The social and biological effects of patient-patient co-presence on health in hospitals using electronic medical records

Jeffrey Lienert

Green Templeton College
University of Oxford

A thesis submitted for the degree of Doctor of Philosophy

Trinity 2018

Abstract

Co-presence, defined as two people being physically proximate to one another, is a ubiquitous and important phenomenon that remains understudied. There is strong reason to believe that co-presence may affect health, but it is likely that these effects are relatively small. Because of this, relatively large sample sizes are needed to reliably detect these effects, and the data to test such hypotheses has only recently become widely-available. In this thesis, I use electronic medical records and hospital administrative data to assess how patient-patient co-presence in a health care system may affect patient health outcomes.

In Chapter 3, I examine the social effects of co-presence on 5-year survival in a group of 4,791 chemotherapy patients. Because no metric for measuring co-presence precisely addressed all the nuances of hospital administrative data, I create a method to detect when patients are co-present more often than expected by chance, terming this consistent co-presence. Consistent co-presence thus allows me to subset co-presence to only that which is likely systematic enough to elicit social influence. Using this, I construct a consistent co-presence network. I then model 5-year survival on 1) whether a patient had any consistent co-presence in the network, 2) the number of patients who survived with whom one was consistently co-present, and 3) and likewise the number patients who did not survive with whom one was consistently co-present. I find that being consistently co-present with at least one other patient increased one's likelihood of 5-year survival compared to being consistently co-present with no one. Being consistently co-present with patients who survived increased one's likelihood of 5-year survival, and being consistently co-present with patients who did not survive decreased one's likelihood of 5-year survival.

In Chapter 4, I assess the ability to predict subsequent infection based on the number of hours a patient spends co-present with another patient suspected of infection. Across five nosocomial infections, I find that this tool has a sensitivity from 0.95 to 1.00, and a specificity from 0.90 to 1.00. If this metric were put in place prospectively, I estimate that it would lead to detecting infections between 4 and 32 hours earlier than the current standard operating procedure. I then use this information, along with biomarker information to detect subclinical infections in Chapter 5. Subclinical infections are those where the bacterial or viral load is below a test's threshold, meaning these infections go undiagnosed. I use a random forest model to perform the classification, and a variety of regression models to examine the validity of said model. I then show that subclinical infections have negative effects both on the affected patients and on the nosocomial disease dynamics, leading to increased infectious outbreak sizes.

As a supplement to support my analyses in Chapter 5, I develop an efficient algorithm to be used in social networks analysis for the colored triad census in Appendix A. I apply this to the outbreak networks observed in Chapter 5 to understand the patterns of connections of subclinically-infected patients.

In sum, I find that co-presence is a useful and informative construct which allows us to better understand patient health in hospitals. Additionally, the outcomes observed here are not exclusive to the health care setting; social influence and infectious disease spread both occur outside of hospitals. As a result, this research opens up a wide variety of future work, including studying these effects in more detail with hospitals and using similar data sources to examine these effects in other populations and settings.

The social and biological effects of patient-patient co-presence on health in hospitals using electronic medical records

Jeffrey Lienert

Green Templeton College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2018

Acknowledgements

This work was funded by grants from the National Human Genome Research Institute, National Institutes of Health (Grant number ZIA HG200335), and the Oxford Martin School, University of Oxford (Grant number LC1213-006). I would also like to acknowledge support received from Green Templeton College and the Biomedical Research Alliance. This thesis would not have been possible without receiving the data from Mr. John Finney.

I first and foremost wish to thank my advisers: Dr. Laura Koehly, Prof. Felix Reed-Tsochas, and Dr. Christopher Steven Marcum. Listing all the ways each was crucial to the completed thesis would take up far too much space. Suffice it to say that all three were instrumental in providing thoughts and comments on my work, as well as on professional development more generally. They all also not only permitted, but actively encouraged me to pursue my research interests, even if that meant my work moved out of their areas of expertise.

I would also like to thank the members of my examination committees. From my first Transfer of Status, Prof. Gesine Reinert and Prof. Steven New provided me with important feedback both on Chapter 3, and on framing the bigger picture of my thesis. Prof. Gesine Reinert and Prof. David Barron provided additional feedback on my second Transfer of Status. From my Confirmation of Status, Prof. Martin Landray and Prof. David Barron provided helpful feedback, particularly with respect to the ethical considerations of my work.

Lab members at both the National Institutes of Health and the University of Oxford were helpful in both solving small, immediate problems and larger picture questions. Specifically, Dr. Andrew Elliott provided helpful feedback on the models in Chapter 3. Dr. Omar Guerrero provided helpful insight on ways to develop and validate the models in Chapter 5. Dr. Jielu Lin brought insight about the models and results in Chapter 3. Ms. Jennifer Cleary and Ms. Hena Thakur were part of many useful conversations on visualization and interpretation of results.

Finally, I would like to thank my family for all their support throughout my doctoral work. My mother, Leslie, has always been supportive of my academic career path, and has took an active interest in my ongoing studies. My partner, Kelly, as well has my greatest thanks for her patience with me as I completed this work, particularly in the final stretch. This was all the more laudable because of my irregular travel patterns which were a part of my work.

Declaration

I hereby declare that this thesis was composed by myself, and the work presented is the result of my own research, with the following acknowledgements. All my supervisors helped develop ideas and provided comments on all of my chapters. The content of this thesis has not been presented in any previous application for a degree.

The chapters of this thesis are in various stages of publication. One chapter has been published, one has been resubmitted following revisions, one is undergoing initial review, and one is in preparation. In addition, the data chapter is being prepared for a special issue on network data collection.

- **Chapter 2**: Lienert J, Marcum CS, Reed-Tsochas F, and Koehly L. Consistent co-presence: A meaningful way to subset affiliation networks to informative dyads. Submitted to special issue in network data collection in *Social Networks*.
- **Chapter 3**: Lienert J, Marcum CS, Finney J, Reed-Tsochas F, and Koehly L. 2017. Social influence on 5-year survival in chemotherapy co-presence network. *Network Science*. DOI: https://doi.org/10.1017/nws.2017.16.
- **Chapter 4**: Lienert J, Reed-Tsochas F, Marcum CS, and Koehly L. Ward co-presence time as a diagnostic indicator for nosocomial infection. Submitted to *JAMA*.
- **Chapter 5**: Lienert J, Reed-Tsochas F, Marcum CS, and Koehly L. Ward co-presence time as a diagnostic indicator for nosocomial infection. In preparation.
- **Appendix A**: Lienert J, Koehly L, Reed-Tsochas F, and Marcum CS. An efficient counting method for the colored triad census. Revised and resubmitted to *Social Networks*.

Abstract

Co-presence, defined as two people being physically proximate to one another, is a ubiquitous and important phenomenon that remains understudied. There is strong reason to believe that co-presence may affect health, but it is likely that these effects are relatively small. Because of this, relatively large sample sizes are needed to reliably detect these effects, and the data to test such hypotheses has only recently become widely-available. In this thesis, I use electronic medical records and hospital administrative data to assess how patient-patient co-presence in a health care system may affect patient health outcomes.

In Chapter 3, I examine the social effects of co-presence on 5-year survival in a group of 4,791 chemotherapy patients. Because no metric for measuring co-presence precisely addressed all the nuances of hospital administrative data, I create a method to detect when patients are co-present more often than expected by chance, terming this consistent co-presence. Consistent co-presence thus allows me to subset co-presence to only that which is likely systematic enough to elicit social influence. Using this, I construct a consistent co-presence network. I then model 5-year survival on 1) whether a patient had any consistent co-presence in the network, 2) the number of patients who survived with whom one was consistently co-present, and 3) and likewise the number patients who did not survive with whom one was consistently co-present. I find that being consistently co-present with at least one other patient increased one's likelihood of 5-year survival compared to being consistently co-present with no one. Being consistently co-present with patients who survived increased one's likelihood of 5-year survival, and being consistently co-present with patients who did not survive decreased one's likelihood of 5-year survival.

In Chapter 4, I assess the ability to predict subsequent infection based on the number of hours a patient spends co-present with another patient suspected of infection. Across five nosocomial infections, I find that this tool has a sensitivity from 0.95 to 1.00, and a specificity from 0.90 to 1.00. If this metric were put in place prospectively, I estimate that it would lead to detecting infections between 4 and 32 hours earlier than the current standard operating procedure. I then use this information, along with biomarker information to detect subclinical infections in Chapter 5. Subclinical infections are those where the bacterial or viral load is below a test's threshold, meaning these infections go undiagnosed. I use a random forest

model to perform the classification, and a variety of regression models to examine the validity of said model. I then show that subclinical infections have negative effects both on the affected patients and on the nosocomial disease dynamics, leading to increased infectious outbreak sizes.

As a supplement to support my analyses in Chapter 5, I develop an efficient algorithm to be used in social networks analysis for the colored triad census in Appendix A. I apply this to the outbreak networks observed in Chapter 5 to understand the patterns of connections of subclinically-infected patients.

In sum, I find that co-presence is a useful and informative construct which allows us to better understand patient health in hospitals. Additionally, the outcomes observed here are not exclusive to the health care setting; social influence and infectious disease spread both occur outside of hospitals. As a result, this research opens up a wide variety of future work, including studying these effects in more detail with hospitals and using similar data sources to examine these effects in other populations and settings.

Contents

List of Figures

List of Tables

List of Abbreviations

BIC . . . . . Bayesian Information Criterion

CMI . . . . . Cell-Mediated Immune response

CPU . . . . . Central Processing Unit

CRP . . . . . C-Reactive Protein

EMR . . . . . Electronic Medical Record

ERGM . . . . Exponential Random Graph Model

EU . . . . . European Union

GDPR . . . . . General Data Protection Regulation

GEE . . . . . Generalized Estimating Equation

HAD . . . . . Hospital Administrative Data

ICD-10 . . . . International Classification of Disease, 10th Edition

IORD . . . . . Infections in Oxfordshire Research Database

MAN . . . . . Mutual Asymmetric Null

MDT . . . . . Multi-Disciplinary Team

MRSA . . . . . Methicillin-Resistant *Staphylococcus aureus*

NHS . . . . . National Health Service

OUH . . . . . Oxford University Hospitals

PC . . . . . Personal Computer

PCP . . . . . Primary Care Physician

PHI . . . . . Personal Health Information

RAM . . . . . Random Access Memory

RCT . . . . . Randomized Clinical Trial

RFID . . . . . Radio-frequency identification

SAOM . . . . . Stochastic Actor-Oriented Model

SES . . . . . Socio-Economic Status

SETIR . . . . Susceptible Exposed Test-susceptible Infectious Recovered disease dynamic model

tERGM . . . . Temporal Exponential Random Graph Model

UK . . . . . United Kingdom

US . . . . . United States of America

VA . . . . . Veteran's Administration

1 Introduction

Contents

1.1 Abstract

In this chapter, I describe the motivation for the thesis, the appropriateness of the data sources I use, and the specific aims that I address in the thesis. The motivation for this work includes a summary of the existing literature on how co-presence can impact health, the gaps in this literature, and how Big Data can enhance our

understanding thereof. In examining the literature, I draw from a variety of fields, including sociology, psychology, and epidemiology, which have all used co-presence data in different ways. I show that despite past work looking at this problem from a variety of angles, there remains a gap in the literature in using and quantifying co-presence as a potential cause of health effects.

