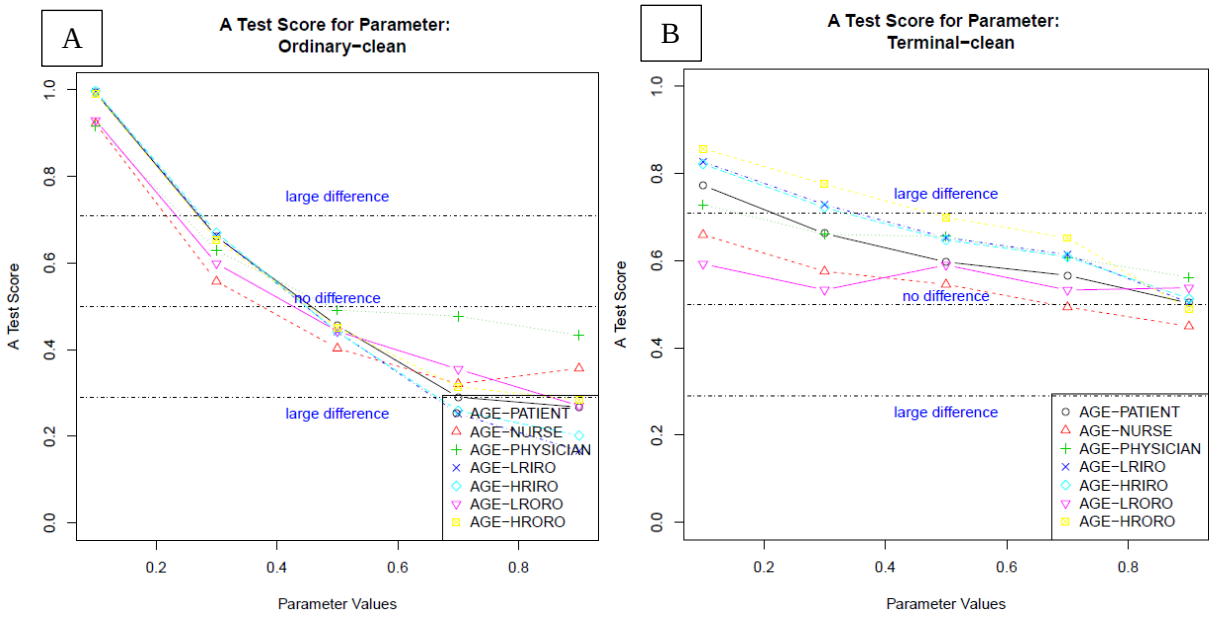


Figure E-3 A-Test scores of max bacteria age for housekeeping parameters

E.2 The effects of hand hygiene compliance

E.2.1 Incidence rate and infection risk

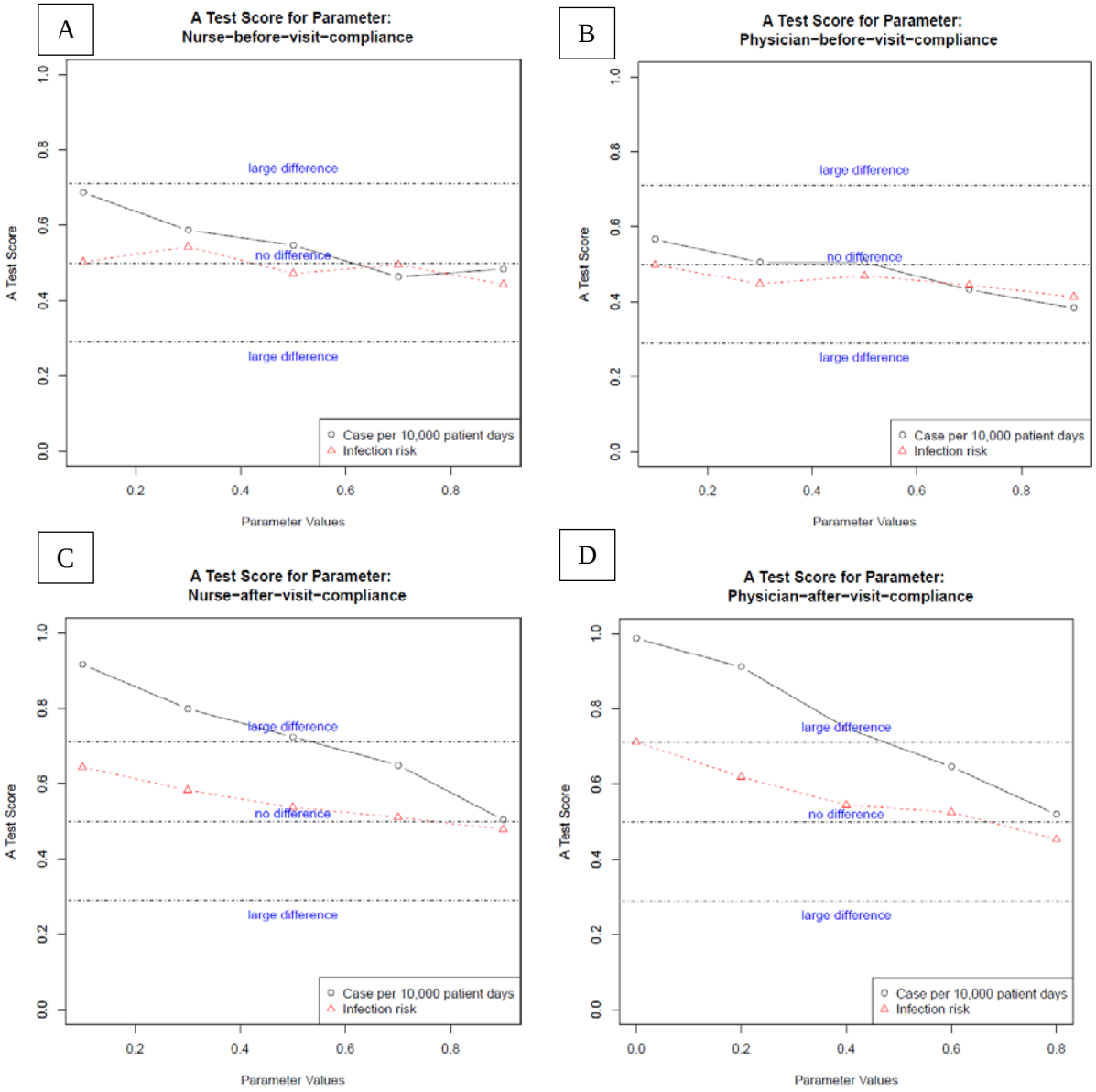


Figure E-4 A-Test scores of incidence rate and infection risk for hand hygiene parameters

E.2.2 Sources of infection

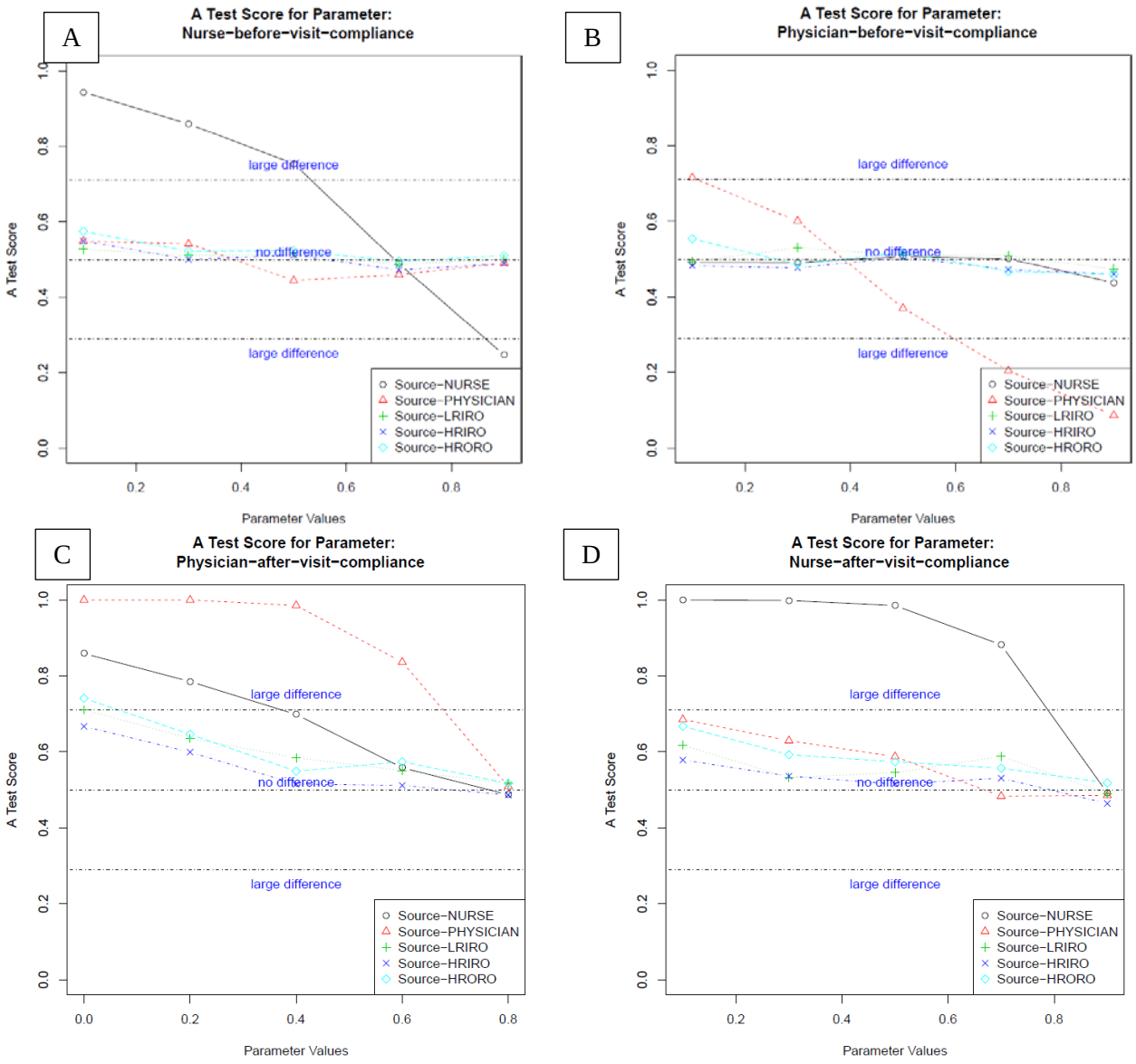


Figure E-5 A-Test scores of source of infection for hand hygiene parameters

E.2.3 Age of bacteria

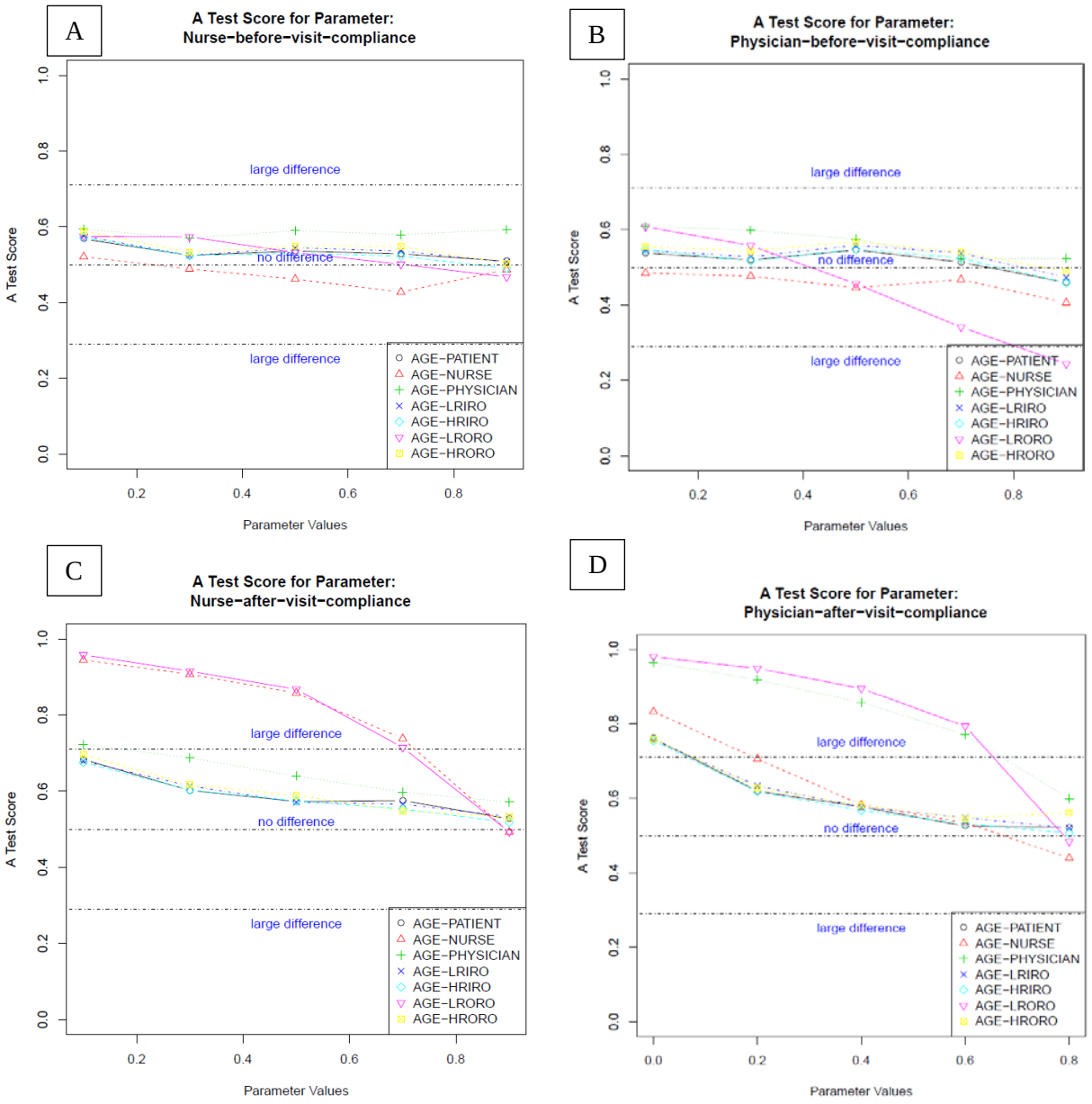


Figure E-6 A-Test scores of max bacteria age for hand hygiene parameters

E.3 The effects of patient turnover

E.3.1 Incidence rate and infection risk

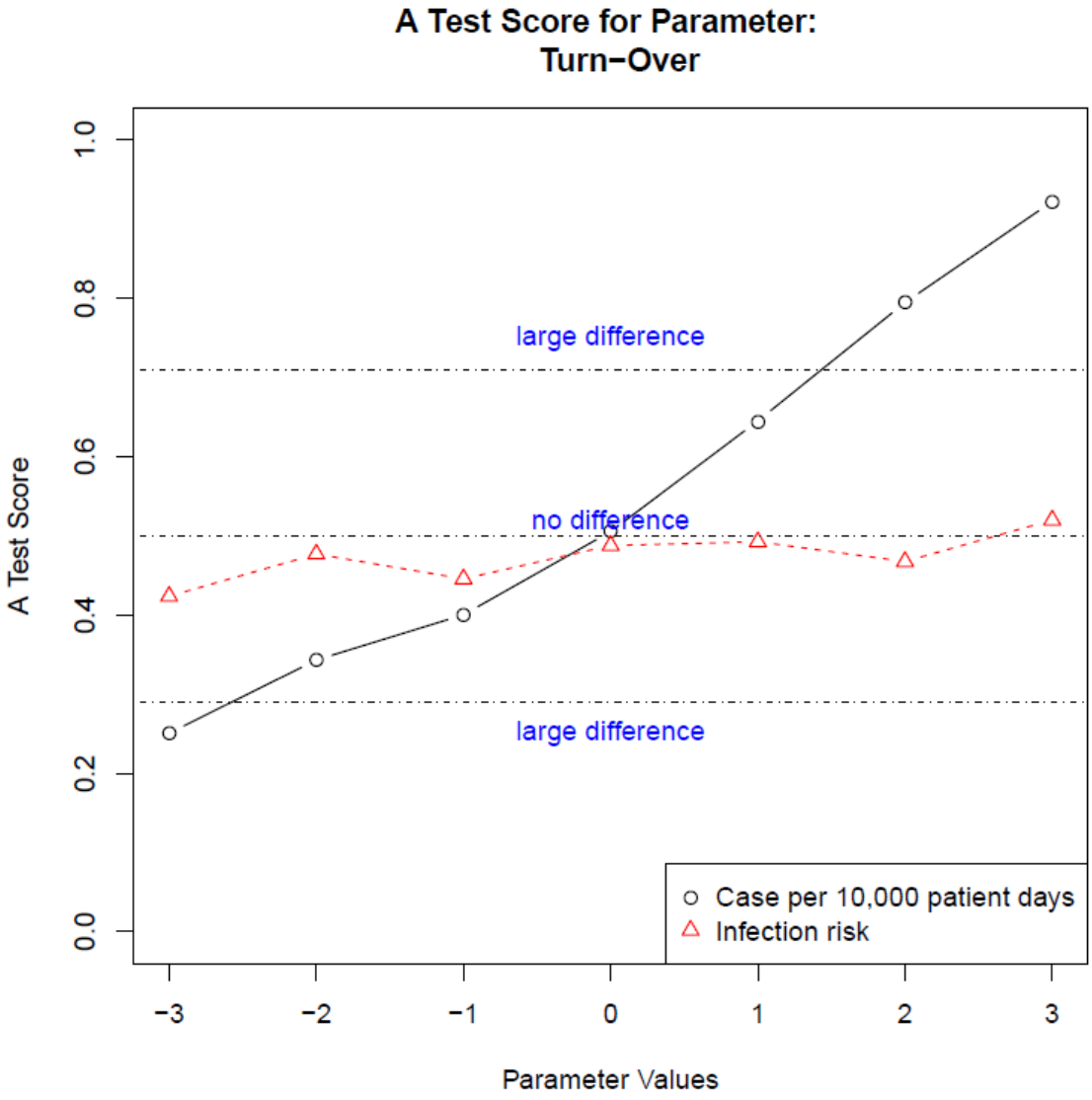


Figure E-7 A-Test scores of incidence rate and infection risk for patient turnover parameter

E.3.2 Sources of infection

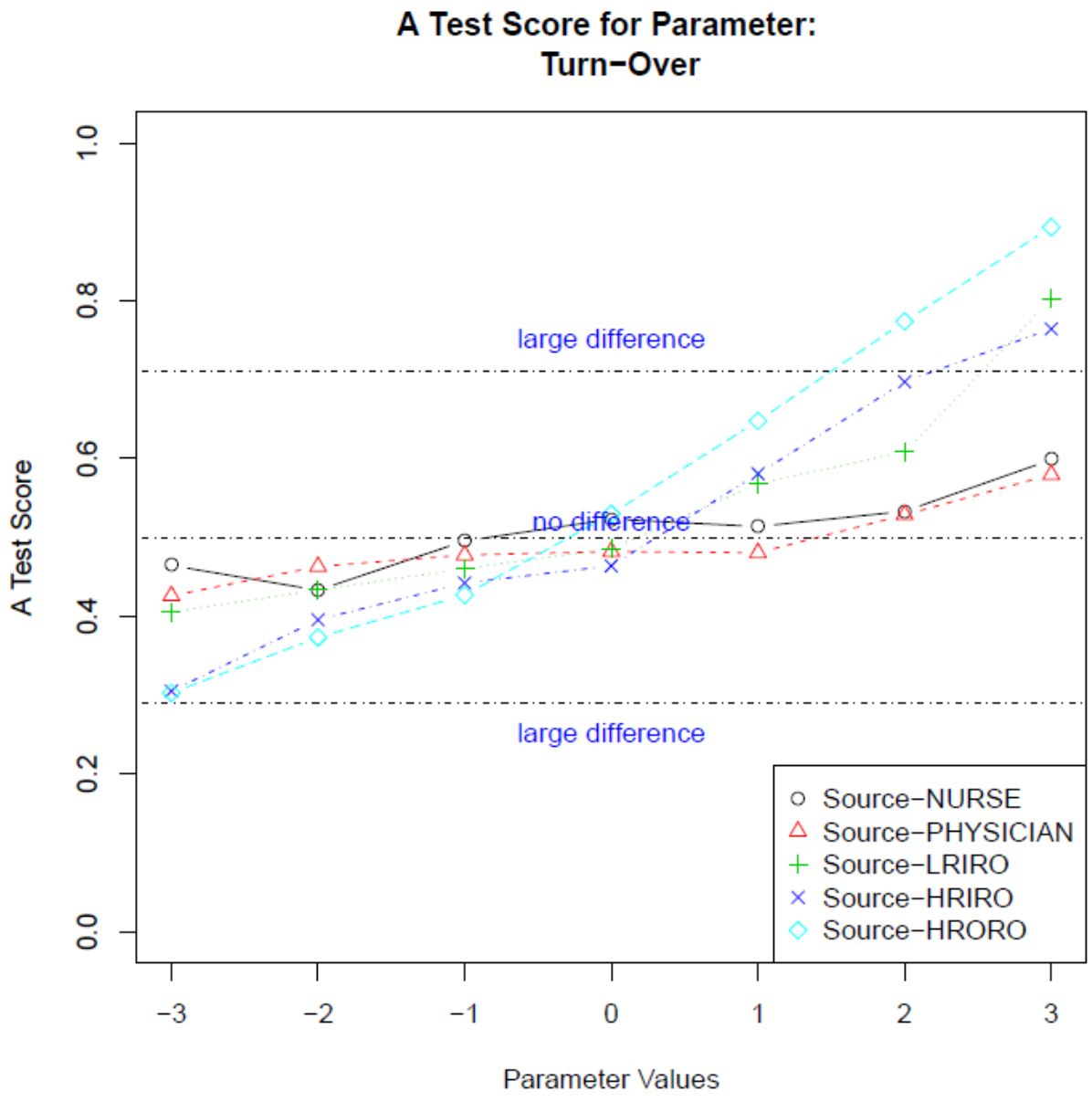


Figure E-8 A-Test scores of source of infection for patient turnover parameter

E.3.3 Age of bacteria

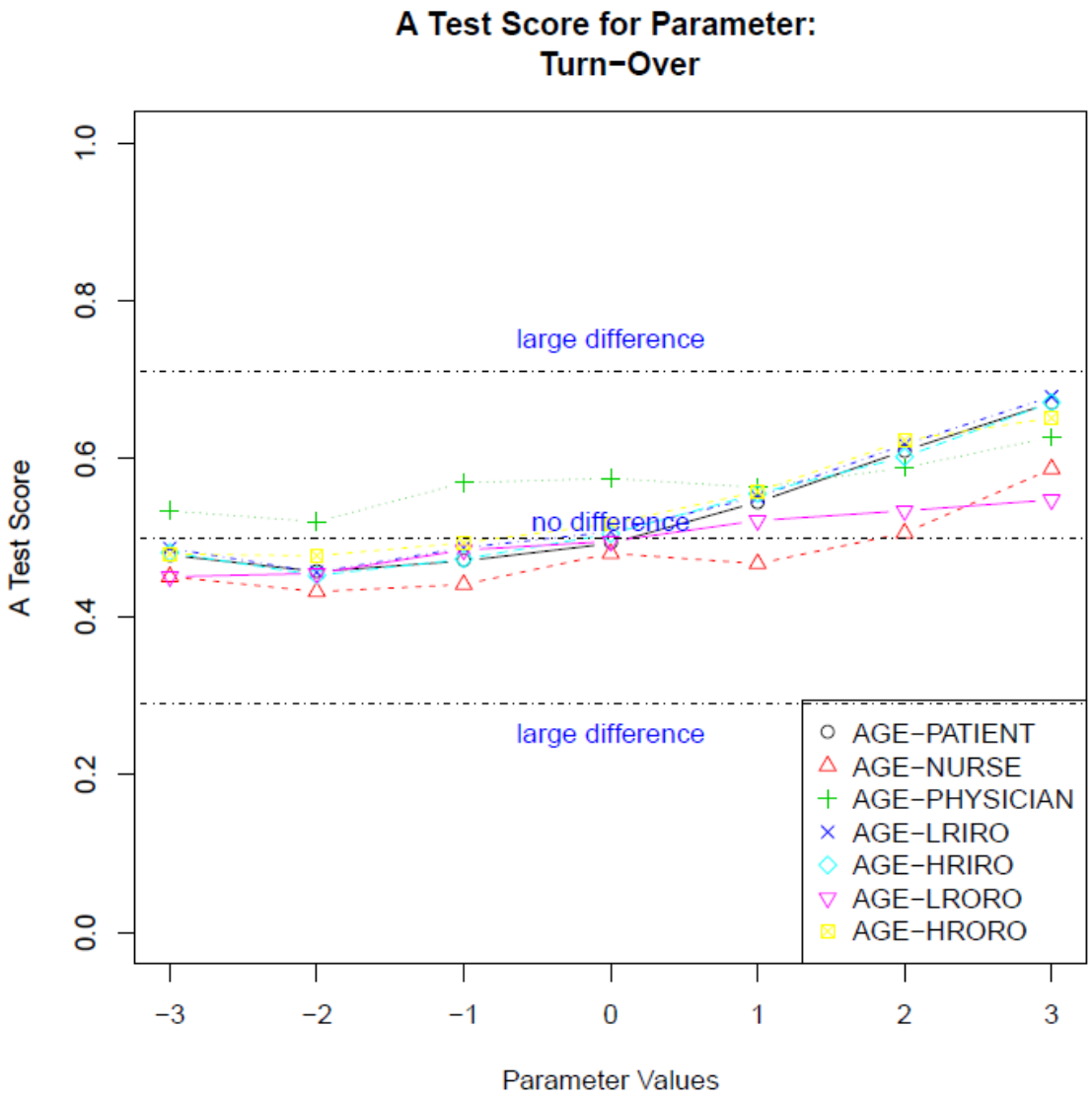


Figure E-9 A-Test scores of max bacteria age for patient turnover parameter

E.4 The effects of antibiotic pressure

E.4.1 Incidence rate and infection risk

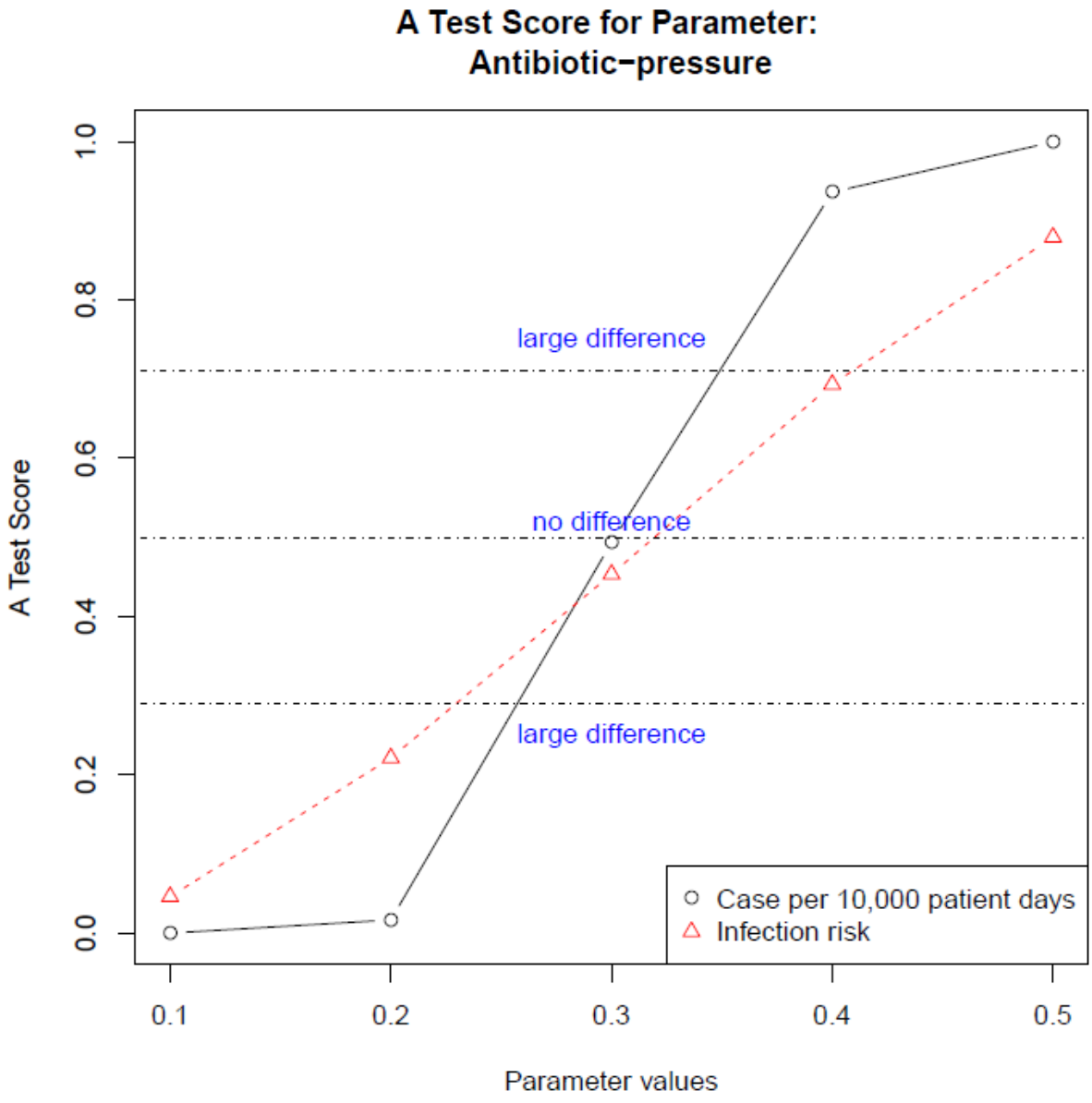


Figure E-10 A-Test scores of incidence rate and infection risk for antibiotic pressure parameter