To understand how co-presence can relate to health, I begin by looking at the more throughly-studied question of how social interaction impacts health. This includes pathways which do not necessarily require co-presence, but often occur when individuals are co-present. Indeed, this correlation is often strong enough that co-presence is used as a proxy indicator of social interaction. Although co-presence is largely used as a proxy in this context, it gives a basis for understanding how co-presence on its own may lead to health effects. I examine how co-presence, irrespective of social interaction, can relate to health outcomes, both biologically and psychologically. Importantly, I show where there are gaps in this work, which I aim to address with this thesis.

I then turn to different analytical methods and types of data and research approaches which could potentially address the relationship between co-presence and health. I discuss social networks analysis, and how it is particularly useful in looking at co-presence. There are many potential ways to collect data on co-presence, and they each have strengths and weaknesses. "Big Data" is one such approach, which is becoming increasingly common due to its increasing ubiquity. I therefore discuss the strengths and weaknesses of Big Data generally. These include hefty limitations centered around the lack of ownership by researchers. I dissect these limitations, what they mean for researchers, and how we can best proceed in the face of them. Fortunately, some of these limitations are mitigated in the specific case of electronic medical records and administrative data, and I discuss why this is so. This is important, as these data sources form the basis of analyses in this thesis.

Finally, based on the identified gaps in the literature and strengths of electronic medical records and hospital administrative data, I delineate my aims for the remainder of the thesis, and how they correspond to the subsequent chapters. In

short, I aim to assess how patient-patient co-presence, as measured using electronic medical records and hospital administrative data, affects patient health while in the hospital.

1.2 The relationship between co-presence and health

Co-presence, defined by two people being physically close to one another, is something that happens to most people many times each day. The definition of "near" can alter the observed effect of co-presence. In the context of hospital administrative data, I specify "near" to mean "within the same hospital ward bay", where a ward bay is a subspace of a hospital ward, often coinciding with a single room. Additionally, not all co-presence is created equal; co-presence in different contexts or with different individuals has differing potential to affect health. Repeated co-presence in particular may be more likely to affect an individual. To examine the effects of such a ubiquitous exposure is difficult, just as discovering the health effects of smoking in the 1940's was complicated by the fact that such a large proportion of adults smoked[1]. To understand how co-presence can affect health, it therefore makes sense to first turn to social interaction, a relatively well-understood exposure which is highly-related to co-presence.

1.2.1 Social interaction

Humans are inherently social organisms, and this has important ramifications for health and wellbeing. Social interaction, occurring as a result of our social tendencies, can take the form of both positive and negative effects, such as in social support[2] and infectious disease spread, respectively[3]. Researchers have found that social interaction affects health across space[4], age[5,6], environment[7], type of relationship[8], type of interaction[9], and demographic factors[10].

Social interaction, be it with friends, family, or colleagues, has a strong impact on our lives. In general, increasing in-person interaction was linked with reduced mortality[11], particularly in the elderly[11]. This has been examined across different

types of ties as well, affecting both health outcomes such as mortality and economic outcomes such as job-seeking[12,13].

In addition to physical health, social interaction plays a role in mental health. Interactions can reduce stress, which subsequently improves mental health[9,2,14]. Social interaction in the form of care giving can also have positive effects on mental well-being[9]. However, this is not always the case; some benefit more from social interaction than others. For instance, the same interaction of care-giving can negatively impact mental health in some individuals[9].

Two major mechanisms have been proposed to explain the relationship between social interaction and health. These are relevant, as there are clear parallels as to how co-presence may impact health through similar pathways. Other than briefly in Chapter 3, I do not examine mechanisms in this thesis, as the data I use do not allow for a thorough disentanglement of mechanisms. However, showing the *potential* mechanisms by which co-presence may impact health leads to my specific hypotheses about how effects of co-presence on health may manifest, as well as suggesting how follow-up work might try to distinguish between different putative mechanisms.

First of these potential mechanisms is the stress buffering hypothesis, which posits that positive social interaction reduces stress[8]. Stress can lead to negative coping mechanisms, such as smoking or drug abuse, or can activate physiological pathways which in turn impact health, such as the parasympathetic nervous system[15].

However, social interaction is also associated with improved health in the absence of stress. This led to the main effect hypothesis, where social interaction itself leads to health improvement irrespective of stress. This can occur through the reinforcement of normative behaviors which are themselves health-promoting, through an increase in social integration[16], or through informational support which increases healthy behaviors[17].

One subset of the main effect hypothesis is that increased social interaction reduces susceptibility to infectious diseases[18]. Specifically, increased social interaction is associated with increased natural killer cell and helper T-cell activity[19]. This pathway is of particular import to this thesis, as it connects the social and biological

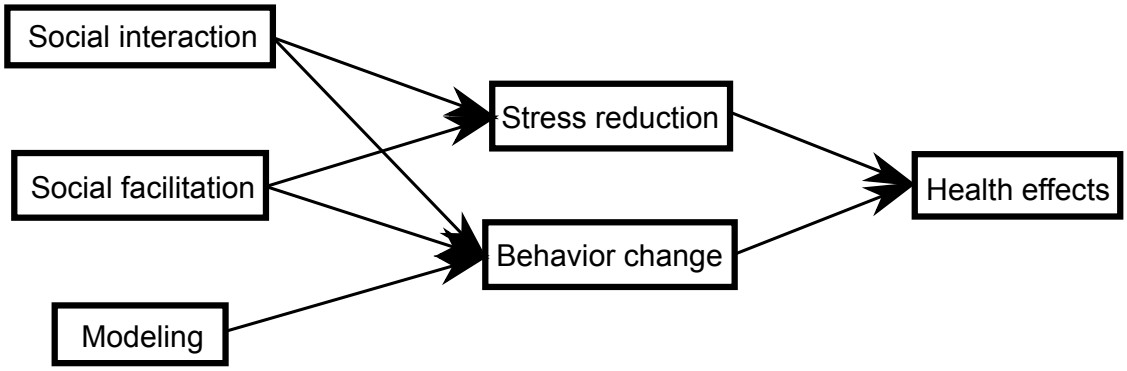

Figure 1.1: Schematic of mechanisms by which social interaction and co-presence can lead to health effects. An arrow from one item to another indicates that the first item has been shown to item in changes in the second object. The multiple, overlapping mechanisms for social interaction and social facilitation make separating the contributions of each effects difficult without data designed for the purpose. Because I cannot definitively disentangle these effects, the important feature here is the overall connection between co-presence leading to various psychosocial mechanisms and the subsequent health effects.

pathways of the effect on co-presence in health. Also, it suggests that we can't simply think of social and biological mechanisms as separable and additive, since there are circumstances where they might interact. I will return to how co-presence relates to infectious disease spread later in this chapter.

1.2.2 Social facilitation and modeling

Given the growing body of evidence that social interaction can affect health, it is important to disentangle the effects of co-presence from those of social interaction more generally. Social interaction is defined as two individuals communicating with one another in some form, be it through speech, text, or non-verbal body language, etc. Studies have shown an effect of co-presence on health independent of social interaction. The theories of social facilitation and modeling explain how co-presence alone could affect health. These are phenomena where social interaction is not actually required, only that individuals are co-present. The general mechanisms by which social facilitation, modeling, and social interaction, can impact health are shown in Figure 1.1. Again, due to the nature of the data, I am unable to conclusively separate these mechanisms. Rather, they make it clear that co-presence can affect health independently of social interaction.

Social facilitation is the idea that being co-present with those who are performing the same task will increase one's own performance of that task[20]. Social facilitation has been tested in a variety of settings and with a variety of tasks being performed. Research has shown that when near other cyclists, cyclists increase their average speed via better maintenance of speed throughout[21]. This affects intellectual as well as physical tasks, as research has shown test performance improves when around others[22]. Social facilitation can also lead to potentially negative outcomes: people have been shown to eat more when around others, even if they cannot see what those others are eating[23]. In all cases, the presence of another performing the same task leads to a change in performance. This effect is also generally independent of the actual proficiency of the others' performance[21].

Examining how the presence of others leads to the effects observed via social facilitation has lead to the hypothesized mechanism that arousal increases when someone is around others performing the same task[24]. This arousal includes the release of neurotransmitters such as adrenaline, which can improve stamina and performance[25]. Similar to the stress buffering hypothesis of social interaction, a reduction in stress based on co-presence alone can also lead to the observed effects of social facilitation[26] (arrow from social facilitation to stress reduction in Figure 1.1). This is particularly pronounced when someone is in the presence of family and friends as opposed to strangers[27].

Differential behaviors and performance can have a marked impact on health. Such impact can be direct, as in the case of increased caloric intake leading to obesity and health problems[28]. It can also be indirect, as increased competitiveness in physical activity can lead to additional confidence, further leading to increased physical activity which can improves cardiovascular health[29,30] (arrow from social facilitation to behavior change in Figure 1.1). Stress itself can also lead to reduced health, so reductions in stress via social facilitation can lead to improved health[31,14]. Social facilitation, occurring because individuals are co-present, can therefore impact health outcomes.

Modeling, on the other hand, is the theory by which one increases performance by mimicking someone else performing the same task, again without the need for social interaction[23]. Modeling has been tested in a variety of settings generally finding that task performance (good or bad) increases when modeling occurs[32,33,23]. This can directly lead to health effects when the behavior in question is related to health. As with social facilitation this can include physical exercise[34] and caloric intake[35]. It also includes behaviors such as seat-belt use, a seemingly-minor behavior that can prevent death or serious injury[36]. Although the effects of modeling on immediate behavior can be pronounced, affecting long-term change can be difficult if done through modeling, as behavioral changes are often transient[37].

The mechanism behind the effects of modeling is relatively straightforward: individual's observe others' behavior, and directly alter their own behavior accordingly to matched the observed behavior (the arrow from modeling to behavior change in Figure 1.1. The eventual health outcomes then depend on whether the behavior being modeled is one that promotes wellbeing. This can be complicated if someone observes conflicting behaviors, as often occurs in realistic settings. Therefore, although modeling can affect behavior and subsequent health, it can be difficult to cleanly measure outside of controlled environments. This is something I discuss in more detail in Chapter 3.