E.4.2 Sources of infection

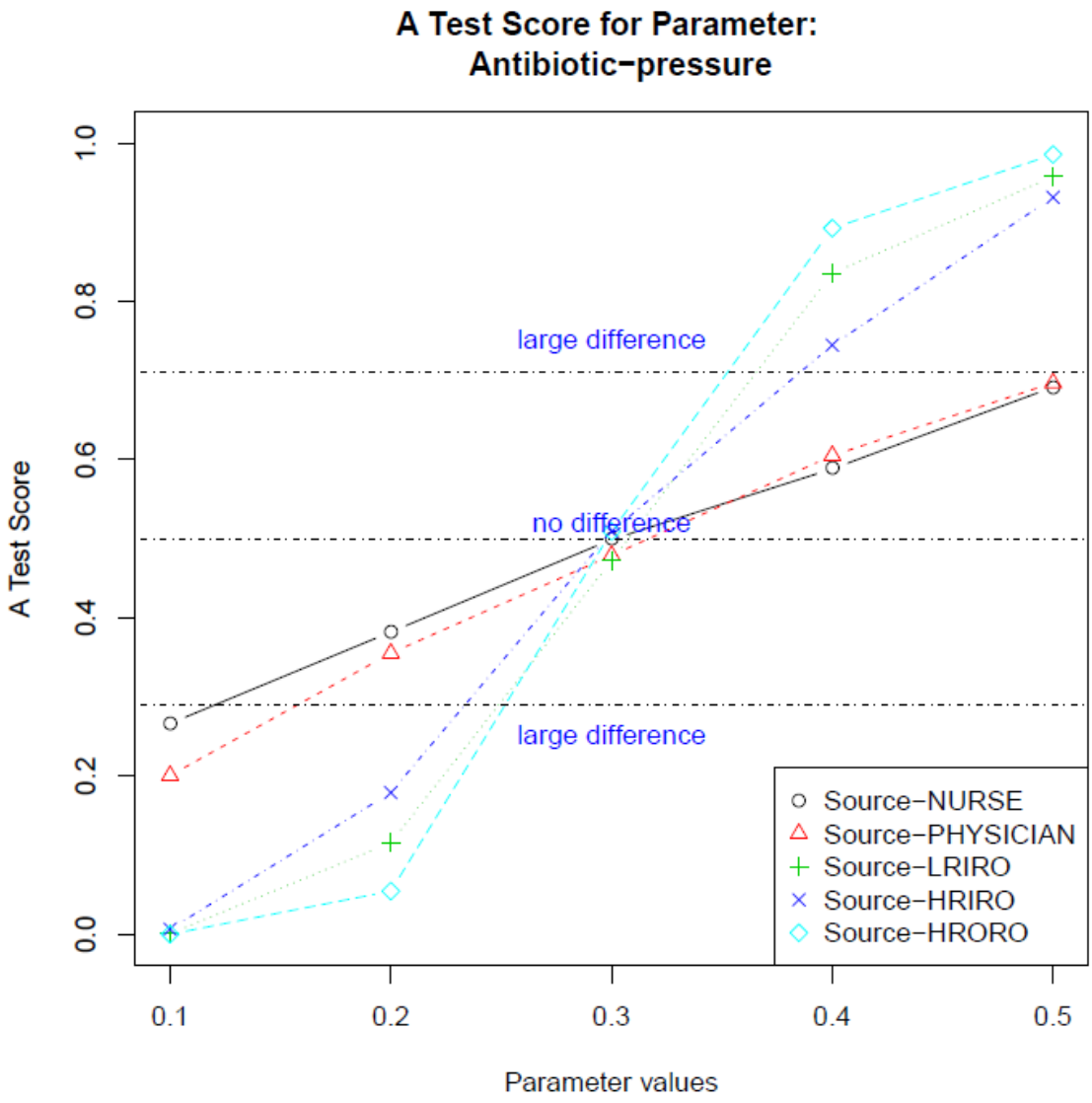


Figure E-11 A-Test scores of source of infection for antibiotic pressure parameter

E.4.3 Age of bacteria

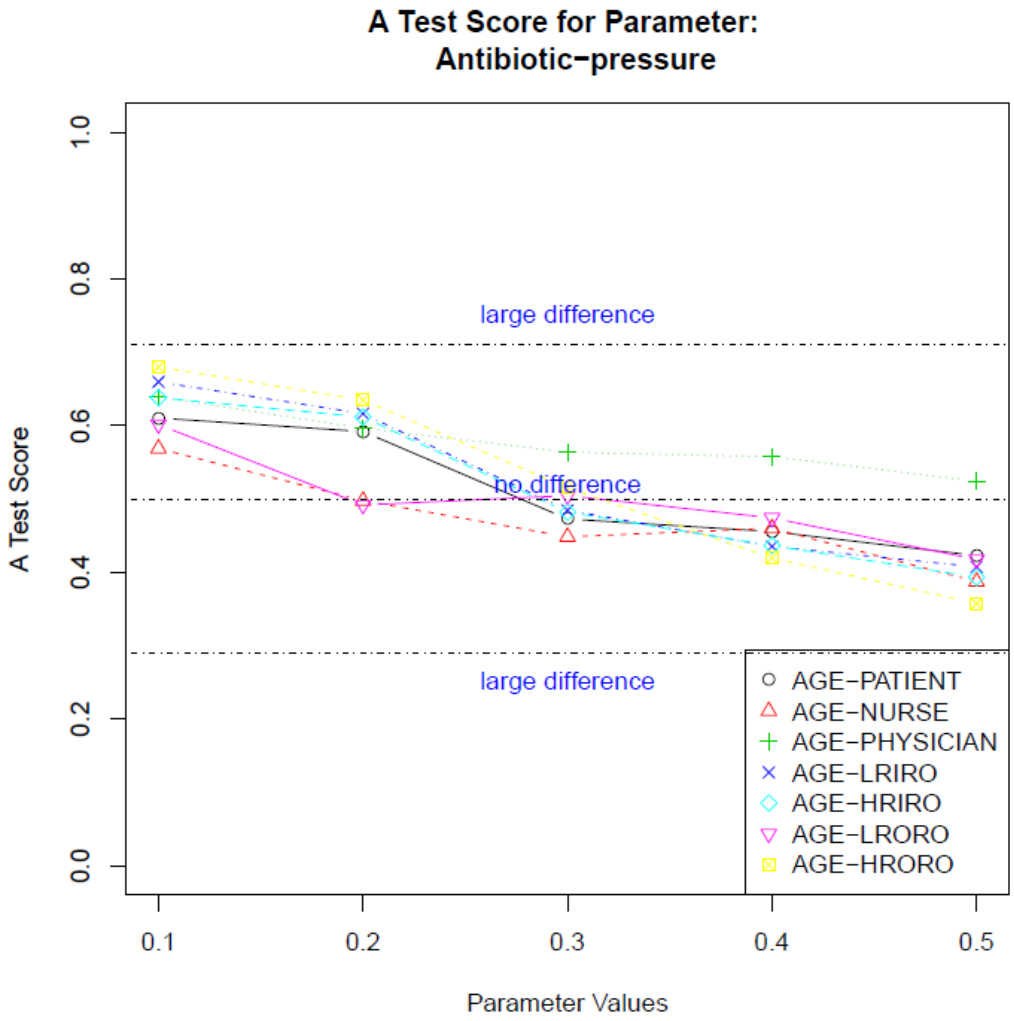


Figure E-12 A-Test scores of max bacteria age for antibiotic pressure parameter

Appendix F. Statistics of the pattern in network study

StartDate	EndDate	Window	CDI admission	HAI onset	Support	Confidence	Significance
01/01/2010	01/05/2014	1	187	124	0.1182	0.0909	0.1371
01/01/2010	01/05/2014	2	187	124	0.1182	0.1444	0.2177
01/01/2010	01/05/2014	3	187	124	0.1182	0.1818	0.2742
01/01/2010	01/05/2014	4	187	124	0.1182	0.2353	0.3548
01/01/2010	01/05/2014	5	187	124	0.1182	0.2674	0.4032
01/01/2010	01/05/2014	6	187	124	0.1182	0.2888	0.4355
01/01/2010	01/05/2014	7	187	124	0.1182	0.3155	0.4758
01/01/2010	01/05/2014	8	187	124	0.1182	0.3369	0.5081
01/01/2010	01/05/2014	9	187	124	0.1182	0.3583	0.5403
01/01/2010	01/05/2014	10	187	124	0.1182	0.3797	0.5726
01/01/2010	01/05/2014	11	187	124	0.1182	0.3904	0.5887
01/01/2010	01/05/2014	12	187	124	0.1182	0.3904	0.5887
01/01/2010	01/05/2014	13	187	124	0.1182	0.3957	0.5968
01/01/2010	01/05/2014	14	187	124	0.1182	0.4011	0.6048
01/01/2010	01/05/2014	15	187	124	0.1182	0.4064	0.6129
01/01/2010	01/05/2014	16	187	124	0.1182	0.4064	0.6129
01/01/2010	01/05/2014	17	187	124	0.1182	0.4118	0.6210
01/01/2010	01/05/2014	18	187	124	0.1182	0.4118	0.6210
01/01/2010	01/05/2014	19	187	124	0.1182	0.4118	0.6210
01/01/2010	01/05/2014	20	187	124	0.1182	0.4118	0.6210
01/01/2010	01/05/2014	21	187	124	0.1182	0.4118	0.6210
01/01/2010	01/05/2014	22	187	124	0.1182	0.4118	0.6210
01/01/2010	01/05/2014	23	187	124	0.1182	0.4118	0.6210
01/01/2010	01/05/2014	24	187	124	0.1182	0.4118	0.6210
01/01/2010	01/05/2014	25	187	124	0.1182	0.4118	0.6210
01/01/2010	01/05/2014	26	187	124	0.1182	0.4118	0.6210
01/01/2010	01/05/2014	27	187	124	0.1182	0.4118	0.6210
01/01/2010	01/05/2014	28	187	124	0.1182	0.4118	0.6210
01/01/2010	01/05/2014	29	187	124	0.1182	0.4118	0.6210
01/01/2010	01/05/2014	30	187	124	0.1182	0.4118	0.6210

Appendix G. Explanation for the diagnosis type

M Most Responsible Diagnosis (MRDx)

The Most Responsible Diagnosis (M) is the one diagnosis or condition that can be described as being most responsible for the patient's stay in a facility. If there is more than one such condition, the one held most responsible for the greatest portion of the length of stay or greatest use of resources (for example, operating room time or investigative technology) is selected.

If no interventions were performed, select the first-listed diagnosis as the Most Responsible Diagnosis.

If no definite diagnosis was made, the main symptom, abnormal finding or problem should be selected as the MRDx.

1 Pre-Admit Comorbidity

Diagnosis Type 1 is conditions that existed prior to admission and satisfies the requirements for determining comorbidity.

2 Post-Admit Comorbidity

Diagnosis Type 2 is conditions that arises post-admission and satisfies the requirements for determining comorbidity. In specific circumstances, diagnosis assigned to Diagnosis Type 2 will also be assigned a Diagnosis Prefix (Group 10 Field 01) of 5 or 6. Also see details on diagnosis Type 2 as a Service Transfer diagnosis.

3 Secondary Diagnosis

Diagnosis Type 3 is secondary diagnoses or conditions for which a patient may or may not have received treatment and/or does not satisfy the requirements for determining comorbidity. Diagnosis Type 3 is also assigned to diagnosis codes that were recorded to provide detail that in themselves do not represent a comorbidity. Diagnoses that are listed only on the front sheet, discharge summary, death certificate, history and physical or pre-operative anaesthetic consults qualify as Diagnosis Type 3. If there is physician documentation elsewhere in the chart that the condition affected the treatment received or required treatment beyond maintenance of the pre-existing condition or increased the length of stay (LOS) by 24 hours or more, it then must be determined if it is a comorbidity that should be assigned as Type 1 or Type 2. When Entry Code (Group 04 Field 06) is N (newborn), Diagnosis Type 3 cannot be applied to any code on the newborn's abstract. Also see details on asterisk code assigned Diagnosis Type 3 in Diagnosis Type 6 (Proxy Most Responsible Diagnosis).

W, X, Y Service Transfer Diagnosis

Service Transfer Diagnoses are codes that are assigned to diagnoses associated with a service transfer. The use of this Diagnosis Type is determined at the provincial/territorial or facility level because service transfer diagnoses are optional, except for a service transfer to Alternate Level of Care (ALC). Assign W, X or Y to the ICD-10-CA code associated with the first (W), second (X) or third (Y) Service Transfer line respectively. When a diagnosis is recorded with a Service Transfer Diagnosis (W, X, or Y), it is equivalent to a Diagnosis Type 1. Do not to repeat the service transfer Diagnosis Code (Group 10 Field 02) on the abstract as a Diagnosis Type 1. When a diagnosis is recorded as a Diagnosis Type 2 and also qualifies as a Service Transfer Diagnosis, it is mandatory to record the diagnosis as a Diagnosis Type 2. Facilities choosing to capture Service Transfer diagnoses must record the diagnosis twice—as Diagnosis Type 2 as well as a service transfer diagnosis (W, X or Y).

4 Morphology Code

Diagnosis Type 4 morphology codes are derived from ICD-O (International Classification of Diseases—Oncology) codes describing the type and behaviour of neoplasm. These codes are found in Chapter XXII: Morphology of Neoplasms of the Canadian Coding Standards for ICD-10-CA and CCI.

5 Admitting Diagnosis

Diagnosis Type 5 can be used to code the admitting diagnosis when it differs from the most responsible Diagnosis Code (Type M). Its use is determined at the provincial/territorial or facility level. Refer to the DAD provincial/territorial sections and facility policies to determine the use of this Diagnosis Type.

6 Proxy Most Responsible Diagnosis (MRDx)

Diagnosis Type 6 is assigned to a designated asterisk code in a dagger/asterisk convention when the condition it represents fulfills the requirements stated in the definition for Diagnosis Type (M). In morbidity coding, asterisk codes are manifestations of an underlying condition and, according to the World Health Organization (WHO) rules, must be sequenced following the code for the underlying cause. The underlying cause codes are identified with a dagger symbol in the ICD-10-CA classification. Diagnosis Type 6 is used on the second line of the diagnosis field of the abstract to indicate that the manifestation is the condition most responsible for the patient's stay in a facility. When the underlying condition meets the criteria for MRDx, or when it would be difficult to delineate whether it is the underlying condition or the manifestation that meets the criteria for MRDx, the asterisk code is assigned Diagnosis Type 3 (secondary diagnosis). The purpose of using Diagnosis Type 6 is to ensure that the case is grouped to a clinically appropriate Case Mix Group (CMG) within the CMG+ grouping methodology.

9 External Cause of Injury Code

A Diagnosis Type 9 is assigned to an external cause of injury code (Chapter XX: External Causes of Morbidity and Mortality in the Canadian Coding Standards for ICD-10-CA and CCI), place of occurrence code (U98.–Place of occurrence) or activity code (U99.–Activity). Chapter XX codes are mandatory for

use with codes in the range S00 to T98 Injury, poisoning and certain other consequences of external causes. Category U98.–Place of occurrence is mandatory with codes in the range W00 to Y34, with the exception of Y06 and Y07. Recording with Category U99.–(Activity) is optional.

0 Newborn

Diagnosis Type 0 is applicable to newborn codes only and when Admit Category (Group 04 Field 05) is N (Newborn). Healthy infant: where a code from category Z38. (Live born infant according to place of birth) is the MRDx (type M), all other diagnoses on the newborn abstract must be assigned Diagnosis Type (0). Unhealthy infant: where a code from the range P00-P96, or any other code indicating a significant condition (for example, any condition that meets the criteria for a comorbidity) in the newborn is the MRDx (type M), then Z.38 must be assigned Diagnosis Type (0). In this circumstance, Diagnosis Type (0) can be used to record any additional insignificant conditions that do not affect the newborn's treatment or length of stay and do not satisfy the requirements for determining comorbidity. Additional conditions that meet the criteria of comorbidity are assigned Diagnosis Types 1, 2, W, X or Y as indicated by the documentation in the chart. Diagnosis Type 3 cannot be applied to any code on a newborn's abstract.

Appendix H. Drug classification and coding system for the feature engineering of machine learning study

H.1 Antibiotics

Source: http://www.emedexpert.com/lists/antibiotics.shtml

Generic Name	Class
Amikacin	Aminoglycoside
Gentamicin	Aminoglycoside
Kanamycin	Aminoglycoside
Neomycin	Aminoglycoside
Netilmicin	Aminoglycoside
Paromomycin	Aminoglycoside
Streptomycin	Aminoglycoside
Tobramycin	Aminoglycoside
Cilastatin	Carbapenems
Doripenem	Carbapenems
Ertapenem	Carbapenems
Imipenem	Carbapenems
Meropenem	Carbapenems
Cefacetrile	Cephalosporins1
Cefadroxil	Cephalosporins1
cefadroxyl	Cephalosporins1
Cefalexin	Cephalosporins1
Cefaloglycin	Cephalosporins1
Cefalonium	Cephalosporins1
Cefaloridine	Cephalosporins1
Cefalotin	Cephalosporins1
Cefapirin	Cephalosporins1
Cefatrizine	Cephalosporins1
Cefazaflur	Cephalosporins1
Cefazedone	Cephalosporins1
Cefazolin	Cephalosporins1
Cefradine	Cephalosporins1

Cefroxadine	Cephalosporins1
Ceftezole	Cephalosporins1
Cephacetrile	Cephalosporins1
cephalexin	Cephalosporins1
Cephaloglycin	Cephalosporins1
Cephalonium	Cephalosporins1
Cephaloradine	Cephalosporins1
Cephalothin	Cephalosporins1
Cephapirin	Cephalosporins1
Cephazolin	Cephalosporins1
Cephradine	Cephalosporins1
Cefaclor	Cephalosporins2
Cefamandole	Cephalosporins2
Cefmetazole	Cephalosporins2
Cefonicid	Cephalosporins2
Cefotetan	Cephalosporins2
Cefoxitin	Cephalosporins2
Cefproxil	Cephalosporins2
Cefprozil	Cephalosporins2
Cefuroxime	Cephalosporins2
Cefuzonam	Cephalosporins2
Cefcapene	Cephalosporins3
Cefdaloxime	Cephalosporins3
Cefdinir	Cephalosporins3
Cefditoren	Cephalosporins3
Cefetamet	Cephalosporins3
Cefixime	Cephalosporins3
Cefmenoxime	Cephalosporins3
Cefodizime	Cephalosporins3
Cefoperazone	Cephalosporins3
Cefotaxime	Cephalosporins3
Cefpimizole	Cephalosporins3
Cefpodoxime	Cephalosporins3
Ceftazidime	Cephalosporins3
Cefteram	Cephalosporins3
Ceftibuten	Cephalosporins3
Ceftiofur	Cephalosporins3
Ceftiolene	Cephalosporins3
Ceftizoxime	Cephalosporins3
Ceftriaxone	Cephalosporins3
Cefclidine	Cephalosporins4

Cefepime	Cephalosporins4
Cefluprenam	Cephalosporins4
Cefoselis	Cephalosporins4
Cefozopran	Cephalosporins4
Cefpirome	Cephalosporins4
Cefquinome	Cephalosporins4
Ceftaroline	Cephalosporins5
Ceftobiprole	Cephalosporins5
Cefaclomezine	Cephalosporins6
Cefaloram	Cephalosporins6
Cefaparole	Cephalosporins6
Cefcanel	Cephalosporins6
Cefedrolor	Cephalosporins6
Cefempidone	Cephalosporins6
Cefetrizole	Cephalosporins6
Cefivitril	Cephalosporins6
Cefmatilen	Cephalosporins6
Cefmepidium	Cephalosporins6
Cefovecin	Cephalosporins6
Cefoxazole	Cephalosporins6
Cefrotil	Cephalosporins6
Cefsumide	Cephalosporins6
Ceftioxide	Cephalosporins6
Cefuracetime	Cephalosporins6
Teicoplanin	Glycopeptides
Clindamycin	Lincosamides
Lincomycin	Lincosamides
Telavancin	Lipoglycopeptides
Azithromycin	Macrolide
Clarithromycin	Macrolide
Dirithromycin	Macrolide
Erythromycin	Macrolide
Ketolides	Macrolide
Roxithromycin	Macrolide
Telithromycin	Macrolide
Aztreonam	Monobactams
Chloramphenicol	OtherAntibiotic
Fluoroquinolone	OtherAntibiotic
Lipoglycopeptide	OtherAntibiotic
Lipopeptide	OtherAntibiotic
Macrocyclics	OtherAntibiotic

Nitrofurantoin	OtherAntibiotic
Cycloserine	Oxazolidinones
Linezolid	Oxazolidinones
Amoxicillin	Penicillins
Ampicillin	Penicillins
Bacampicillin	Penicillins
Carbenicillin	Penicillins
Cloxacillin	Penicillins
Dicloxacillin	Penicillins
Flucloxacillin	Penicillins
Mezlocillin	Penicillins
Nafcillin	Penicillins
Oxacillin	Penicillins
Penicillin G	Penicillins
Penicillin V	Penicillins
Piperacillin	Penicillins
Pivampicillin	Penicillins
Pivmecillinam	Penicillins
Ticarcillin	Penicillins
Bacitracin	Polypeptides
Polymyxin	Polypeptides
Flumequine	Quinolone1
Nalidixic acid	Quinolone1
Oxolinic acid	Quinolone1
Pipemidic acid	Quinolone1
Piromidic acid	Quinolone1
Rosoxacin	Quinolone1
Ciprofloxacin	Quinolone2
Enoxacin	Quinolone2
Lomefloxacin	Quinolone2
Nadifloxacin	Quinolone2
Norfloxacin	Quinolone2
Ofloxacin	Quinolone2
Pefloxacin	Quinolone2
Rufloxacin	Quinolone2
Balofloxacin	Quinolone3
Gatifloxacin	Quinolone3
Grepafloxacin	Quinolone3
Levofloxacin	Quinolone3
Moxifloxacin	Quinolone3
Pazufloxacin	Quinolone3

Sparfloxacin	Quinolone3
Temafloxacin	Quinolone3
Tosufloxacin	Quinolone3
Besifloxacin	Quinolone4
Clinafloxacin	Quinolone4
Gemifloxacin	Quinolone4
Prulifloxacin	Quinolone4
Sitafloxacin	Quinolone4
Trovafloxacin	Quinolone4
Rifabutin	Rifamycins
Rifampin	Rifamycins
Rifapentine	Rifamycins
Dalfopristin	Stretogramins
Pristinamycin	Stretogramins
Quinupristin	Stretogramins
Sulfamethizole	Sulfonamides
Sulfamethoxazole	Sulfonamides
Sulfamethoxazole	Sulfonamides
Sulfisoxazole	Sulfonamides
Trimethoprim	Sulfonamides
Demeclocycline	Tetracycline
Doxycycline	Tetracycline
Doxycycline	Tetracycline
Minocycline	Tetracycline
Oxytetracycline	Tetracycline
Tetracycline	Tetracycline
Tigecycline	Tetracycline
Capreomycin	Tuberactinomycins
Viomycin	Tuberactinomycins
Metronidazole	ZZMetronidazole
Nitazoxanide	ZZNitazoxanide
Ramoplanin	ZZRamoplanin
Rifaximin	ZZRifaximin
Tinidazole	ZZTinidazole
Tolevamer	ZZTolevamer
Vancomycin	ZZVancomycin
CHLORHEXIDINE	OtherAntibiotic
CIPRO	OtherAntibiotic
DAPSONE	OtherAntibiotic
DAPTOMYCIN	OtherAntibiotic
ETHAMBUTOL	OtherAntibiotic

FUSIDATE	OtherAntibiotic
FUSIDIC	OtherAntibiotic
ISONIAZID	OtherAntibiotic
MUPIROCIN	OtherAntibiotic
NITROFURANTOIN	OtherAntibiotic
PENTAMIDINE	OtherAntibiotic
POLYSPORIN	OtherAntibiotic
PYRAZINAMIDE	OtherAntibiotic
SILVER SULFASALAZINE	OtherAntibiotic
SULFASALAZINE	OtherAntibiotic

H.2 Antiviral

Generic Name	**Class**
ACYCLOVIR	AntiViral
FAMCICLOVIR	AntiViral
GANCICLOVIR	AntiViral
HYDROXYCHLOROQUINE	AntiViral
KALETRA	AntiViral
LAMIVUDINE	AntiViral
OSELTAMIVIR	AntiViral
QUININE	AntiViral

H.3 Antifungal

Generic Name	Class
AMPHOTERICIN	AntiFungal
CASPOFUNGIN	AntiFungal
CLOTRIMAZOLE	AntiFungal
CLOTRIMAZOLE	AntiFungal
CLOTRIMAZOLE	AntiFungal
FLUCONAZOLE	AntiFungal
ITRACONAZOLE	AntiFungal
KETOCONAZOLE	AntiFungal
MICONAZOLE	AntiFungal

NYSTATIN	AntiFungal
TERBINAFINE	AntiFungal

H.4 PPI

Generic Name	Class
ESOMEPRAZOLE	PPI
LANSOPRAZOLE	PPI
OMEPRAZOLE	PPI
PANTOPRAZOLE	PPI
RABEPRAZOLE	PPI

H.5 Immunosuppression

Generic Name	**Class**
BETAMETHASONE	Immunosuppressive
CHLOROQUINE	Immunosuppressive
HYDROCORTISONE	Immunosuppressive
ORABASE	Immunosuppressive

H.6 Corticosteroid

Generic Name	**Class**
CLOBETASOL	Corticosteroid
FLUOCINONIDE	Corticosteroid
TRIAMCINOLONE	Corticosteroid