In addition to providing a theoretical reason to believe that co-presence may lead to health effects, there are also gaps in the literature with respect to study design and setting in evaluating co-presence. Modeling and social facilitation are often studied in cases where there is a clear task to be performed, and participants are often cognizant of this task. The effects of modeling and social facilitation in settings where this isn't the case are unknown. In Chapter 3, I therefore leverage the fact that although they know their data is being collected, patients' primary concern when receiving chemotherapy is their treatment, allowing me to test whether effects such as modeling and social facilitation may occur in settings such as this.

1.2.3 Proxy for social interaction

In many of the situations described previously, social interaction often occurs when the participants are co-present. Co-presence is therefore intricately tied to social interaction, although not exclusively. Many technologies facilitating interaction exist which do not require physical proximity, such as cell phones and the Internet. Indeed, these interactions can have important impacts on health as well[38]. However, they are not the focus of this thesis, as I am primarily interested in how *co-presence* impacts health. Despite these methods to interact without co-presence, many of our most important interactions still take place with physical proximity[39]. Furthermore, many of our digital interactions are used to maintain contacts we have previously grounded in face-to-face interaction[40]. Because of this, co-presence remains inextricably tied to social interaction for our strongest interactions. As a result, co-presence can be used as a strong proxy for social interaction. In this way, even though co-presence may have an independent contribution towards health, it can also reinforce our understanding of social interaction.

One of the earliest studies to examine co-presence was the Southern Women's study, where co-presence at various events helped the authors to understand the society of the American South[41]. Although Davis et al. did not have any data on the actual social interactions of his study population, they knew when women attended the same social event. Information on the women's co-presence was arguably sufficient to use as a proxy for more finely-grained social interaction, and resulted in meaningful conclusions. This assumption that co-presence was highly correlated to social interaction lead to important conclusions while increasing the efficiency of data collection.

Theories of how social interaction impact health, such as social support, also often depend on co-presence. While information can be sent digitally, providing care-giving support that requires physical help cannot, thus necessitating co-presence. Additionally, the amount of time two people spend together is correlated with how much social support they often exchange[42]. As a result, co-presence can serve as an important proxy for certain types of social interaction, such as social support.

Social support is the provision of help to members of one's social network[2]. Increasing perceived and observed social support generally coincides with better health outcomes[2]. Social support also takes a variety of functions: instrumental, emotional, and informational[43]. Instrumental social support refers to tangible support one provides, such as babysitting. Emotional social support is defined as making someone feel cared for, bolstering their sense of self-worth. Informational support is defined as sharing beneficial information with someone to help them navigate their situation. Although generally positive, these types of social support can also be negative in their effects. For instance, in those with mild intellectual disability, social support can lead to social strain, negatively affecting mental health[44].[1]

More recently, the growth of "Big Data"[2] and digital sources of data collection have made co-presence data more common, even when data on specific interactions are not available. In these cases, using co-presence as a proxy for social interactions may give useful information on its own. For instance, cell phones can be used to monitor when two people are near one another[45]. This data can then be used to determine how co-presence, and social interaction by proxy, may relate to various behaviors or outcomes[46]. This final example shows one of the main uses of co-presence in the future: gleaning it from sources of Big Data where co-presence can be easily determined and monitored, and use it to understand social interaction in situations where the co-presence may be meaningful .

1.2.4 Infectious disease spread

In addition to being a useful proxy for social interaction, co-presence also may independently impact health through biological mechanisms. Although the exact mechanism affecting health may differ, the underlying exposure of co-presence affects health in both cases. The most intuitive of these is the spread of communicable

[1]For a more thorough discussion of social support, and how it may affect patients on a chemotherapy ward, see Chapter 3.

[2]Throughout this thesis, I use "Big Data" as a general term referring to data sources that either contain many observations, or many variables per observation. It is not meant to delineate any specific datasets, or define a cutoff for what qualifies vs. what doesn't, but rather speak to the trend towards larger datasets with more observations and variables. I will also use the term "large observational datasets" synonymously, as this is the parlance often used by epidemiologists.

disease. Many infectious vectors do not survive long outside of the human body, and so direct, or near-direct, contact is required[47]. Co-presence with infected persons, when measured at a fine enough level, may effectively quantify a person's risk of subsequent infection[45].

This observation is the basis of contact tracing, where a person with a confirmed infection has their recent social contacts (i.e. those with whom they were co-present) examined for the infectious vector[48]. Ring vaccination (Figure 1.2), vaccinating those who are identified via contact tracing, was instrumental in the elimination of smallpox[49]. Co-presence, therefore, is a valuable tool in understanding the spread of infectious disease.

Despite the well-known correlation between co-presence and subsequent infection, the relationship has generally neither been quantified nor assessed prospectively. Although R_0, an estimate of the average number of infections caused by those initially infected by a disease, is often dependent on co-presence for the spread of infection, it only implicitly assesses co-presence rather than directly. The reasons for this have largely been practical - R_0 is an important epidemiological parameter that adequately models many infections, and there have been a paucity of data sources which allow for accurate quantification of co-presence on an entire population. This is shifting with the increasing ubiquity of electronic medical records and hospital administrative data. Addressing this gap in the literature is one of the goals of this thesis (Chapter 4). Additionally, using co-presence data as well electronic medical record data, I attempt to infer infections where the bacterial load is below standard tests' thresholds (Chapter 5).

1.3 Using networks to understand social processes

Although co-presence can be assessed at a dyadic level, the co-presence shared by a pair of individuals independent of other individuals, additional insight can be gained from taking a networks perspective. Social networks analysis is a field that has existed for over 80 years[51], and is aimed at understanding how humans interact with one another. A more recent, related, field, is network science, which

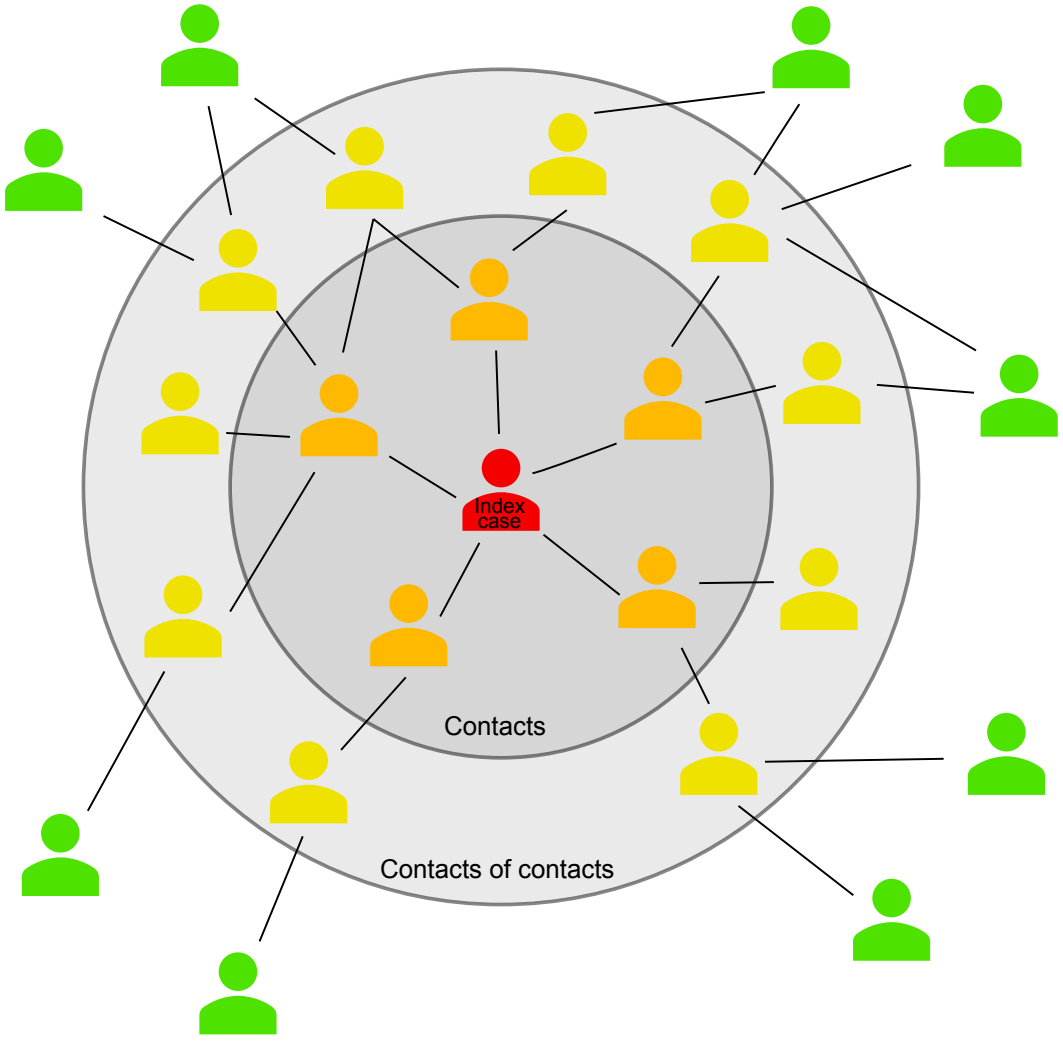

Figure 1.2: Schematic of ring vaccination. An index case infected with an infectious disease, is identified (red individual). Edges between individuals indicates co-presence. People with whom the index case has come into contact (orange), and the contacts of their contacts (yellow) are subsequently identified and vaccinated. If done in time, potential secondary infections are prevented (green). Adapted from Gerlier [50].

has existed for 20 years, which also aims to understand how humans interact with one another. It has also expanded to encompass a more general set of nodes and edges, and typically uses a different set of analytical techniques[52]. In addition to understanding how social networks and human interaction occur and lead to a variety of different outcomes, these fields have developed numerous techniques to address the crucial assumption of many statistical models, which is that that observations are independent. Below, I briefly describe the basics of social network analysis (SNA), which I draw on in later chapters.

Constructing a network involves defining a set of individuals, or nodes, and the connections between them, or edges[53]. There are a number of different terms across fields for the same thing, such as nodes also being called vertices or actors, and edges also being called links or arcs. In the case of a co-presence network, the edges are formed when two individuals are co-present with one another. The first goal of SNA is often description of a network, using a variety of well-understood metrics with summary statistics. These include attributes of the network as a whole, such as density, which is the fraction of observed edges of all possible edges that could exist in the network. This also includes node-level attributes, such as degree centrality, which is the number of edges each node is incident upon. In co-presence networks, these metrics can impart important information about the patterns of co-presence between people and their subsequent health. For instance, using measures of a person's degree in a co-presence network strongly predicted mortality[54].[3]

More complex analytical methods can address the dependence issues directly. For example, network structure can be modeled via Exponential Random Graph Models (ERGMs), which can use any number of structural properties and nodal attributes the likelihood of a network observation (given a Bernoulli baseline and the vector of model coefficients, assuming a non-degenerate model)[55]. This includes the tendency for triangles to form (triplets of nodes all connected to one another), the likelihood for nodes with similar attributes to connect, or to propensity for a

[3]In Orth-Gomer and Johnson[54], the population is a random sample from Sweden, and as such avoids the dependency issue. I use a similar approach to sample only unconnected individuals as a sensitivity analysis in Chapter 3.

node to have many incoming edges. Modeling a network in this way allows one to tease apart the tendency for different types of connections and structures to form. When related to social phenomena, this can lead to insights about human social interaction[56]. For instance, modeling a co-presence network with an ERGMs can lead to insights about the ways in which people become co-present[57].[4]

More complex still, when network data are longitudinal, methods such as temporal ERGMS (tERGMS)[58] and Stochastic Actor-Oriented Models, model the change over time in a network[59]. These models typically use panel data, a series of cross-sectional network data at specific points in time, and include specific functions for how nodes create or destroy edges between time points. As a result, they are an important bridge between static networks, and true temporal networks which include data on a continuous time scale. Temporal ERGMs can lead to important conclusions about how individuals become co-present on a more precise temporal resolution than using static networks. Because co-presence is an often transient occurrence, the networks of co-presence are able to change relatively rapidly, necessitating the use of temporal methods. Although they are powerful models, and can model time in a continuous manner, they work best with panel data, or a series of cross-sections, which is different than the data I use in this thesis. As a result, I use models focusing on networks where the data exists in continuous time, which allows for a richer understanding of the underlying processes.

Networks where the data are in continuous time are otherwise known as temporal networks[60]. By continuous time, I mean that each edge between two individuals is present for a set period of time, rather than wholly extant or wholly absent. This has important consequences for processes on networks when the rate of edge formation or deletion is of a similar time-scale as the process occurring on the network. For instance, research showed that there was a greater difference in results when comparing a static and a temporal network, than when comparing static networks to a homogeneously-mixing population (that is, considering a completely-connected network)[61]. In the case of infections and co-presence networks, the

[4] I use ERGMs to form the basis of a null model, or baseline model, for simulations in Chapter 5.

infection and changing co-presence occur at similar time-scales. In other words, the rate at which an individual's co-presence changes is of a similar time scale as to the rate at which one's infection status changes. If the rates are wholly different (e.g. kinship edges and infection), then a static network is often adequate. As a result, a temporal network is the most appropriate approach in the situation of co-presence and infection, and I use them to help understand nosocomial spread in Chapter 5. The methodology for analyzing temporal networks is less settled, as this a very active area of research.

Social networks analysis therefore affords a key perspective and approach to examining co-presence data. In the next section I turn to different ways of collecting that data, and how utilizing electronic medical record and hospital administrative data is a novel approach that can lead to important conclusions about co-presence and human health.

1.3.1 Bipartite networks

As a brief aside, the analysis of bipartite networks is due special consideration. A bipartite network is one where instead of a single type of node, there are two types of nodes, which can form connections between, but not within, types. This is of particular import, because co-presence networks as I define them here begin as bipartite networks with patients and wards or ward bays as the two node types. This mainly comes in the form of the unipartite, or one-mode, projection from the bipartite network. This is done by manipulating the adjacency matrix. The adjacency matrix of a one-mode network is square (has the same number of rows and columns), with a value of 1 in the i^{th} row and j^{th} column indicating an existing edge from the i^{th} node to the j^{th} node.

The adjacency matrix of a bipartite network, A, is an $m \times n$ matrix with nodes of two different types, arrayed along the rows and columns of A. Each entry in the matrix A_{ij} is one if and only if there is an edge from node i to node j. By multiplying the adjacency matrix by its transpose ($A \times A^T$), one is left with a one-mode square adjacency matrix that connects only one of the two types of nodes

in the bipartite graph. The values of this matrix can be greater than one if the two nodes have edges to more than one common node, but this is often simplified to be a maximum of one. This resultant square matrix is the adjacency matrix of the unipartite projection of the original bipartite network. The nodes are connected if they share edges to at least one node of the second type.

Much network analysis on bipartite graphs, including the work in the following chapters, uses these unipartite projections. This has specific implications for many aspects of the unipartite network[62]. Constructing the unipartite projection typically induces specific peculiarities which need to be accounted for. Therefore, when using one-mode networks derived from bipartite networks, one needs to be cognizant of this fact. Newman et al.[63] developed a method to determine the amount of transitivity that would occur in a unipartite projection of a *random* bipartite network, and this can be used to determine if observed transitivity is greater than expected by chance. In later chapters, I address this issue by taking care to either create networks from very precise ward bays (Chapters 4 and 5), or reducing the number of edges by subsetting to only the strongest connections (Chapter 3).

1.4 Measuring co-presence

As stated, examining co-presence and its impact on health, particularly the spread of infectious diseases, is not new. To reiterate, co-presence is when two people are simultaneously in a space together. How the space is defined, what distance constitutes co-presence, how long two people must be there for it to qualify as co-presence, etc. are all aspects that must be defined by a researcher. Co-presence has been measured using several approaches, including surveys[64], synthetic populations[65], and more recently, sensor data[45]. These methods can also be combined to gain unique information from each approach[66]. There is also administrative data, which is the approach I take. However, to give an idea as to why I do not take these other approaches, I first describe them as well as their strengths and limitations, finally showing why this leads me to use Big Data[67,68].

Surveys or diaries can be used to collect co-presence data by asking participants to recall with whom they were co-present recently[69]. The exact wording of the questions varies based on the research questions at hand. The wording of the question can also limit or exacerbate certain biases[70]. For instance, research comparing surveys and sensor data found that shorter-term contacts are not as well-represented in surveys as longer-term contacts[71].

Functions creating simulated data by simulating co-presence networks from well-known characteristics of human social behavior has resulted in explaining a variety of outcomes without ever directly collecting co-presence data[65,72]. Studies have shown that synthetic datasets can very closely approximate the results of large-scale surveys on co-presence, indicating that this method can behave quite well[73]. However, because the data are synthetic, there is no way to know whether or not the data accurately represent *co-presence*.

Sensor data consists of giving participants some form of wearable device or app that records when it is in close proximity to another device, often by Bluetooth, Wifi, or RFID[74]. This allows for an accounting of co-presence unaffected by some of the biases inherent in other approaches, such as recall bias. Additionally, it can collect co-presence data at a much finer temporal resolution than survey data, allowing for richer analysis[75]. However, sensor data also has flaws: it only works on participants wearing the sensor, which can lead to unknown biases if only a subset of people are fitted with one[75].

In addition to being measured by multiple methods, co-presence has also been assessed on populations or samples in a variety of locations, including academics at conferences[74], employees within office buildings[76], students in schools[77] and patients and health care professionals in hospitals[78]. These populations provide particularly appropriate uses of co-presence data, as it ameliorates some of the limitations regarding biases in the aforementioned methods. These multiple populations have allowed us to see how co-presence varies based on setting, generally finding that the aggregate patterns of co-presence look very similar across settings[67]. This means

that findings from co-presence in one context may generalize to others, expanding the impact of studies using co-presence data.

One important issue in the collection of co-presence data is the large heterogeneity of the quantities of co-presence involved. The time spent co-present can be a difference of orders of magnitude from one end of the continuum to the other[79]. This makes a number of standard quantitative techniques ineffective, such as using a mean to describe central tendency, or applying an arbitrary cutoff to dichotomize the distribution. One approach to addressing this issue is the contact matrix of distributions, which fits a negative binomial distribution to the co-presence times between people with a variety of predefined roles, or types of nodes[80]. However, if only one type of node exists, then this method loses much of its ability to capture important heterogeneity, as there are no between-class parameters to estimate. This is one open problem to which I propose a potential solution in Chapter 3.

All of these methods have important limitations which limit their ability to assess how co-presence affects health for patients in hospitals. Therefore, I turn to Big Data, specifically electronic medical records (EMR) and hospital administrative data (HAD). Although having limitations of its own, which I describe in more detail in the discussions of Chapters 3, 4, and 5, quantifying co-presence using this data ameliorates many of the limitations of the other methods described. Like sensor data, it eliminates many biases inherent in self-reporting. Unlike much of sensor data, because it is collected on the entire population of patients, it also removes many potential biases from having non-representative samples. However, human error may still be present. Next, I discuss some of the positives and negatives of Big Data in general, and dive more specifically into how they work when using electronic medical records and hospital administrative data. From a broader perspective, the novelty in this thesis lies neither in using electronic medical records nor in studying how co-presence affects health, but in the intersection of these two ideas.

1.5 Big Data

To better understand how co-presence impacts health, it is important to use data suitable for the questions at hand. Historically, this has meant collecting data through surveys[81] or observing social interaction over time via ethnographic work[82]. However, these data sources are not necessarily equipped to answer several open research questions related to co-presence effects on health which may be relatively small. Large observational datasets, often called "Big Data", are becoming ubiquitous and their use in answering scientific questions is becoming more prevalent. Importantly, this is not to say that Big Data is superior to other discussed methods in all aspects. Much of the work I present in later chapters would be augmented by the use of other approaches due to their specific strengths. In this section, I will discuss the various benefits and downsides of using Big Data, specifically as it pertains to electronic medical records and administrative data and to answering questions about co-presence and health. Big Data, although referring to a large variety of quite heterogeneous data sources, often indicates a shared set of characteristics. These characteristics can often be broken down into those that are generally positive, and those that are generally negative.

1.5.1 The benefits of Big Data

Strengths of Big Data include that it 1) has many observations and 2) is constantly updating[83]. These benefits allow one to use the data to answer questions not typically feasible with other datasets.

The fact that Big Data is big is not surprising - it is stated in the name. Even so, the evident importance of its size merits mention. It allows one to have a much larger study population than would otherwise be feasible. For instance, what would in the past have been only cost-effective in a case-control design, can now be done for a similar cost using electronic medical records[84]. This is because the sheer number of patients in many data sets allows even rare conditions to have large enough sample sizes to provide adequate statistical evidence.

Additionally, the size of the datasets allows one to detect relatively small effects. This is because a large sample size will decrease the standard error of estimated effects, allowing one to discern effects that would likely not otherwise reach statistical significance. However, specific to the health care setting, the overriding force of mortality due to the underlying morbidity is likely to overshadow any health changes due to co-presence[85]. Although this is an exciting opportunity, this can lead to the misuse and over-promising of Big Data. With a large enough sample size, almost any effect will become significant, even if it is not clinically meaningful[86]. Researchers must be cognizant of this possibility when conducting studies with Big Data.

The effects of co-presence on health are likely to fall within one of these categories - rare or small effects - for a number of reasons. First, if they were large or common, classical methods would have likely detected them sooner[87]. Second, the likely mechanisms of the effect, such as stress reduction, are not large effects in social interaction studies, and therefore are unlikely to be large when studying co-presence[31]. Because of this, using Big Data to measure these effects means an increased likelihood of detecting a signal relative to the noise.