H.7 Others

Generic Name	Class
ANUSOL PLUS	Others
PROCTODAN-HC	Others
CALAMINE LOTION	Others
CALMOSEPTINE	Others
SILICONE CREAM	Others
TUCKS	Others
UREMOL 10% CREAM	Others
VORICONAZOLE	Others
ANTIPHLOGISTINE	Others
UREA 10% CREAM	Others
OLIVE OIL	Others
LIDOCAINE	Others
TARO BASE	Others
EUCERIN 10% LOTN	Others
WHITE PETROLATUM	Others
LIDOCAINE-PRILOCAINE 2.5 %-2.5 %	Others
VITAMIN E	Others
MINERAL OIL LIGHT	Others
POVIDONE-IODINE 10 %	Others
ZINC OXIDE 15 %	Others
PERMETHRIN 1 %	Others
PYRETHRINS-PIPERONYL BUTOXIDE	Others
UDDERLY SMOOTH CREAM	Others
AVEENO 43 % PWD	Others
DERMABASE	Others
COAL TAR 3 %	Others
DIPHENHYDRAMINE 2 %	Others
CAPSAICIN	Others
GLAXAL BASE	Others
IODINE TINCTURE 2%	Others
FOSCARNET	Others

Appendix I. Predictors in machine learning models

Index	Name
1	ADMIT_YEAR
2	LOS_DAY
3	AGE
4	ADMIT_JAN
5	ADMIT_FEB
6	ADMIT_MAR
7	ADMIT_APR
8	ADMIT_MAY
9	ADMIT_JUNE
10	ADMIT_JULY
11	ADMIT_AUG
12	ADMIT_SEP
13	ADMIT_OCT
14	ADMIT_NOV
15	ADMIT_DEC
16	PRIOR_ADM_YES
17	PRIOR_ADM_UNK
18	TYPE_CAP
19	TYPE_CHR
20	TYPE_INP
21	TYPE_IPO
22	TYPE_OBS
23	TYPE_PED
24	TYPE_RHB
25	GENDER_F
26	C_ICD10_MR
27	D_ICD10_MR
28	E_ICD10_MR
29	F_ICD10_MR
30	G_ICD10_MR
31	H_ICD10_MR
32	I_ICD10_MR
33	J_ICD10_MR
34	K_ICD10_MR
35	L_ICD10_MR
36	M_ICD10_MR
37	N_ICD10_MR

38	O_ICD10_MR
39	P_ICD10_MR
40	Q_ICD10_MR
41	R_ICD10_MR
42	S_ICD10_MR
43	T_ICD10_MR
44	Z_ICD10_MR
45	C_ICD10_PRE_COM
46	A_ICD10_PRE_COM
47	B_ICD10_PRE_COM
48	D_ICD10_PRE_COM
49	E_ICD10_PRE_COM
50	F_ICD10_PRE_COM
51	G_ICD10_PRE_COM
52	H_ICD10_PRE_COM
53	I_ICD10_PRE_COM
54	J_ICD10_PRE_COM
55	K_ICD10_PRE_COM
56	L_ICD10_PRE_COM
57	M_ICD10_PRE_COM
58	N_ICD10_PRE_COM
59	O_ICD10_PRE_COM
60	P_ICD10_PRE_COM
61	Q_ICD10_PRE_COM
62	R_ICD10_PRE_COM
63	S_ICD10_PRE_COM
64	T_ICD10_PRE_COM
65	U_ICD10_PRE_COM
66	Z_ICD10_PRE_COM
67	D_ICD10_POST_COM
68	E_ICD10_POST_COM
69	F_ICD10_POST_COM
70	G_ICD10_POST_COM
71	H_ICD10_POST_COM
72	I_ICD10_POST_COM
73	J_ICD10_POST_COM
74	K_ICD10_POST_COM
75	L_ICD10_POST_COM
76	M_ICD10_POST_COM
77	N_ICD10_POST_COM
78	O_ICD10_POST_COM

79	R_ICD10_POST_COM
80	S_ICD10_POST_COM
81	T_ICD10_POST_COM
82	U_ICD10_POST_COM
83	Z_ICD10_POST_COM
84	C_ICD10_SEC
85	D_ICD10_SEC
86	E_ICD10_SEC
87	F_ICD10_SEC
88	G_ICD10_SEC
89	H_ICD10_SEC
90	I_ICD10_SEC
91	J_ICD10_SEC
92	K_ICD10_SEC
93	L_ICD10_SEC
94	M_ICD10_SEC
95	N_ICD10_SEC
96	O_ICD10_SEC
97	P_ICD10_SEC
98	Q_ICD10_SEC
99	R_ICD10_SEC
100	S_ICD10_SEC
101	T_ICD10_SEC
102	U_ICD10_SEC
103	Z_ICD10_SEC
104	DIAG_ALL
105	CCI_1A
106	CCI_1B
107	CCI_1E
108	CCI_1F
109	CCI_1G
110	CCI_1I
111	CCI_1J
112	CCI_1K
113	CCI_1M
114	CCI_1N
115	CCI_1O
116	CCI_1P
117	CCI_1Q
118	CCI_1R
119	CCI_1S

120	CCI_1T
121	CCI_1U
122	CCI_1V
123	CCI_1W
124	CCI_1Y
125	CCI_2E
126	CCI_2F
127	CCI_2G
128	CCI_2M
129	CCI_2N
130	CCI_2O
131	CCI_2P
132	CCI_2R
133	CCI_2S
134	CCI_2V
135	CCI_5C
136	CCI_5P
137	ABD_INTVN
138	DRUG_Aminoglycoside_DOSE
139	DRUG_AntiFungal_DOSE
140	DRUG_AntiViral_DOSE
141	DRUG_Carbapenems_DOSE
142	DRUG_Cephalosporins1_DOSE
143	DRUG_Cephalosporins2_DOSE
144	DRUG_Cephalosporins3_DOSE
145	DRUG_Corticosteriod_DOSE
146	DRUG_Immunosuppressive_DOSE
147	DRUG_Lincosamides_DOSE
148	DRUG_Macrolide_DOSE
149	DRUG_OtherAntibiotic_DOSE
150	DRUG_Others_DOSE
151	DRUG_Oxazolidinones_DOSE
152	DRUG_Penicillins_DOSE
153	DRUG_PPI_DOSE
154	DRUG_Quinolone2_DOSE
155	DRUG_Quinolone3_DOSE
156	DRUG_Rifamycins_DOSE
157	DRUG_Sulfonamides_DOSE
158	DRUG_Tetracycline_DOSE
159	DRUG_Aminoglycoside_DURATION
160	DRUG_AntiFungal_DURATION

161	DRUG_AntiViral_DURATION
162	DRUG_Carbapenems_DURATION
163	DRUG_Cephalosporins1_DURATION
164	DRUG_Cephalosporins2_DURATION
165	DRUG_Cephalosporins3_DURATION
166	DRUG_Corticosteriod_DURATION
167	DRUG_Immunosuppressive_DURATION
168	DRUG_Lincosamides_DURATION
169	DRUG_Macrolide_DURATION
170	DRUG_OtherAntibiotic_DURATION
171	DRUG_Others_DURATION
172	DRUG_Oxazolidinones_DURATION
173	DRUG_Penicillins_DURATION
174	DRUG_PPI_DURATION
175	DRUG_Quinolone2_DURATION
176	DRUG_Quinolone3_DURATION
177	DRUG_Rifamycins_DURATION
178	DRUG_Sulfonamides_DURATION
179	DRUG_Tetracycline_DURATION

References

Adrie, C., Francais, A., Alvarez-Gonzalez, A., Mounier, R., Azoulay, E., Zahar, J.-R., . . . Descorps-Declere, A. (2009). Model for predicting short-term mortality of severe sepsis. *Crit Care, 13*(3), R72.

Agarwal, D., McGregor, A., Phillips, J. M., Venkatasubramanian, S., & Zhu, Z. (2006). *Spatial scan statistics: approximations and performance study.* Paper presented at the Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining.

Aggarwal, C. C. (2011). *An introduction to social network data analytics*: Springer.

Aguiar, F. S., Almeida, L. L., Ruffino-Netto, A., Kritski, A. L., Mello, F. C., & Werneck, G. L. (2012). Classification and regression tree (CART) model to predict pulmonary tuberculosis in hospitalized patients. *BMC pulmonary medicine, 12*(1), 40.

Alden, K., Read, M., Timmis, J., Andrews, P. S., Veiga-Fernandes, H., & Coles, M. (2013). Spartan: A comprehensive tool for understanding uncertainty in simulations of biological systems. *PLoS computational biology, 9*(2), e1002916.

Alfa, M. J., Kabani, A., Lyerly, D., Moncrief, S., Neville, L. M., Al-Barrak, A., Embil, J. M. (2000). Characterization of a toxin A-negative, toxin B-positive strain of *Clostridium difficile* responsible for a nosocomial outbreak of *Clostridium difficile*-associated diarrhea. *Journal of clinical microbiology, 38*(7), 2706-2714.

Al-Hasan, M. N., Lahr, B. D., Eckel-Passow, J. E., & Baddour, L. M. (2013). Predictive scoring model of mortality in Gram-negative bloodstream infection. *Clinical Microbiology and Infection, 19*(10), 948-954.

Allegranzi, B., & Pittet, D. (2007). Healthcare-associated infection in developing countries: simple solutions to meet complex challenges. *Infection control and hospital epidemiology, 28*(12), 1323-1327.

Annex, C. (2013). Testing, Surveillance and Management of *Clostridium difficile*.

Apte, M., Landers, T., Furuya, Y., Hyman, S., & Larson, E. (2011). Comparison of two computer algorithms to identify surgical site infections. *Surgical infections, 12*(6), 459-464.

Bansal, S., Grenfell, B. T., & Meyers, L. A. (2007). When individual behaviour matters: homogeneous and network models in epidemiology. *Journal of the Royal Society Interface, 4*(16), 879-891.

Barnes, S. (2012). An Agent-Based Modeling Approach to Reducing Pathogenic Transmission in Medical Facilities and Community Populations.

Barnes, S., Golden, B., & Wasil, E. (2010). *A dynamic patient network model of hospital-acquired infections.* Paper presented at the Simulation Conference (WSC), Proceedings of the 2010 Winter Simulation Conference.

Bennett, J. V., Jarvis, W. R., & Brachman, P. S. (2007). *Bennett and Brachman's Hospital Infections, 5e*: Lippincott Williams & Wilkins.

Berner, E. S. (2007). *Clinical Decision Support Systems*: Springer.

Best, E. L., Fawley, W. N., Parnell, P., & Wilcox, M. H. (2010). The potential for airborne dispersal of *Clostridium difficile* from symptomatic patients. *Clinical Infectious Diseases, 50*(11), 1450-1457.

Biswal, S. (2014). Proton pump inhibitors and risk for *Clostridium difficile* associated diarrhea. *Biomedical journal, 37*(4), 178.

Bouzbid, S., Gicquel, Q., Gerbier, S., Chomarat, M., Pradat, E., Fabry, J., . . . Metzger, M.-H. (2011). Automated detection of nosocomial infections: evaluation of different strategies in an intensive care unit 2000–2006. *Journal of Hospital Infection, 79*(1), 38-43.

Bradley, C. A., Rolka, H., Walker, D., & Loonsk, J. (2005). BioSense: implementation of a national early event detection and situational awareness system. *MMWR Morb Mortal Wkly Rep, 54*(Suppl), 11-19.

Brandes, U., & Erlebach, T. (2005). *Network analysis: methodological foundations* (Vol. 3418): Springer.

Breslow, N. E. (1996). Statistics in epidemiology: the case-control study. *Journal of the American Statistical Association, 91*(433), 14-28.

Broderick, A., Mori, M., Nettleman, M. D., Streed, S. A., & Wenzel, R. P. (1990). Noaoxomial infections: validation of surveillance and computer modeling ot indentify patient at risk. *American journal of epidemiology, 131*(4), 734-742.

Brown, E., Talbot, G. H., Axelrod, P., Provencher, M., & Hoegg, C. (1990). Risk factors for *Clostridium difficile* toxin-associated diarrhea. *Infection Control and Hospital Epidemiology*, 283-290.

Buckeridge, D. L. (2007). Outbreak detection through automated surveillance: A review of the determinants of detection. *Journal of Biomedical Informatics, 40*(4), 370-379. doi: 10.1016/j.jbi.2006.09.003

Buckeridge, D. L., Burkom, H., Campbell, M., Hogan, W. R., & Moore, A. W. (2005). Algorithms for rapid outbreak detection: a research synthesis. *Journal of Biomedical Informatics, 38*(2), 99-113.

Burke, J. P. (2003). Infection control-a problem for patient safety. *New England Journal of Medicine, 348*(7), 651-656.

Campbell, K., Phillips, M., Stachel, A., Bosco III, J., & Mehta, S. (2013). Incidence and risk factors for hospital-acquired *Clostridium difficile* infection among inpatients in an orthopaedic tertiary care hospital. *Journal of Hospital Infection, 83*(2), 146-149.

Canadian Classification of Health Interventions. (2014) Retrieved Oct, 31, 2014, from http://www.cihi.ca/CIHI-ext-portal/internet/en/document/standards+and+data+submission/standards/classification+and+coding/codingclass_cci

Carnevale, R. J., Talbot, T. R., Schaffner, W., Bloch, K. C., Daniels, T. L., & Miller, R. A. (2011). Evaluating the utility of syndromic surveillance algorithms for screening to detect potentially clonal hospital infection outbreaks. *Journal of the American Medical Informatics Association, 18*(4), 466-472.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly Detection: A Survey. *Acm Computing Surveys, 41*(3). doi: 10.1145/1541880.1541882

Chandra, S., Latt, N., Jariwala, U., Palabindala, V., Thapa, R., Alamelumangapuram, C. B., . . . Jani, N. (2012). A cohort study for the derivation and validation of a clinical prediction scale for hospital-onset Clostridium difficile infection. *Canadian Journal of Gastroenterology, 26*(12), 885.

Chen, H., Zeng, D., & Yan, P. (2010). *Infectious disease informatics: syndromic surveillance for public health and bio-defense* (Vol. 21): Springer Science & Business Media.

Choudhuri, J. A., Pergamit, R. F., Chan, J. D., Schreuder, A. B., McNamara, E., Lynch, J. B., & Dellit, T. H. (2011). An electronic catheter-associated urinary tract infection surveillance tool. *Infection control and hospital epidemiology, 32*(8), 757-762.

Chung, Y., Lo, Y.-S., Lee, W.-S., Hsu, M.-H., & Liu, C.-T. (2009). *Faster and Active Surveillance of Hospital-Acquired Infections: A Model for Settings with High Sensitivity Predictors.* Paper presented at the Bioinformatics and BioEngineering, 2009. BIBE'09. Ninth IEEE International Conference on.

CIHI. (2012). Highlights of 2010-2011 Inpatient Hospitalizations and Emergency Department Visits. https://secure.cihi.ca/estore/productFamily.htm?locale=en&pf=PFC1840.

Cimiotti, J. P., Aiken, L. H., Sloane, D. M., & Wu, E. S. (2012). Nurse staffing, burnout, and health care–associated infection. *American Journal of Infection Control, 40*(6), 486-490.

Clabots, C. R., Johnson, S., Olson, M. M., Peterson, L. R., & Gerding, D. N. (1992). Acquisition of *Clostridium difficile* by hospitalized patients: evidence for colonized new admissions as a source of infection. *Journal of infectious diseases, 166*(3), 561-567.

Clements, A. C. A., SOARES MAGALHAES, R. J., TATEM, A., Paterson, D. L., & Riley, T. V. (2010). *Clostridium difficile* PCR ribotype 027: assessing the risks of further worldwide spread. *Lancet. Infectious diseases, 10*(6), 395-404.

Codella, J., Safdar, N., Heffernan, R., & Alagoz, O. (2014). An Agent-based Simulation Model for *Clostridium difficile* Infection Control. *Medical Decision Making*, 0272989X14545788.

Cohen, B., Hyman, S., Rosenberg, L., & Larson, E. (2012). Frequency of patient contact with health care personnel and visitors: implications for infection prevention. *Joint Commission journal on quality and patient safety/Joint Commission Resources, 38*(12), 560.

Cooper, P. B., Heuer, A. J., & Warren, C. A. (2013). Electronic screening of patients for predisposition to *Clostridium difficile* infection in a community hospital. *American Journal of Infection Control, 41*(3), 232-235.

Curtis, D., Kanade, G., Pemmaraju, S., Polgreen, P., & Segre, A. (2009). *Analysis of hospital health-care worker contact networks.* Paper presented at the 5th UK Social Networks Conference.

Cusumano-Towner, M., Li, D. Y., Tuo, S., Krishnan, G., & Maslove, D. M. (2013). A social network of hospital acquired infection built from electronic medical record data. *Journal of the American Medical Informatics Association*, amiajnl-2012-001401.

Cvach, M. (2012). Monitor alarm fatigue: an integrative review. *Biomedical Instrumentation & Technology, 46*(4), 268-277.

Dancer, S. (2008). Considering the introduction of universal MRSA screening. *Journal of Hospital Infection, 69*(4), 315-320.

Daneman, N., Simor, A. E., & Redelmeier, D. A. (2009). Validation of a modified version of the national nosocomial infections surveillance system risk index for health services research. *Infection control and hospital epidemiology, 30*(6), 563-569.

Das, G., Lin, K.-I., Mannila, H., Renganathan, G., & Smyth, P. (1998). *Rule Discovery from Time Series.* Paper presented at the KDD.

Das, K. (2009). *Detecting patterns of anomalies*: ProQuest.

de Blank, P., Zaoutis, T., Fisher, B., Troxel, A., Kim, J., & Aplenc, R. (2013). Trends in *Clostridium difficile* Infection and Risk Factors for Hospital Acquisition of *Clostridium difficile* among Children with Cancer. *The Journal of Pediatrics*.

de Bruin, J. S., Seeling, W., & Schuh, C. (2014). Data use and effectiveness in electronic surveillance of healthcare associated infections in the 21st century: a systematic review. *Journal of the American Medical Informatics Association*, amiajnl-2013-002089.

de Oliveira, A. C., Ciosak, S. I., Ferraz, E. M., & Grinbaum, R. S. (2006). Surgical site infection in patients submitted to digestive surgery: risk prediction and the NNIS risk index. *American journal of infection control, 34*(4), 201-207.

Didelot, X., Eyre, D. W., Cule, M., Ip, C., Ansari, M. A., Griffiths, D., Batty, E. M. (2012). Microevolutionary analysis of *Clostridium difficile* genomes to investigate transmission. *Genome Biol, 13*(12), R118.

Dietterich, T. G. (2000). Ensemble methods in machine learning *Multiple classifier systems* (pp. 1-15): Springer.

Donskey, C. J. (2010). Preventing transmission of *Clostridium difficile*: is the answer blowing in the wind? *Clinical Infectious Diseases, 50*(11), 1458-1461.

Dubberke, E. R., Reske, K. A., Noble-Wang, J., Thompson, A., Killgore, G., Mayfield, J., McDonald, L. C. (2007). Prevalence of *Clostridium difficile* environmental contamination and strain variability in multiple health care facilities. *American Journal of Infection Control, 35*(5), 315-318.

Dubberke, E. R., Reske, K. A., Yan, Y., Olsen, M. A., McDonald, L. C., & Fraser, V. J. (2007). *Clostridium difficile*—associated disease in a setting of endemicity: identification of novel risk factors. *Clinical Infectious Diseases, 45*(12), 1543-1549.

Dubberke, E. R., Yan, Y., Reske, K. A., Butler, A. M., Doherty, J., Pham, V., & Fraser, V. J. (2011). Development and validation of a *Clostridium difficile* infection risk prediction model. *Infection control and hospital epidemiology: the official journal of the Society of Hospital Epidemiologists of America, 32*(4), 360-366.

Eckstein, B. C., Adams, D. A., Eckstein, E. C., Rao, A., Sethi, A. K., Yadavalli, G. K., & Donskey, C. J. (2007). Reduction of *Clostridium difficile* and vancomycin-resistant Enterococcus contamination of environmental surfaces after an intervention to improve cleaning methods. *BMC infectious diseases, 7*(1), 61.

Eggertson, L. (2005). *C. difficile* may have killed 2000 in Quebec: study. *Canadian Medical Association Journal, 173*(9), 1020-1021.

Ehrentraut, C., Tanushi, H., Dalianis, H., & Tiedemann, J. (2012). *Detection of Hospital Acquired Infections in sparse and noisy Swedish patient records.* Paper presented at the Proceedings of the Sixth Workshop on Analytics for Noisy Unstructured Text Data (AND 2012).

Elias, J., Heuschmann, P. U., Schmitt, C., Eckhardt, F., Boehm, H., Maier, S., . . . Weisser, C. (2013). Prevalence dependent calibration of a predictive model for nasal carriage of methicillin-resistant Staphylococcus aureus. *BMC infectious diseases, 13*(1), 1-11.

Erasmus, V., Daha, T. J., Brug, H., Richardus, J. H., Behrendt, M. D., Vos, M. C., & van Beeck, E. F. (2010). Systematic review of studies on compliance with hand hygiene guidelines in hospital care. *Infection Control and Hospital Epidemiology, 31*(3), 283-294.

Escolano, S., Golmard, J. L., Korinek, A. M., & Mallet, A. (2000). A multi-state model for evolution of intensive care unit patients: prediction of nosocomial infections and deaths. *Statistics in medicine, 19*(24), 3465-3482.

Eyre, D. W., Cule, M. L., Wilson, D. J., Griffiths, D., Vaughan, A., O'Connor, L., Finney, J. M. (2013). Diverse sources of *C. difficile* infection identified on whole-genome sequencing. *New England Journal of Medicine, 369*(13), 1195-1205.

Eyre, D. W., Griffiths, D., Vaughan, A., Golubchik, T., Acharya, M., O'Connor, L., Peto, T. E. (2013). Asymptomatic *Clostridium difficile* Colonisation and Onward Transmission. *PloS one, 8*(11), e78445.

Fang, W. F., Yang, K. Y., Wu, C. L., Yu, C. J., Chen, C. W., Tu, C.-Y., & Lin, M. C. (2011). Application and comparison of scoring indices to predict outcomes in patients with healthcare-associated pneumonia. *Crit Care, 15*(1), R32.

Fernando, J. I., & Dawson, L. L. (2009). The health information system security threat lifecycle: An informatics theory. *International Journal of Medical Informatics, 78*(12), 815-826.

Filetoth, Z. (2008). *Hospital-Acquired Infections*: Wiley.

Freeman, R., Moore, L., García Álvarez, L., Charlett, A., & Holmes, A. (2013). Advances in electronic surveillance for healthcare-associated infections in the 21st Century: a systematic review. *Journal of Hospital Infection, 84*(2), 106-119.

Frost, S. D., Pybus, O. G., Gog, J. R., Viboud, C., Bonhoeffer, S., & Bedford, T. (2014). Eight challenges in phylodynamic inference. *Epidemics*.

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

Garner, J. S., Jarvis, W. R., Emori, T. G., Horan, T. C., & Hughes, J. M. (1988). CDC definitions for nosocomial infections, 1988. *American journal of infection control, 16*(3), 128-140.

Glaz, J., Pozdnyakov, V., & Wallenstein, S. (2009). *Scan statistics: methods and applications*: Springer.

Goulenok, T., Ferroni, A., Bille, E., Lécuyer, H., Join-Lambert, O., Descamps, P., . . . Zahar, J. (2013). Risk factors for developing ESBL *E. coli*: can clinicians predict infection in patients with prior colonization? *Journal of Hospital Infection, 84*(4), 294-299.

Gravel, D., Miller, M., Simor, A., Taylor, G., Gardam, M., McGeer, A., . . . Boyd, D. (2009). Health care-associated *Clostridium difficile* infection in adults admitted to acute care hospitals in Canada: a Canadian Nosocomial Infection Surveillance Program Study. *Clinical Infectious Diseases, 48*(5), 568-576.

Gravel, D., Taylor, G., Ofner, M., Johnston, L., Loeb, M., Roth, V., . . . Matlow, A. (2007). Point prevalence survey for healthcare-associated infections within Canadian adult acute-care hospitals. *Journal of Hospital Infection, 66*(3), 243-248.

Graves, N., Barnett, A. G., & Rosenthal, V. D. (2011). Open versus closed IV infusion systems: a state based model to predict risk of catheter associated blood stream infections. *BMJ open, 1*(2).

Grimm, V., & Railsback, S. F. (2005). *Individual-based modeling and ecology*: Princeton university press.

Haley, R. W., Culver, D. H., White, J. W., Morgan, W. M., Emori, T. G., Munn, v. P., & Hooton, T. M. (1985). The efficacy of infection surveillance and control programs in preventing nosocomial infections in US hospitals. *American journal of epidemiology, 121*(2), 182-205.

Halpin, H., Shortell, S. M., Milstein, A., & Vanneman, M. (2011). Hospital adoption of automated surveillance technology and the implementation of infection prevention and control programs. *American journal of infection control, 39*(4), 270-276.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*(1), 29-36.

Harbarth, S., & Samore, M. H. (2012). *Clostridium difficile* : transmission? *Plos Medicine, 9*(2), e1001171.

Harbarth, S., Sax, H., Uckay, I., Fankhauser, C., Agostinho, A., Christenson, J. T., . . . Pittet, D. (2008). A Predictive Model for Identifying Surgical Patients at Risk of Methicillin-Resistant *Staphylococcus aureus* Carriage on Admission. *Journal of the American College of Surgeons, 207*(5), 683-689.

Harrison, W. A., Griffith, C. J., Ayers, T., & Michaels, B. (2003). Bacterial transfer and cross-contamination potential associated with paper-towel dispensing. *American Journal of Infection Control, 31*(7), 387-391.

Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Friedman, J., & Tibshirani, R. (2009). *The elements of statistical learning* (Vol. 2): Springer.

Hautemanière, A., Florentin, A., Hunter, P. R., Bresler, L., & Hartemann, P. (2013). Screening for surgical nosocomial infections by crossing databases. *Journal of infection and public health, 6*(2), 89-97.

He, J., Tong, H., & Carbonell, J. (2012). An effective framework for characterizing rare categories. *Frontiers of Computer Science, 6*(2), 154-165. doi: 10.1007/s11704-012-2861-9

Helton, J. C., & Davis, F. J. (2003). Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems. *Reliability Engineering & System Safety, 81*(1), 23-69.

Hendrich, A. L., & Lee, N. (2004). Intra-unit patient transports: time, motion, and cost impact on hospital efficiency. *Nursing economics, 23*(4), 157-164, 147.

Henrich, T. J., Krakower, D., Bitton, A., & Yokoe, D. S. (2009). Clinical risk factors for severe *Clostridium difficile*–associated disease. *Emerging infectious diseases, 15*(3), 415.

Herman, T., Pemmaraju, S. V., Segre, A. M., Polgreen, P. M., Curtis, D. E., Fries, J., . . . Severson, M. (2009). *Wireless applications for hospital epidemiology.* Paper presented at the Proceedings of the 1st ACM international workshop on Medical-grade wireless networks.

Hethcote, H. W. (2000). The mathematics of infectious diseases. *SIAM review, 42*(4), 599-653.

Hirsch, E. B., Cottreau, J. M., Chang, K. T., Caeiro, J. P., Johnson, M. L., & Tam, V. H. (2012). A model to predict mortality following *Pseudomonas aeruginosa* bacteremia. *Diagnostic microbiology and infectious disease, 72*(1), 97-102.

Honda, H., & Dubberke, E. R. (2014). The changing epidemiology of *Clostridium difficile* infection. *Current opinion in gastroenterology, 30*(1), 54-62.

Hookman, P., & Barkin, J. S. (2009). *Clostridium difficile* associated infection, diarrhea and colitis. *World Journal of Gastroenterology, 15*(13), 1554-1580. doi: 10.3748/wjg.15.1554

Horan, T. C., Gaynes, R. P., Martone, W. J., Jarvis, W. R., & Grace Emori, T. (1992). CDC definitions of nosocomial surgical site infections, 1992: a modification of CDC definitions of surgical wound infections. *American journal of infection control, 20*(5), 271-274.

Hornbuckle, K., Chak, A., Lazarus, H., Cooper, G., Kutteh, L., Gucalp, R., . . . Salata, R. (1998). Determination and validation of a predictive model for Clostridium diĄcile diarrhea in hospitalized oncology patients. *Annals of oncology, 9*, 307-311.