The fact that Big Data sources are typically constantly-updating and receiving new data and observations allows for many exciting analyses. For instance, many longitudinal study designs using surveys are panels - a series of cross-sectional surveys at predetermined times[88]. Many Big Data sources instead allow this information to be collected in real or near-real time, providing a much more accurate and realistic picture of the underlying changes in whatever is under study. As stated previously, this allows one to use more complex methods, such as temporal networks instead of tERGMS, which can allow for richer insight. Additionally, once a system is in place to collect this data, collecting additional data is often relatively trivial, and carries minimal costs.

This advantage carries an associated cost - that of data overload. This occurs in two ways - lack of appropriate methodology and required computational power. There are many methods for dealing with longitudinal data from a panel design. Methods such as hierarchical linear models, growth curve models, and generalized

estimating equations can appropriately model the association induced by having multiple observations on a single person or unit of observation[89]. In approaches where there are correlations within individuals over time and between individuals due to an underlying social network structure, there are likewise advanced methods for modeling panel network data, such as SAOMs[59] or tERGMs[58]. However, in both cases, the methods for continuous-time data are less well-developed. To some extent this necessitates a bespoke approach to problems of continuous-time data. The computational power required to conduct these analyses is also a hurdle to overcome which is induced by the use of Big Data. To an extent, this can be addressed using optimization methods. I demonstrate this approach in Appendix A. Other than developing methods specifically with an eye to minimizing computational complexity, a large amount of time can be eliminated from analyses using proper techniques[90]. For instance, many networks are sparse, or contain an overwhelming majority of absent edges, and sparse matrix methods can leverage this to dramatically reduce computational time.

1.5.2 Limitations of Big Data

Although Big Data comes with important advantages, there are also potential downsides, based on the generalization that Big Data is 3) non-representative of some underlying population 4) controlled by parties with goals differing from the researcher, and 5) often replete with sensitive data.

Big Data sources often contain information on a subset of individuals, and the ways in which they enter the dataset means there is often some unknown selection into the dataset. This can make generalizing to a larger population difficult, particularly when researchers are used to working with probability samples[91]. For instance, when looking at Twitter, only people who join Twitter would be included in that dataset, and how Twitter users differ from non-Twitter users is not well-understood[92].

The sheer size of many large observational datasets means that they can only be created and maintained by organizations with sufficient infrastructure and resources

to do so. These parties generally control these datasets, and their priorities often supersede those using the data for research, which can carry a host of difficulties if a researcher is not careful. First and foremost, the data are rarely publicly available, so getting access can be difficult. Second, the data are often incomplete with respect to included variables, as what is of interest to a researcher may not be of interest to the organization. Third, the organization's interests can change over time, and with it, the very structure of the dataset. Variables that were once included may no longer be included because the organization decided such variables were not needed. Fourth, the way the dataset works may have endogenous functions that inherently confound the data. For instance, a well-known social phenomenon is that of triadic closure - an individual's friends who are initially unknown to one another are more likely to come to know one another as a result of their common third-party[93]. However, one cannot easily ask whether this occurs in Facebook friendships, since Facebook endogenously recommends friends-of-friends as potential friends[94]. Although all of these downsides can be mitigated if the dataset is under the researcher's control, the cost of implementing, running, and maintaining such a dataset can be prohibitive, and so researchers must often work with the owners of these datasets. Many of these limitations can be avoided if a researcher works with the owning party, and shows how removing some of these limitations may benefit those owning the data.

In addition to the concerns of their owners, large observational datasets necessitate concern about the individuals composing them, as such datasets may contain sensitive information on said individuals. Hospital administrative data has Personal Health Information (PHI), mobile phone data has private records of who-called-whom, etc. Very recently, Facebook has come under heavy scrutiny for their releasing data to third-party researchers, such as Cambridge Analytica, without getting individuals' consent[95]. Again, the very size of these datasets can be problematic, as the sheer amount of data on individuals may be enough to eliminate anonymization[96]. Coincidentally, the European Union (EU) government had passed a law in 2016 which fully came into effect in 2018, shortly after the Cambridge

Analytica scandal. This is the General Data Protection Regulation (GDPR), and aimed to put the consent for data usage back in the hands of consumers[97]. This regulation will impact researchers using large observational datasets in the future, including those using health care data. Irrespective of regulation, researchers need to be aware of potential breaches of anonymity when using large datasets, and proactive in mitigating these risks.

Of course, all the discussion of data used to collect co-presence information has so far presupposed that one is conducting an observational study. There is also the potential for a randomized clinical trial (RCT), the gold standard of assessing causal effects. Any of the data sources thus far mentioned could be used as part of an RCT, with co-presence appropriately randomized. However, RCTs are expensive, and are mainly designed to pinpoint an effect size rather than determine whether an effect exists[98]. Here, I ask questions that have not been asked before, and therefore an RCT is not appropriate at this stage of the research trajectory. Once I have established that there appear to be clinically-significant effects of co-presence on health by the end of this thesis, then an RCT becomes a potential next step to examine the precision of the results determined herein.

1.5.3 Hospital Administrative Data and Electronic Medical Records

Above, I have outlined the positives and negatives of using Big Data for research. In this thesis, I use Hospital Administrative Data (HAD) and Electronic Medical Records (EMR). How these benefits and negatives apply to this data merits a discussion of the strengths and limitations specific to this type of Big Data.

Although they are often used interchangeably, HAD and EMR refer to subtly different things[99,100]. EMR is data that is directly taken from a patient on the status of their health. This includes information such as diagnoses and test results. HAD, on the other hand, is data collected by the hospital for the purpose of accounting for the patient in the system, and often for billing. This includes information such as when a patient entered and left a hospital ward.

An important distinction then, is between diagnoses, which are the actual medical morbidity as identified by a doctor, and between coding systems like the International Classification of Disease, 10th edition (ICD-10), which document conditions and treatments[101]. The diagnoses themselves are encoded in a patient's EMR, while the ICD-10 code, which is often used for billing purposes, is included in the HAD. EMRs may also have more complete information; for instance, an EMR may contain the results of a blood test, whereas the HAD would only note that a blood test was ordered. Both types of data are therefore needed for a complete understanding of a patient's health and the hospital's response to said health. Because the ICD-10 are often used for billing, there is not a one-to-one correspondence between the specific medical condition and the documented ICD-10 code. Because of this, the ICD-10 codes often serve as an important proxy for diagnoses[102].

Fortunately, both of the major benefits of Big Data hold up in HAD and EMR. The data contain many observations, and are consistently updated. One minor caveat is that unlike many datasets that rely only on digital products and can be updated in perfectly real-time, HAD and EMR must interface with a physical system. Because of this, much of the data need be entered by a human coder, and as such is susceptible to error. In addition, it can be affected by various human tendencies, such as integer rounding and typos or non-uniform shorthand. For instance, ward entry and exit times are "heaped", or clustered at even numbers like the top of the hour. Fully automated systems would be able to capture the exact time, but there is likely some error in the time entry of these data systems[103]. For this reason, in this thesis, I generally perform analyses at hour intervals, as that is the most precise unit of time where measurements are accurate[104].

While the benefits of Big Data apply to HAD and EMR, some of the negatives are ameliorated in the specific case of HAD and EMR. Most importantly is that of generalizability and the patient population. Whereas the users of Facebook and Twitter are not a random sample of the underlying population, patients in a hospital are more likely to be representative of patients from the hospital's catchment area. This is particularly true in countries with single-payer insurance systems like the

UK, where low Socio-Economic Status (SES) patients are no less likely to use the health care system than higher-SES patients[105,106]. To be clear, that is not to say there are important systemic disparities in health care in these countries, only that the issue of *access* is not a major cause of these disparities. In these countries, the population in the dataset is therefore often very close to the general population, and so the results are highly generalizable.

The issue of sensitivity of personal information is also less problematic when using this data. Not because the data are less sensitive, but precisely because they are very sensitive, and have been recognized as such for a long time. There have been rules in place on the proper handling of PHI for many years, and these have generally transitioned to EMR and HAD[97,107]. This does not mean researchers can become complacent when using EMR and HAD, but it does mean that potential pitfalls will often be caught before they can cause harm.

The other downsides of Big Data, stemming from the ownership of the data are generally not ameliorated in the case of EMR and HAD. Researchers still need to gain access to these datasets. Researchers often have little say over what variable is or is not included. The dataset will change over time, particularly as the hospital system is altered based on policy changes at the national level[97]. Smaller changes, such as renovation, may also implicitly impact the data without any clear indication that it has occurred. For instance, in the data I use, a chemotherapy ward was closed in 2009, and a brand-new one with a different layout opened. Had I been unaware of this, the data in Chapter 3 would have been intractably confounded with this change midway through the data. However, as long as the researcher is cognizant of these potential issues, and reacts accordingly, their impact can be minimized or absorbed.

1.6 Summary

From the above, it should be clear that a number of health outcomes may depend on co-presence, and that Big Data in the form of HAD and EMR offers an exciting new way to measure co-presence to further understand its effect on health. Although co-presence could be measured in a hospital without using HAD, it would be

immensely inefficient, and this is likely a large part of why the above gaps in the literature exist. At the same time, using HAD to study exposures relating to social interaction other than co-presence would be difficult, if not impossible. This is because hospitals set up their EHR and HAD to capture the information that is of interest to them, and not necessarily to researchers. This does not include data detailing who-talked-to-whom or about a patient's feelings - these are beyond the general scope of physicians' interactions with patients. In this manner, Big Data and co-presence in hospitals go hand-in-hand, and can be used to answer many research questions across disciplines that have up to now been elusive.

1.7 Aims

Based on the specific gaps in the literature I have identified in previous section, combined with the benefits of using HAD and EMR above, I propose the following aims for this thesis. I will address each aim in a separate chapter. Aim 5 will be addressed in an appendix, as it is a methodological improvement rather than a substantive question.

Aim 1: Describe a large database comprising administrative data and electronic medical records. In addition to describing this specific dataset, I describe more generally the ways in which networks can be constructed from data of this type. I do this in Chapter 2.

Aim 2: Determine whether there is evidence of social influence between chemotherapy patients based on co-presence (Chapter 3). Given the open layout of a chemotherapy ward, the timing of patient schedules, and the importance of social support outside the chemotherapy ward, I believe that this setting is the most likely place to observe social influence based on patient-patient co-presence. To assess this, I develop a novel method to detect significant co-presence, or when two patients are co-present more than expected by chance. I also perform a number of sensitivity analyses to see if other factors could have lead to the same results.

Aim 3: Assess how well a count of the hours of co-presence with a patient suspected of an infection performs as a screening test for infection (Chapter 4).

Although co-presence is a known risk factor for infection, and forms the basis of containment strategies like ring vaccination, the exact relationship between *quantity* of co-presence and risk of infection is not known. I quantify this relationship for five important communicable diseases, both bacterial and viral, and with different modes of transmission, to assess the efficacy of this test. I also quantify how many hours earlier each patient's infection may have been detected if using this test in real-time.

Aim 4: Develop a model to detect subclinical infection based on EMR and HAD, particularly including co-presence with infected patients (Chapter 5). By subclinical infection, I mean a bloodstream infection that does not result in symptoms and has a bacterial/viral load below the standard test's minimum threshold. To do so, I build a random forest model (a machine learning method for classification) containing demographics, health information, biomarker levels, and co-presence with infected individuals. I assess whether identified subclinical patients experience negative outcomes to themselves, and to subsequently affect nosocomial outbreak dynamics. I also perform validation analyses to assess whether the model works as intended.

Aim 5: As part of understanding subclinical infection, I develop a computationally efficient method for the colored triad census on networks (Appendix A). This method builds on a paper by Moody[108], and uses matrix multiplication to calculate the triad census. It takes only 1.5 days on a network of $10,000$ nodes, much improved relative to the standard approach of counting each triad individually. I apply the algorithm to the Zachary Karate Club network[109], and find that the method allows one to understand homophily, bridging, and their intersection all within one analysis simultaneously.

2

Data and study population

Contents

2.1 Abstract

In this chapter I aim to orient the reader to the data I use throughout the rest of the thesis. I first describe the data and the types of variables it contains. I then contextualize the data with information regarding the catchment population in Oxfordshire, UK. Next, I characterize the data with a wide variety of statistics and visualizations to provide a general impression of how it may or may not relate to

the underlying population. Diving further into visualizations and the strengths of electronic medical records and administrative data, I explore in detail the ways these types of data can be used to relate observations to one another in the form of networks. This includes networks such as patient-patient co-presence networks (the basis of this thesis), as well as co-morbidity networks, and other types of networks. Finally, I show how I divide the data into subsets which form the basis of the study populations in the subsequent substantive chapters in the thesis.

2.2 Data

The data used throughout this thesis are subsets of the Infections in Oxfordshire Research Database (IORD)[110]. This database was created by merging administrative data and electronic medical records of all the patients seen in the Oxford University Hospitals (OUH) System from 2000 to 2015. Records were linked using standard methods, and errors were identified and corrected using graphical analysis[111]. Records are anonymized at the patient level[1].

For each patient utilizing the health care system in Oxfordshire, the database contains data on their basic demographics (sex and birth date). Each patient's flow into and out of the health care system is included: each patients entry to and exit from the system constitutes one stay. Each stay is divided into ward spells, with one spell comprising an entry to and exit from a ward. All of these entry and exit times are recorded, as are the reasons for the transfer (e.g. admission and discharge). Prior to 2011, ward bay is the highest spatial resolution of patient location included. From 2011 onwards, a patient's bed number is included.

Patients' health records are also recorded, including ICD-10 codes[101]. ICD-10 codes are primarily a coding scheme used for billing purposes, but have been

[1]It is important to note that although patient records do not have names attached, it is quite likely that individuals' identities could be reconstructed from the precise nature of the data here. Because of this, sharing this data, even in its anonymized form, could cause harm to patients through the risk of identification. Even in aggregate data, an individual may be identified if the number of patients in a specific category is small enough. For this reason, in collaboration with colleagues I only shared aggregate data with at least 5 patients in a given category.

shown to be highly correlated with actual diagnoses[112,102]. I therefore use ICD-10 codes as a proxy for diagnoses.

Blood and microbiological test orders and their results are included. Both tests list the time, type of test, value of the result. In addition to a positive/negative result, microbiological tests that return positive further show any antibiotic resistances detected in the strain.

Finally, a patient's admitting physician is known. The admitting physician is on a patient's multi-disciplinary team (MDT), and is at least partly responsible for decisions regarding the patient's health. For this reason, I use admitting physician as a proxy to account for physician-specific effects. Similar information is not available regarding nursing staff, which is a limitation of these data.

2.3 Oxfordshire population

The catchment area of the IORD is the county of Oxfordshire. People residing in this area have a primary care physician (PCP) and a hospital for all specialties within the county. The county of Oxfordshire contains 678,000 people, including students and military[113]. In addition to those residing in Oxfordshire, the OUH is responsible for those temporarily in the county.

The population is aging, but has a large proportion of University-aged students (Figure 2.1). In the past decade, the population has aged, which has important implications for the utilization of health care, both within Oxfordshire and across the country. Migration has also increased in Oxfordshire since 2009, accounting for at least one third of total population growth (Figure 2.2). This has had important consequences on the utilization of health care, particularly for morbidities not common in the English population of Caucasian ancestry. The non-zero migration into Oxford shows that the population in Oxfordshire is an open, rather than a closed, cohort. This precludes certain analyses that require a closed cohort. However, if an individual was born in Oxfordshire, they will appear in the dataset with a record on their birth date and an ICD-10 code corresponding to birth. In this manner, the closed cohort of individuals born in Oxfordshire can be inferred

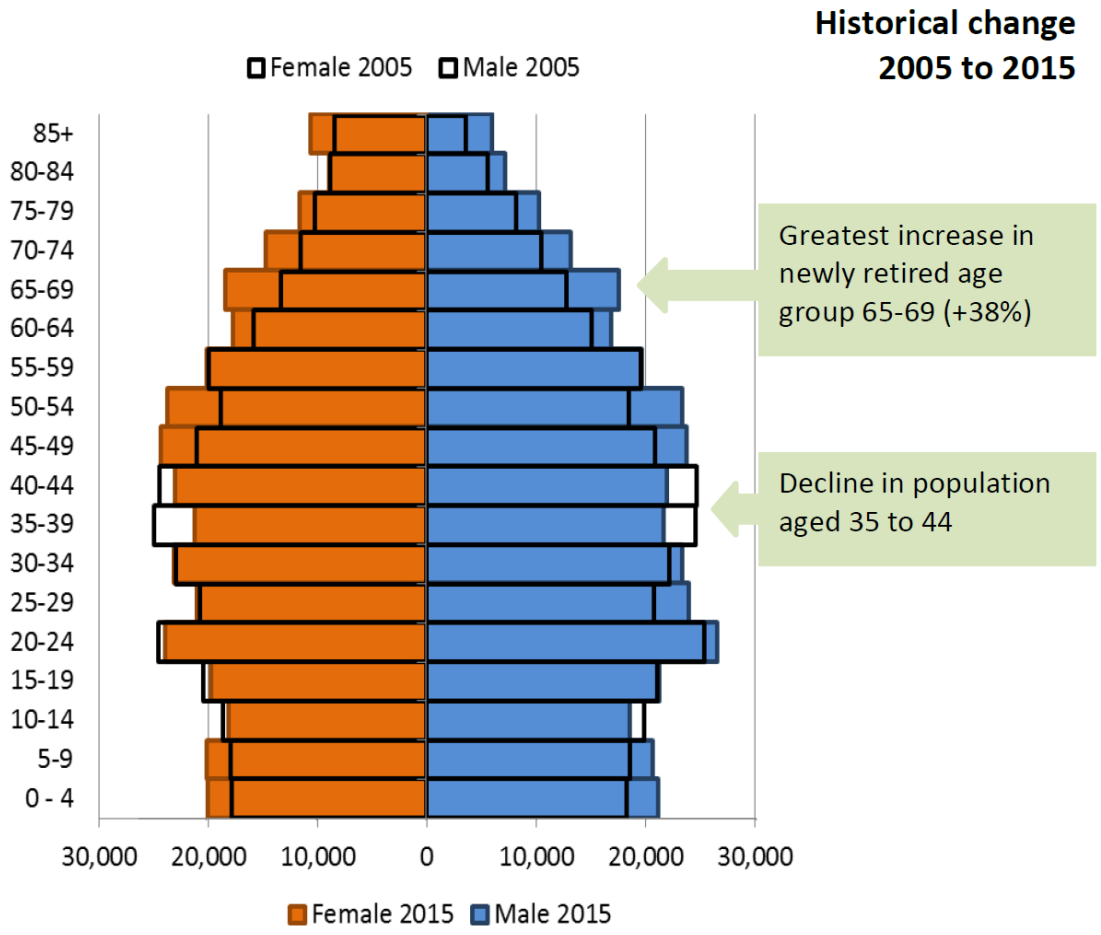

Figure 2.1: Comparison of age pyramids of Oxfordshire 2015 and 2005 populations. Orange and blue bars represent female and male populations in 2015, respectively. Black outlines represent populations in 2005. Reprinted from Oxfordshire's Health and Wellbeing Board [113].

from the data. This information can therefore be used to separate out natural change (i.e. births and deaths) from migration.

2.4 Data population

Oxford University Hospitals NHS Trust provides >90% of hospital care and all acute services in Oxfordshire. It includes two large acute-care teaching hospitals, one specialist orthopedic hospital, and a number of smaller community hospitals in Oxford and one district hospital 35 miles north.

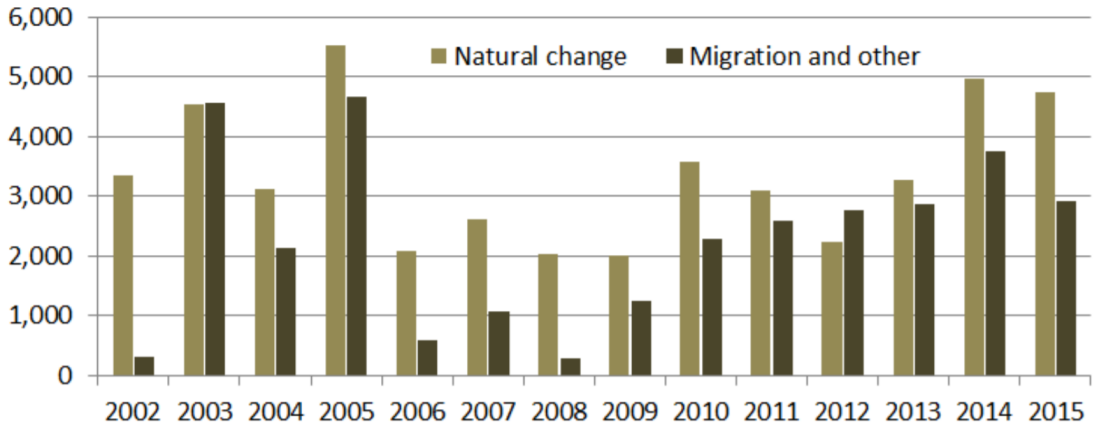

Figure 2.2: Comparison of sources of population growth from 2002 to 2015. Population growth is divided into "Natural change" which includes births and deaths, and "Migration and other" which includes any movement into or out of Oxfordshire, including to or from elsewhere in the UK. Reprinted from Oxfordshire's Health and Wellbeing Board [113].