Horvitz, E. (2010). From Data to Predictions and Decisions: Enabling Evidence-Based Healthcare. *Microsoft research*.

Hsu, C. C., Lin, Y. E., Chen, Y. S., Liu, Y. C., & Muder, R. R. (2008). Validation study of artificial neural network models for prediction of methicillin-resistant Staphylococcus aureus carriage. *Infection control and hospital epidemiology, 29*(7), 607-614. https://secure.cihi.ca/estore/productFamily.htm?pf=PFC1046&lang=en&media=0.

Huang, S. S., Yokoe, D. S., Stelling, J., Placzek, H., Kulldorff, M., Kleinman, K., . . . Platt, R. (2010). Automated Detection of Infectious Disease Outbreaks in Hospitals: A Retrospective Cohort Study. *Plos Medicine, 7*(2). doi: 10.1371/journal.pmed.1000238

Hurr, H., Hawley, H. B., Czachor, J. S., Markert, R. J., & McCarthy, M. C. (1999). APACHE II and ISS scores as predictors of nosocomial infections in trauma patients. *American journal of infection control, 27*(2), 79-83.

Hutwagner, L., Browne, T., Seeman, G. M., & Fleischauer, A. T. (2005). Comparing aberration detection methods with simulated data. *Emerging infectious diseases, 11*(2), 314.

Information, C. I. f. H. (2005). Inpatient Hospitalizations and Average Length of Stay Trends in Canada, 2003-2004 and 2004-2005 Retrieved from

Jackson, M. O. (2010). *Social and economic networks*: Princeton University Press.

Jackson, M. O., & Watts, A. (2002). The evolution of social and economic networks. *Journal of Economic Theory, 106*(2), 265-295.

Jiménez, J. M., Lewis, B., & Eubank, S. (2013). Hospitals as Complex Social Systems: Agent-Based Simulations of Hospital-Acquired Infections *Complex Sciences* (pp. 165-178): Springer.

Johnson, S., Clabots, C., Linn, F., Olson, M., Peterson, L., & Gerding, D. (1990). Nosocomial *Clostridium difficile* colonisation and disease. *The Lancet, 336*(8707), 97-100.

Join, C. Health care associated infections: a backgrounder.

Joshi, N., Localio, A. R., & Hamory, B. H. (1992). A predictive risk index for nosocomial pneumonia in the intensive care unit. *The American journal of medicine, 93*(2), 135-142.

Kachrimanidou, M., & Malisiovas, N. (2011). *Clostridium difficile* infection: a comprehensive review. *Critical reviews in microbiology, 37*(3), 178-187.

Kahn, M., Steib, S., Fraser, V., & Dunagan, W. (1993). *An expert system for culture-based infection control surveillance.* Paper presented at the Proceedings of the Annual Symposium on Computer Application in Medical Care.

Kaier, K., Luft, D., Dettenkofer, M., Kist, M., & Frank, U. (2011). Correlations between bed occupancy rates and *Clostridium difficile* infections: a time-series analysis. *Epidemiology and infection, 139*(03), 482-485.

Keeling, M. J., & Eames, K. T. (2005). Networks and epidemic models. *Journal of the Royal Society Interface, 2*(4), 295-307.

Keeling, M. J., & Rohani, P. (2008). *Modeling infectious diseases in humans and animals*: Princeton University Press.

Keller, J. P., Diefes, R., Graham, K., Meyers, M., & Pelczarski, K. (2011). Why clinical alarms are a'top ten'hazard: how you can help reduce the risk. *Biomedical Instrumentation & Technology, 45*(s1), 17-23.

Khanafer, N., Touré, A., Chambrier, C., Cour, M., Reverdy, M. E., Argaud, L., & Vanhems, P. (2013). Predictors of Clostridium difficile infection severity in patients hospitalised in medical intensive care. *World journal of gastroenterology: WJG, 19*(44), 8034.

Kim, H. T. (2007). Cumulative incidence in competing risks data and competing risks regression analysis. *Clinical Cancer Research, 13*(2), 559-565.

Kinlin, L. M., Kirchner, C., Zhang, H., Daley, J., & Fisman, D. N. (2010). Derivation and validation of a clinical prediction rule for nosocomial pneumonia after coronary artery bypass graft surgery. *Clinical infectious diseases, 50*(4), 493-501.

Klevens, R. M., Edwards, J. R., Richards, C. L., Horan, T. C., Gaynes, R. P., Pollock, D. A., & Cardo, D. M. (2007). Estimating health care-associated infections and deaths in US hospitals, 2002. *Public health reports, 122*(2), 160.

Kohavi, R. (1995). *A study of cross-validation and bootstrap for accuracy estimation and model selection.* Paper presented at the IJCAI.

Kolaczyk, E. D. C., Gábor. (2014). *Statistical analysis of network data with R*: Springer.

Krall, S., O'Connor, R., & Maercks, L. (2009). Higher inpatient medical surgical bed occupancy extends admitted patients' stay. *The western journal of emergency medicine, 10*(2), 93.

Kristjánsson, M., Samore, M. H., Gerding, D. N., DeGirolami, P. C., Bettin, K. M., Karchmer, A. W., & Arbeit, R. D. (1994). Comparison of restriction endonuclease analysis, ribotyping, and pulsed-field gel electrophoresis for molecular differentiation of *Clostridium difficile* strains. *Journal of clinical microbiology, 32*(8), 1963-1969.

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software, 28*(5), 1-26.

Kuhn, M. (2012). Variable importance using the caret package.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling* Springer.

Kutty, P. K., Woods, C. W., Sena, A. C., Benoit, S. R., Naggie, S., Frederick, J., . . . McDonald, L. C. (2010). Risk Factors for and Estimated Incidence of Community-associated *Clostridium difficile* Infection, North Carolina, USA1. *Emerging infectious diseases, 16*(2), 198.

Lanzas, C., & Dubberke, E. R. (2014). Effectiveness of Screening Hospital Admissions to Detect Asymptomatic Carriers of *Clostridium difficile*: A Modeling Evaluation. *Infection Control, 35*(08), 1043-1050.

Lanzas, C., Dubberke, E., Lu, Z., Reske, K., & Grohn, Y. (2011). Epidemiological model for *Clostridium difficile* transmission in healthcare settings. *Infection Control and Hospital Epidemiology, 32*(6), 553-561.

Leal, J., & Laupland, K. (2008). Validity of electronic surveillance systems: a systematic review. *Journal of Hospital Infection, 69*(3), 220-229.

Lee, J. T., Hertz, M. I., Dunitz, J. M., Kelly, R. F., D'Cunha, J., Whitson, B. A., & Shumway, S. J. (2013). The rise of Clostridium difficile infection in lung transplant recipients in the modern era. *Clinical transplantation, 27*(2), 303-310.

Leekha, S., Aronhalt, K. C., Sloan, L. M., Patel, R., & Orenstein, R. (2013). Asymptomatic *Clostridium difficile* colonization in a tertiary care hospital: admission prevalence and risk factors. *American Journal of Infection Control, 41*(5), 390-393.

Lessa, F. C., Gould, C. V., & McDonald, L. C. (2012). Current status of *Clostridium difficile* infection epidemiology. *Clinical Infectious Diseases, 55*(suppl 2), S65-S70.

Leth, R., & Møller, J. (2006). Surveillance of hospital-acquired infections based on electronic hospital registries. *Journal of Hospital Infection, 62*(1), 71-79.

Levy, S. B., & Marshall, B. (2004). Antibacterial resistance worldwide: causes, challenges and responses. *Nature medicine, 10*, S122-S129.

Lewis, R. J. (2000). *An introduction to classification and regression tree (CART) analysis.* Paper presented at the Annual Meeting of the Society for Academic Emergency Medicine in San Francisco, California.

Linsky, A., Gupta, K., Lawler, E. V., Fonda, J. R., & Hermos, J. A. (2010). Proton pump inhibitors and risk for recurrent *Clostridium difficile* infection. *Archives of internal medicine, 170*(9), 772.

Lo, Y. S., Lee, W. S., & Liu, C. T. (2013). Utilization of Electronic Medical Records to Build a Detection Model for Surveillance of Healthcare-Associated Urinary Tract Infections. *Journal of medical systems, 37*(2), 1-6.

Lodise, T. P., McKinnon, P. S., & Rybak, M. (2003). Prediction model to identify patients with Staphylococcus aureus bacteremia at risk for methicillin resistance. *Infection control and hospital epidemiology, 24*(9), 655-661.

Longadge, R., & Dongre, S. (2013). Class Imbalance Problem in Data Mining Review. *arXiv preprint arXiv:1305.1707*.

Loo, V. G., Bourgault, A.-M., Poirier, L., Lamothe, F., Michaud, S., Turgeon, N., . . . Gilca, R. (2011). Host and pathogen factors for *Clostridium difficile* infection and colonization. *New England Journal of Medicine, 365*(18), 1693-1703.

Lopes, J. M., Goulart, E., Siqueira, A. L., Fonseca, I. K., de Brito, M. V., & Starling, C. E. (2009). Nosocomial infections in brazilian pediatric patients: using a decision tree to identify high mortality groups. *Brazilian Journal of Infectious Diseases, 13*(2), 111-117.

Louie, T. J., & Meddings, J. (2004). *Clostridium difficile* infection in hospitals: risk factors and responses. *Canadian Medical Association Journal, 171*(1), 45-46.

Marino, S., Hogue, I. B., Ray, C. J., & Kirschner, D. E. (2008). A methodology for performing global uncertainty and sensitivity analysis in systems biology. *Journal of theoretical biology, 254*(1), 178-196.

Mayhall, C. G. (2012). *Hospital epidemiology and infection control*: LWW.

McArdle, F., Lee, R., Gibb, A., & Walsh, T. (2006). How much time is needed for hand hygiene in intensive care? A prospective trained observer study of rates of contact between healthcare workers and intensive care patients. *Journal of Hospital Infection, 62*(3), 304-310.

McCollum, D. L., & Rodriguez, J. M. (2012). Detection, Treatment, and Prevention of *Clostridium difficile* Infection. *Clinical Gastroenterology and Hepatology, 10*(6), 581-592. doi: 10.1016/j.cgh.2012.03.008

McFarland, L. V., Clarridge, J. E., Beneda, H. W., & Raugi, G. J. (2007). Fluoroquinolone use and risk factors for *Clostridium difficile*-associated disease within a Veterans Administration health care system. *Clinical Infectious Diseases, 45*(9), 1141-1151.

McFarland, L. V., Surawicz, C. M., & Stamm, W. E. (1990). Risk factors for *Clostridium difficile* carriage and *C. difficile*-associated diarrhea in a cohort of hospitalized patients. *Journal of infectious diseases, 162*(3), 678-684.

McGlone, S., Bailey, R., Zimmer, S., Popovich, M., Tian, Y., Ufberg, P., . . . Lee, B. (2012). The economic burden of *Clostridium difficile*. *Clinical Microbiology and Infection, 18*(3), 282-289.

Meng, Y. (2009). *Application of Agent-based Simulation to the Modelling and Management of Hospital-acquired Infections.* University of Warwick.

Meng, Y., Davies, R., Hardy, K., & Hawkey, P. (2010). An application of agent-based simulation to the management of hospital-acquired infection. *Journal of Simulation, 4*(1), 60-67.

Merrill, R. M. (2012). *Introduction to epidemiology*: Jones & Bartlett Publishers.

Meyers, L. (2007). Contact network epidemiology: Bond percolation applied to infectious disease prediction and control. *Bulletin of the American Mathematical Society, 44*(1), 63-86.

Meyers, L. A., Newman, M., Martin, M., & Schrag, S. (2003). Applying network theory to epidemics: control measures for Mycoplasma pneumoniae outbreaks. *Emerging infectious diseases, 9*(2), 204.

Meystre, S. M., Savova, G. K., Kipper-Schuler, K. C., & Hurdle, J. F. (2008). Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform, 35*, 128-144.

Miller, M. A., Hyland, M., Ofner-Agostini, M., Gourdeau, M., & Ishak, M. (2002). Morbidity, mortality, and healthcare burden of nosocomial *Clostridium difficile*-associated diarrhea in Canadian hospitals. *Infection Control and Hospital Epidemiology, 23*(3), 137-140.

Mirsaeidi, M., Peyrani, P., & Ramirez, J. A. (2009). Predicting mortality in patients with ventilator-associated pneumonia: The APACHE II score versus the new IBMP-10 score. *Clinical infectious diseases, 49*(1), 72-77.

Missaghi, B., Valenti, A. J., & Owens Jr, R. C. (2008). *Clostridium difficile* infection: A critical overview. *Current Infectious Disease Reports, 10*(3), 165-173. doi: 10.1007/s11908-008-0028-5

Morales, C. H., Escobar, R. M., Villegas, M. I., Castaño, A., & Trujillo, J. (2011). Surgical site infection in abdominal trauma patients: risk prediction and performance of the NNIS and SENIC indexes. *Canadian Journal of Surgery, 54*(1), 17.

Moss, F., McSwiggan, D., McNicol, M., & Miller, D. (1981). Survey of antibiotic prescribing in a district general hospital I. Pattern of use. *The Lancet, 318*(8242), 349-352.

Mutters, R., Nonnenmacher, C., Susin, C., Albrecht, U., Kropatsch, R., & Schumacher, S. (2009). Quantitative detection of *Clostridium difficile* in hospital environmental samples by real-time polymerase chain reaction. *Journal of Hospital Infection, 71*(1), 43-48.

Najmi, A.-H., & Magruder, S. F. (2005). An adaptive prediction and detection algorithm for multistream syndromic surveillance. *BMC medical informatics and decision making, 5*(1), 33.

Nelson, B. L., Carson, J. S., & Banks, J. (2001). *Discrete event system simulation*: Prentice hall.

Newman, M. E. (2002). Spread of epidemic disease on networks. *Physical review E, 66*(1), 016128.

Ohst, J., Liljeros, F., Stenhem, M., & Holme, P. (2014). The network positions of methicillin resistant Staphylococcus aureus affected units in a regional healthcare system. *EPJ Data Science, 3*(1), 1-15.