2.4.1 Demographics

Overall, the dataset comprises 779,624 individuals, 365,860 of whom are male (46.93%), and 413,674 of whom are female (53.06%). Looking at the most recent year in the data (2015), there are 120,809 individuals, 55,137 of whom are male (45.64%), and 65,670 of whom are female (54.36%). With respect to age, as of their earliest use of the health care system in 2014, patients were an average of 46.73 years old (SD=27.17). The distribution of patients' age and sex can be seen in Figure 2.3. This distribution is highly skewed towards older patients and infants. For females, there is a spike in health care utilization from age 20-45, which is likely the result of pregnancy-related health care. I also see the rates of male hospitalization outpace those of females beginning at age 55, and continuing until age 80, the life expectancy of men.

Importantly, the data do not contain information about race or location. Post codes in the UK are highly specific, and having access to the post codes would compromise anonymity. Race is not collected by the NHS. The population of Oxfordshire is highly homogeneous, and race would potentially not include sufficient variability to be inferentially informative.

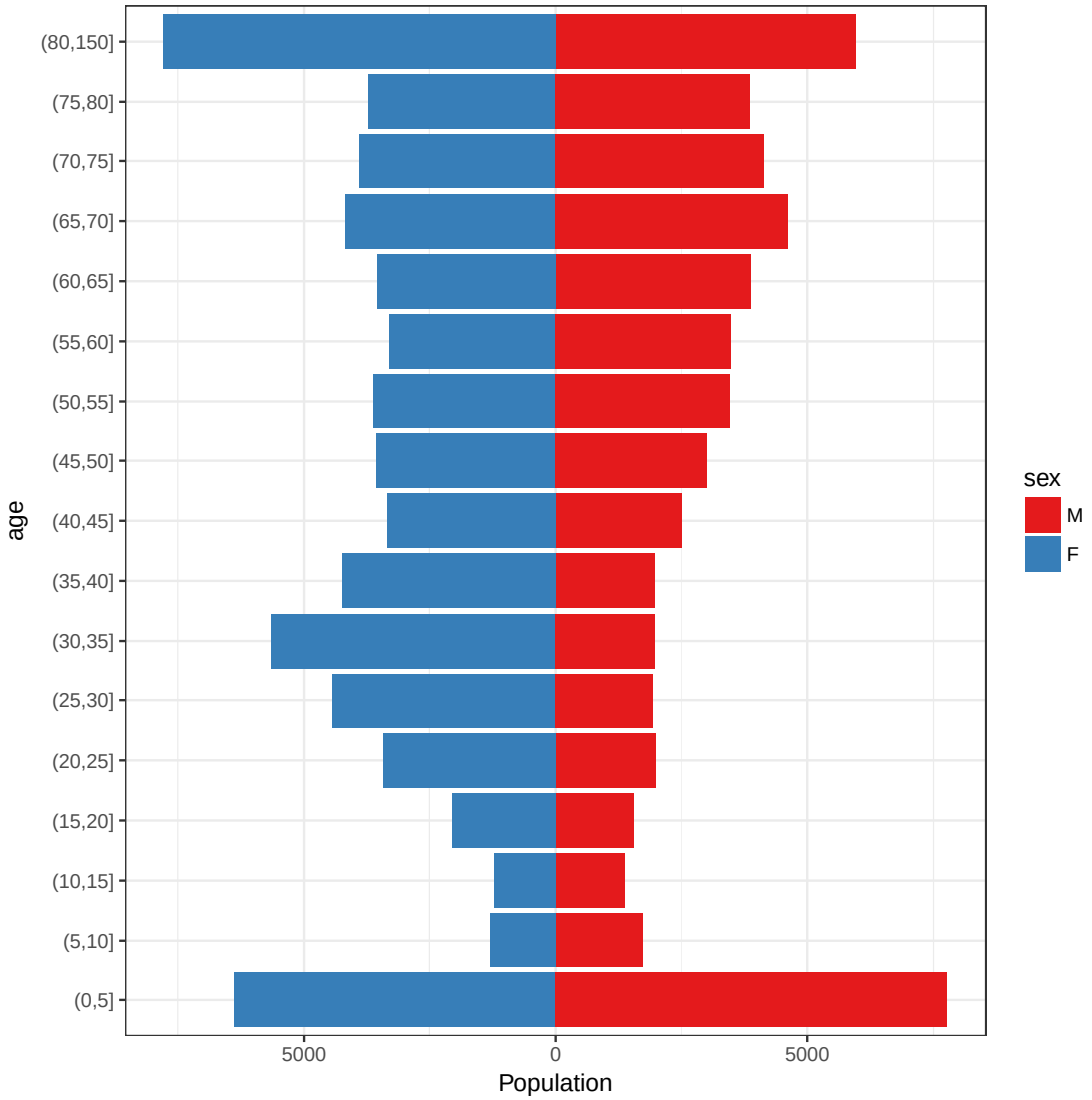

Figure 2.3: Age and sex distribution of patients in the IORD in 2014. Patients were included if they had a hospital spell beginning or ending in 2014. If a patient was observed multiple times, then their age as of their first visit in 2014 was used.

2.4.2 Mortality and morbidity

Fifty percent of patients survive to age 87 (Figure 2.4). The oldest patient in the dataset lived to past age 110.

In total, there were 4,748,758 ICD-10 code instances (not unique) in the dataset. This comprises 11,197 unique ICD-10 codes, of over 14,400 possible codes in the 5th edition[114]. The distribution of these diagnoses is shown in Figure 2.5. The distribution is heavy-tailed, with many codes appearing very few times, and a

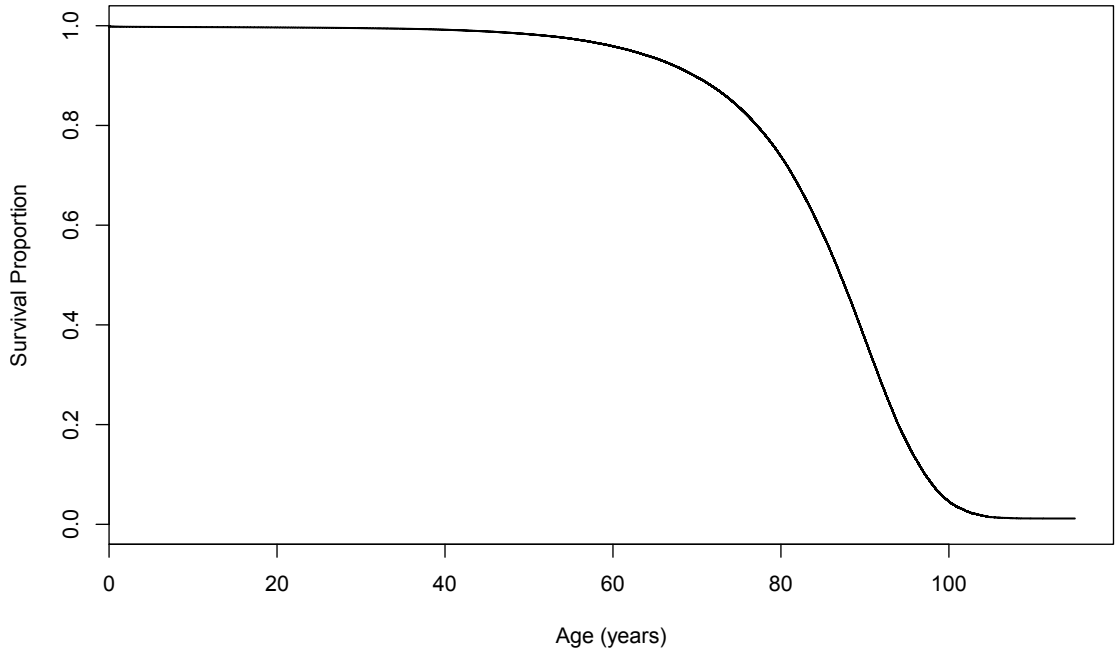

Figure 2.4: Kaplan-Meier curve of survival times for patients observed at least once in the IORD. Death dates are only missing if a person moved outside of the EU and then died. Survival time was calculated by determining the difference in years between birth date and death date or the end of the data (1 Jan 2015), whichever came first. If the latter, patient's survival time was censored.

few codes appearing a very large number of times. However, it is not power-law distributed, as the density plot is not approximately linear on the log-log scale. Instead, the tails are heavier than they would be in a power-law distribution, as the curve is convex. This is important if one were to simulate distributions of ICD-10 codes in patients. This may also give some insight into how diagnoses co-occur within people, and may merit further study. The minimum count is one (although ICD codes exist that did not appear in the dataset, I do not show the number of codes with zero uses), and the maximum count is 107,624.

To investigate the heavy end of the tail, I show the 20 most common diagnoses in Table 2.1. After hypertension, chemotherapy-related ICD-10 codes are the most common. When the diagnoses related to birth, pregnancy, and other non-terminal diagnoses are removed, the table largely reflects the main causes of death in the UK[115]. Interestingly, the difference between live births and live births in hospitals