Olsen, M. A., Higham-Kessler, J., Yokoe, D. S., Butler, A. M., Vostok, J., Stevenson, K. B., . . . Program, C. P. E. (2009). Developing a risk stratification model for surgical site infection after abdominal hysterectomy. *Infection control and hospital epidemiology: the official journal of the Society of Hospital Epidemiologists of America, 30*(11), 1077.

Owens, R. C., Donskey, C. J., Gaynes, R. P., Loo, V. G., & Muto, C. A. (2008). Antimicrobial-associated risk factors for *Clostridium difficile* infection. *Clinical Infectious Diseases, 46*(Supplement 1), S19-S31.

Park, S., Kim, E., Kang, Y., Park, M., Kim, Y., Kim, S., . . . Jung, J. (2013). Validation of a scoring tool to predict drug-resistant pathogens in hospitalised pneumonia patients. *The International Journal of Tuberculosis and Lung Disease, 17*(5), 704-709.

Pearl, A., & Bar-Or, D. (2012). Decision Support in Trauma Management: Predicting Potential Cases of Ventilator Associated Pneumonia. *Quality of Life Through Quality of Information: Proceedings of MIE2012, 180*, 305.

Peled, N., Pitlik, S., Samra, Z., Kazakov, A., Bloch, Y., & Bishara, J. (2007). Predicting Clostridium difficile toxin in hospitalized patients with antibiotic-associated diarrhea. *Infection control and hospital epidemiology, 28*(4), 377-381.

Pirracchio, R., Petersen, M. L., Carone, M., Rigon, M. R., Chevret, S., & van der Laan, M. J. (2015). Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): a population-based study. *The Lancet Respiratory Medicine, 3*(1), 42-52.

Pittet, D., Allegranzi, B., Sax, H., Dharan, S., Pessoa-Silva, C. L., Donaldson, L., & Boyce, J. M. (2006). Evidence-based model for hand transmission during patient care and the role of improved practices. *The Lancet infectious diseases, 6*(10), 641-652.

Polley, E. C., & van der Laan, M. J. (2010). Super learner in prediction.

Poutanen, S. M., & Simor, A. E. (2004). *Clostridium difficile*-associated diarrhea in adults. *Canadian Medical Association Journal, 171*(1), 51-58.

Program, T. C. N. I. S. (2013). Healthcare-Associated-*Clostridium difficile* Infection (HA-CDI) 2007 - 2011, 2014

Proux, D., Hagège, C., Gicquel, Q., Pereira, S., Darmoni, S., Segond, F., & Metzger, M. H. (2011). *Architecture and Systems for Monitoring Hospital Acquired Infec-tions inside a Hospital Information Workflow*. Paper presented at the Proceedings of the Workshop on Biomedical Natural Language Processing. USA: Portland, Oregon.

Rada, R. (2007). *Information systems and healthcare enterprises*: IGI Global.

Railsback, S. F., & Grimm, V. (2011). *Agent-based and individual-based modeling: a practical introduction*: Princeton University Press.

RAY, A., & RIVERA, K. R. (2014). Overview of the Management Of *Clostridium difficile* Infections.

Riddle, D. J., & Dubberke, E. R. (2009). *Clostridium difficile* Infection in the Intensive Care Unit. *Infectious Disease Clinics of North America, 23*(3), 727-743. doi: 10.1016/j.idc.2009.04.011

Riggs, M. M., Sethi, A. K., Zabarsky, T. F., Eckstein, E. C., Jump, R. L., & Donskey, C. J. (2007). 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), 992-998.

Rubin, M. A., Jones, M., Leecaster, M., Khader, K., Ray, W., Huttner, A., . . . Borotkanics, R. J. (2013). A Simulation-Based Assessment of Strategies to Control *Clostridium Difficile* Transmission and Infection. *PloS one, 8*(11), e80671.

Ryan, K. J., & Ray, C. G. (2010). *Sherris medical microbiology*: McGraw Hill Medical New York.

Samore, M. H., DeGirolami, P. C., Tlucko, A., Lichtenberg, D. A., Melvin, Z. A., & Karchmer, A. W. (1994). *Clostridium difficile* colonization and diarrhea at a tertiary care hospital. *Clinical Infectious Diseases, 18*(2), 181-187.

Samore, M. H., Venkataraman, L., DeGirolami, P. C., Arbeit, R. D., & Karchmer, A. W. (1996). Clinical and molecular epidemiology of sporadic and clustered cases of nosocomial *Clostridium difficile* diarrhea. *The American journal of medicine, 100*(1), 32-40.

Sanagou, M., Wolfe, R., Leder, K., & Reid, C. (2013). External validation and updating of a prediction model for nosocomial pneumonia after coronary artery bypass graft surgery. *Epidemiology and infection*, 1-5.

Scheckler, W. E., & Bennett, J. V. (1970). Antibiotic usage in seven community hospitals. *Jama, 213*(2), 264-267.

Sébille, V., & Valleron, A.-J. (1997). A computer simulation model for the spread of nosocomial infections caused by multidrug-resistant pathogens. *Computers and biomedical research, 30*(4), 307-322.

Sendelbach, S. (2012). Alarm fatigue. *Nursing Clinics of North America, 47*(3), 375-382.

Sherman, E. R., Heydon, K. H., John, K. H. S., Eva Teszner, B., Rettig, S. L., Alexander, S. K., . . . Coffin, S. E. (2006). Administrative data fail to accurately identify cases of healthcare-associated infection. *Infection control and hospital epidemiology, 27*(4), 332-337.

Sorensen, H. T., Sabroe, S., & Olsen, J. (1996). A framework for evaluation of secondary data sources for epidemiological research. *International Journal of Epidemiology, 25*(2), 435-442.

Starr, J., Campbell, A., Renshaw, E., Poxton, I., & Gibson, G. (2009). Spatio-temporal stochastic modelling of *Clostridium difficile*. *Journal of Hospital Infection, 71*(1), 49-56.

Steele, S., Bilchik, A., Eberhardt, J., Kalina, P., Nissan, A., Johnson, E., . . . Stojadinovic, A. (2012). Using Machine-Learned Bayesian Belief Networks to Predict Perioperative Risk of *Clostridium Difficile* Infection Following Colon Surgery. *interactive Journal of Medical Research (i-JMR), 1*(2), e6.

Steinmann, J., Knaust, A., Moussa, A., Joch, J., Ahrens, A., Walmrath, H. D., . . . Herr, C. E. W. (2008). Implementation of a novel on-ward computer-assisted surveillance system for device-associated infections in an intensive care unit. *International journal of hygiene and environmental health, 211*(1), 192-199.

Stevens, V., Concannon, C., van Wijngaarden, E., & McGregor, J. (2013). Validation of the chronic disease score-infectious disease (CDS-ID) for the prediction of hospital-associated *clostridium difficile* infection (CDI) within a retrospective cohort. *BMC infectious diseases, 13*(1), 1-8.

Stewardson, A., Fankhauser, C., De Angelis, G., Rohner, P., Safran, E., Schrenzel, J., . . . Harbarth, S. (2013). Burden of Bloodstream Infection Caused by Extended-Spectrum β-Lactamase–Producing Enterobacteriaceae Determined Using Multistate Modeling at a Swiss University Hospital and a Nationwide Predictive Model. *Infection control and hospital epidemiology, 34*(2), 133-143.

Stonedahl, F. J., & Adviser-Wilensky, U. J. (2011). Genetic algorithms for the exploration of parameter spaces in agent-based models.

Su, C.P., Chen, T. H. H., Chen, S. Y., Ghiang, W. C., Wu, G. H. M., Sun, H. Y., . . . Chen, Y. C. (2011). Predictive model for bacteremia in adult patients with blood cultures performed at the emergency department: a preliminary report. *Journal of Microbiology, Immunology and Infection, 44*(6), 449-455.

Tacconelli, E., Cataldo, M. A., De Pascale, G., Manno, D., Spanu, T., Cambieri, A., . . . Cauda, R. (2008). Prediction models to identify hospitalized patients at risk of being colonized or infected with multidrug-resistant Acinetobacter baumannii calcoaceticus complex. *Journal of antimicrobial chemotherapy, 62*(5), 1130-1137.

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

Thibault, A., Miller, M. A., & Gaese, C. (1991). Risk factors for the development of *Clostridium difficile*-associated diarrhea during a hospital outbreak. *Infection Control and Hospital Epidemiology*, 345-348.

Thomas, M., & Viner-Brown, S. (2010). Public Reporting of Hospital-Acquired Infections. *Medicine and Health Rhode Island, 93*(9), 283.

Trick, W. E., Zagorski, B. M., Tokars, J. I., Vernon, M. O., Welbel, S. F., Wisniewski, M. F., . . . Weinstein, R. A. (2004). Computer algorithms to detect bloodstream infections. *Emerging infectious diseases, 10*(9), 1612.

Ueno, T., & Masuda, N. (2008). Controlling nosocomial infection based on structure of hospital social networks. *Journal of theoretical biology, 254*(3), 655-666.

van Kleef, E., Robotham, J. V., Jit, M., Deeny, S. R., & Edmunds, W. J. (2013). Modelling the transmission of healthcare associated infections: a systematic review. *BMC infectious diseases, 13*(1), 294.

van Mourik, M. S., Moons, K. G., van Solinge, W. W., Berkelbach-van Der Sprenkel, J.-W., Regli, L., Troelstra, A., & Bonten, M. J. (2012). Automated detection of healthcare associated infections: external validation and updating of a model for surveillance of drain-related meningitis. *PloS one, 7*(12), e51509.

van Mourik, M. S., Troelstra, A., van Solinge, W. W., Moons, K. G., & Bonten, M. J. (2013). Automated surveillance for healthcare-associated infections: opportunities for improvement. *Clinical infectious diseases, 57*(1), 85-93.

van Walraven, C., & Musselman, R. (2013). The Surgical Site Infection Risk Score (SSIRS): A Model to Predict the Risk of Surgical Site Infections. *PloS one, 8*(6), e67167.

Vargha, A., & Delaney, H. D. (2000). A critique and improvement of the CL common language effect size statistics of McGraw and Wong. *Journal of Educational and Behavioral Statistics, 25*(2), 101-132.

Viana, F. A. (2013). *Things you wanted to know about the Latin hypercube design and were afraid to ask.* Paper presented at the 10th World Congress on Structural and Multidisciplinary Optimization, Orlando, Florida, USA (cf. p. 69).

Visscher, S., Kruisheer, E. M., Schurink, C. A., Lucas, P. J., & Bonten, M. J. (2008). Predicting pathogens causing ventilator-associated pneumonia using a Bayesian network model. *Journal of antimicrobial chemotherapy, 62*(1), 184-188.

Vynnycky, E., & White, R. (2010). *An introduction to infectious disease modelling*: Oxford University Press.

Walker, A. S., Eyre, D. W., Wyllie, D. H., Dingle, K. E., Harding, R. M., O'Connor, L., . . . Wilcox, M. H. (2012). Characterisation of *Clostridium difficile* Hospital Ward–Based Transmission Using Extensive Epidemiological Data and Molecular Typing. *Plos Medicine, 9*(2), e1001172.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications* (Vol. 8): Cambridge university press.

Weber, D. J., Rutala, W. A., Miller, M. B., Huslage, K., & Sickbert-Bennett, E. (2010). Role of hospital surfaces in the transmission of emerging health care-associated pathogens: Norovirus, *Clostridium difficile*, and *Acinetobacter* species. *American Journal of Infection Control, 38*(5), S25-S33.

WHO. (2010). *ICD-10 Version:2010*.

Wiens, J., Guttag, J., & Horvitz, E. (2012). *Patient Risk Stratification for Hospital-Associated C. diff as a Time-Series Classification Task.* Paper presented at the Advances in Neural Information Processing Systems 25.

Wiens, J., Horvitz, E., & Guttag, J. (2012). *Learning evolving patient risk processes for c. diff colonization.* Paper presented at the ICML Workshop on Machine Learning from Clinical Data.

Wisnivesky, J. P., Henschke, C., Balentine, J., Willner, C., Deloire, A. M., & McGinn, T. G. (2005). Prospective validation of a prediction model for isolating inpatients with suspected pulmonary tuberculosis. *Archives of internal medicine, 165*(4), 453-457.

Woeltje, K. (2013). Moving into the future: electronic surveillance for healthcare-associated infections. *Journal of Hospital Infection, 84*(2), 103-105.

Wong, W. K., Moore, A., Cooper, G., & Wagner, M. (2003). WSARE: What's strange about recent events? *Journal of Urban Health-Bulletin of the New York Academy of Medicine, 80*(2), I66-I75.

Yadav, Y., Garey, K., Dao-Tran, T., Kaila, V., Gbito, K. E., & DuPont, H. (2009). Automated system to identify *Clostridium difficile* infection among hospitalised patients. *Journal of Hospital Infection, 72*(4), 337-341.

Yazaki, M., Atsuta, Y., Kato, K., Kato, S., Taniguchi, S., Takahashi, S., . . . Inoue, M. (2009). Incidence and risk factors of early bacterial infections after unrelated cord blood transplantation. *Biology of Blood and Marrow Transplantation, 15*(4), 439-446.

Yip, C., Loeb, M., Salama, S., Moss, L., & Olde, J. (2001). Quinolone use as a risk factor for nosocomial *Clostridium difficile*-associated diarrhea. *Infection Control and Hospital Epidemiology, 22*(9), 572-575.

Zhang, C., & Zhang, S. (2002). *Association rule mining: models and algorithms*: Springer-Verlag.

Zhang, Y., Dang, Y., Chen, H., Thurmond, M., & Larson, C. (2009). Automatic online news monitoring and classification for syndromic surveillance. *Decision Support Systems, 47*(4), 508-517. doi: http://dx.doi.org/10.1016/j.dss.2009.04.016.

Zimmerman, R. K. (1991). Risk factors for *Clostridium difficile* cytotoxin-positive diarrhea after control for horizontal transmission. *Infection Control and Hospital Epidemiology*, 96-100.

Zoutman, D. E., Ford, B. D., Bryce, E., Gourdeau, M., Hébert, G., Henderson, E., & Paton, S. (2003). The state of infection surveillance and control in Canadian acute care hospitals. *American Journal of Infection Control, 31*(5), 266-273.