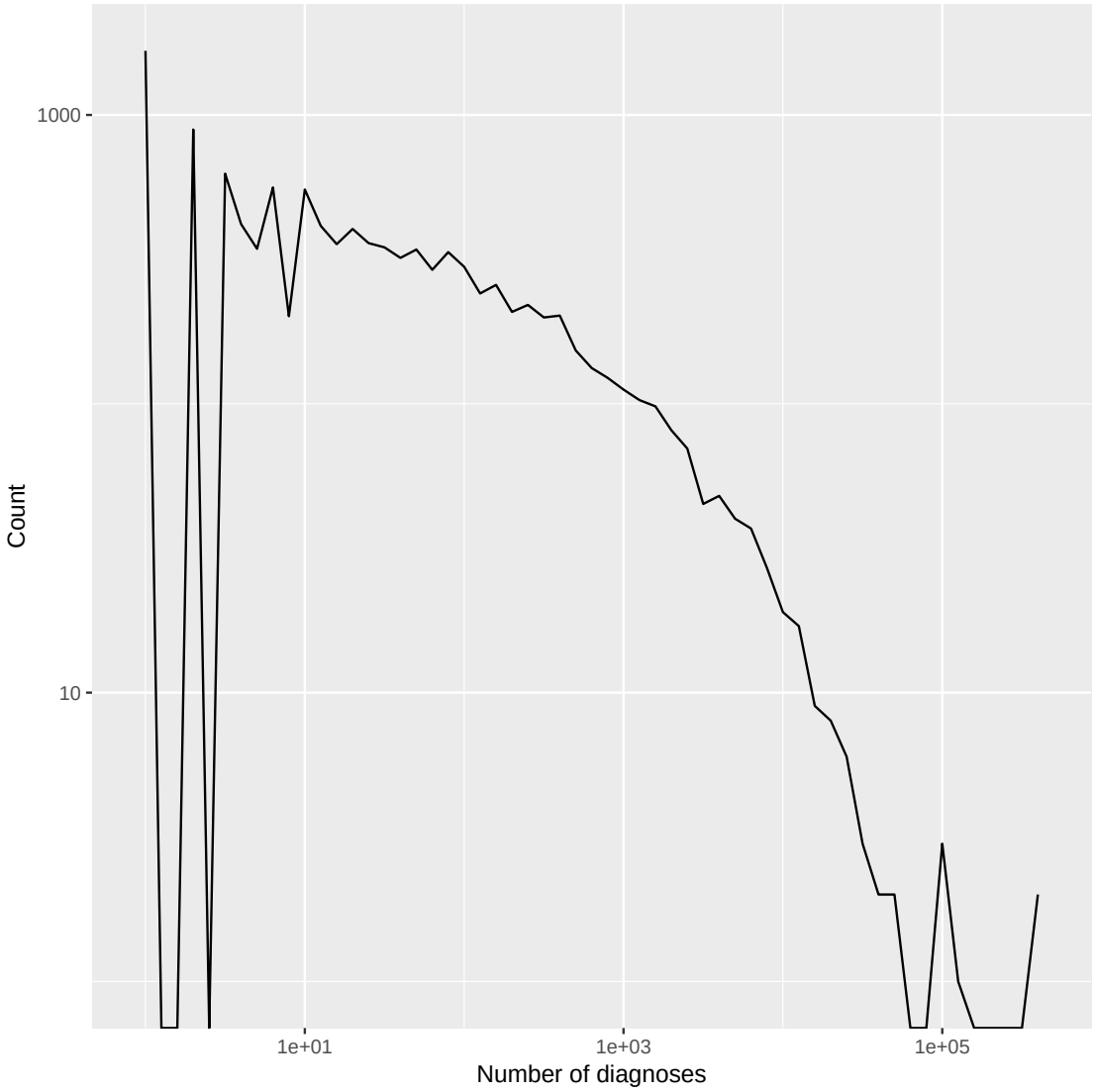

Figure 2.5: Distribution of ICD-10 diagnoses used in the data. Codes were rank-ordered according to their usage, and plotted on a log-log scale. The plot shows the data are heavy-tailed. The minimum count for a diagnosis is 1 rather than 0 because I omit any ICD-10 codes which do not appear in the dataset. The density drops to 0 following 1 and 2 diagnoses due to the log-sclae of the x-axis. There is comparatively more width between these numbers than between latter numbers, allowing the smoothing function to reduce the density to zero in-between these integers.

Diagnosis	Proportion
Essential (primary) hypertension	0.024
Encounter for antineoplastic chemotherapy and immunotherapy	0.023
Single live birth	0.020
Single liveborn infant, born in hospital	0.019
Type 2 diabetes mellitus without complications	0.010
Atrial fibrillation and flutter	0.010
Unspecified cataract	0.008
Atherosclerotic heart disease of native coronary artery	0.008
Pure hypercholesterolemia	0.007
Other and unspecified asthma	0.007
Second degree perineal laceration during delivery	0.006
Malignant neoplasm of breast of unspecified site	0.005
Personal history of diseases of the circulatory system	0.005
Urinary tract infection, site not specified	0.005
Tobacco use	0.005
Noninfective gastroenteritis and colitis, unspecified	0.005
Sleep apnea	0.004
Chronic ischemic heart disease, unspecified	0.004
Nausea and vomiting	0.004
Secondary malignant neoplasm of liver and intrahepatic bile duct	0.004
Anemia, unspecified	0.004
Other specified pregnancy related conditions	0.004
Angina pectoris, unspecified	0.004
Unspecified acute lower respiratory infection	0.003
Secondary malignant neoplasm of bone and bone marrow	0.003

Table 2.1: 25 most common diagnoses in the patient population, as measured by proportion of all ICD-10 diagnoses. ICD-10 codes were left at their most specific value rather than grouping similar codes under common groups of morbidities. The underlying total of ICD-10 codes in the dataset was 4,748,758.

can ostensibly be used as an indicator of how many and which births occur at home or other locations outside of the hospital. This could be a useful way to study the differences between health outcomes of hospital vs. home or outside of Oxfordshire births. With respect to co-presence, this could also be used to observe the effects of patient-patient co-presence for infants.

Subsets of diagnoses play an important role in subsequent chapters, specifically cancer and infectious diseases. I therefore examine just diagnoses pertaining to these diseases, and their relative frequencies. In Table 2.2, there are diagnoses related to cancer of the breast, liver, and bone. There were a total of 366,025

Cancer diagnosis	Proportion
Malignant neoplasm of breast of unspecified site	0.071
Secondary malignant neoplasm of liver and intrahepatic bile duct	0.053
Secondary malignant neoplasm of bone and bone marrow	0.044
Malignant neoplasm of bladder, unspecified	0.043
Acute lymphoblastic leukemia	0.034
Malignant neoplasm of unspecified part of bronchus or lung	0.031
Malignant neoplasm of rectum	0.029
Malignant neoplasm of prostate	0.026
Secondary malignant neoplasm of lung	0.026
Non-Hodgkin lymphoma, unspecified	0.024
Multiple myeloma	0.023
Malignant neoplasm of ovary	0.022
Myelodysplastic syndrome, unspecified	0.020
Other and unspecified malignant neoplasm skin/ and unsp parts of face	0.019
Acute myeloblastic leukemia	0.018
Malignant neoplasm without specification of site	0.017
Malignant neoplasm of colon, unspecified	0.016
Secondary malignant neoplasm of retroperiton and peritoneum	0.016
Malignant neoplasm of esophagus, unspecified	0.014
Malignant neoplasm of sigmoid colon	0.012
Leiomyoma of uterus, unspecified	0.012
Secondary and unspecified malignant neoplasm of intra-abdominal nodes	0.012
Chronic lymphocytic leukemia of B-cell type	0.011
Unspecific malignant neoplasm of axilla and upper limb nodes	0.010
Malignant neoplasm of pancreas, unspecified	0.010

Table 2.2: 25 most common cancer-related diagnoses in the patient population. ICD-10 codes relating to infection were selected by counting all ICD-10 codes in the chapter on neoplasms. ICD-10 codes were left at their most specific value rather than grouping similar codes under common groups of morbidities. The underlying total of ICD-10 codes pertaining to cancer in the dataset was 366,025.

diagnoses pertaining to cancer in the data between 2000 and 2015. When looking at the relative frequencies of the most common cancer-related diagnoses, these therefore take the top 3 positions (Table 2.2). The most common cancers in the dataset generally reflect the most common cancers in the UK[116].

Although three cancer types are relatively common in the data, no infectious disease is seen is Table 2.1. There were a total of 45,390 diagnoses pertaining to infectious diseases in the data. Unspecified viral infection is the most common with 5,429 cases, at an overall rank of 141 (Table 2.3). The top four most common specific species of infectious disease are *S. aureus*, *E. coli*, *C. difficile*, and *Pseudomonas* species. These common bacterial diseases, along with norovirus, form the basis

Infectious disease	Proportion
Viral infection, unspecified	0.120
Sepsis, unspecified organism	0.096
Staphylococcus aureus as the cause of diseases classified elsewhere	0.078
Escherichia coli as the cause of diseases classified elsewhere	0.075
Enterocolitis due to *Clostridium difficile*	0.063
Viral intestinal infection, unspecified	0.045
Infectious gastroenteritis and colitis, unspecified	0.039
Other viral agents as the cause of diseases classified elsewhere	0.035
Chronic viral hepatitis C	0.033
Pseudomonas (mallei) causing diseases classified elsewhere	0.030
Other bacterial agents as the cause of diseases classified elsewhere	0.027
Candidal stomatitis	0.019
Unspecified staphylococcus as the cause of diseases classified elsewhere	0.013
Sepsis due to other Gram-negative organisms	0.013
Candidiasis of other sites	0.013
Unspecified streptococcus as the cause of diseases classified elsewhere	0.011
Acute hepatitis C	0.011
Sepsis due to *Staphylococcus aureus*	0.010
Varicella without complication	0.010
Zoster without complications	0.010
Staphylococcal infection, unspecified site	0.010
Other streptococcus as the cause of diseases classified elsewhere	0.009
Streptococcus, group B, causing diseases classified elsewhere	0.008
Viral meningitis, unspecified	0.008
Klebsiella pneumoniae as the cause of diseases classified elsewhere	0.007

Table 2.3: 25 most common infectious disease diagnoses in the patient population. ICD-10 codes relating to infection were selected by counting all ICD-10 codes in the chapter on infectious diseases. ICD-10 codes were left at their most specific value rather than grouping similar codes under common groups of morbidities. The underlying total of ICD-10 codes pertaining to infection in the dataset was 45,390.

of my work in chapters 4 and 5.

2.4.3 Stay characteristics

The patterns of time patients spend in the hospital is important, as it relates to how patients can overlap with one another. For example, the potential for overlap in in-patient wards is very different from out-patient wards. To examine this, I looked at the general stay and spell properties of patients. To reiterate, a stay is defined as the interval between when a patient enters and leaves the health care system. A spell is likewise the interval between a patient entering and leaving one hospital *ward*.

The mean stay length was 86.43, with a standard deviation of 5,427.71 hours. This large deviation was primarily due to the vastly different stay lengths implied by inpatient vs outpatient visits (Figure 2.6). Additionally, there was heaping observed at 24, 48, 72, and 96 hours, transitioning into a smooth curve after four days of stay length. The trough at ~15 hours likely reflects patients not being discharged overnight (e.g. a patient entering at noon would not be discharged between 6PM and 8AM, or 6-22 hours after entry). This indicates that there is underlying hospital procedure that makes certain stay length more common than others. At a minimum, this is likely partially driven by the working hours of those responsible for discharging patients, as patients can only be discharged when at least one such person is active. For out-patient wards, which generally are not open at night, stays can be no longer than the full opening hours of the ward.

Looking at spell lengths, I observe a similar pattern (Figure 2.7). This is largely because 56.81% of stays are comprised of a single spell. The mean spell time was 64.03 hours with a standard deviation of 214.51 hours. I see similar, but not identical, heaping at the times as observed in Figure 2.6. The times are slightly lower for spells rather than stays, since each stay consists of a minimum of one spell, but often more than that.

2.4.4 Ward information

Understanding not only how patients spend time in the health care system, but also how they traverse the system is important, because the transfers a patient experiences (or doesn't) may have implications for their health and eventual outcomes[117,118,119]. This can be partly understood through the characteristics of the wards themselves. This can be further understood via the ward transfer network, which I describe in the following section.

Ward transfers are limited by the size of the ward. The distribution of the occupancy of all 145 wards is shown below (Figure 2.8). I calculated this from the data by counting the maximum number of patients observed in the ward at any one time. The size of each ward sets a limit on how many patients can be

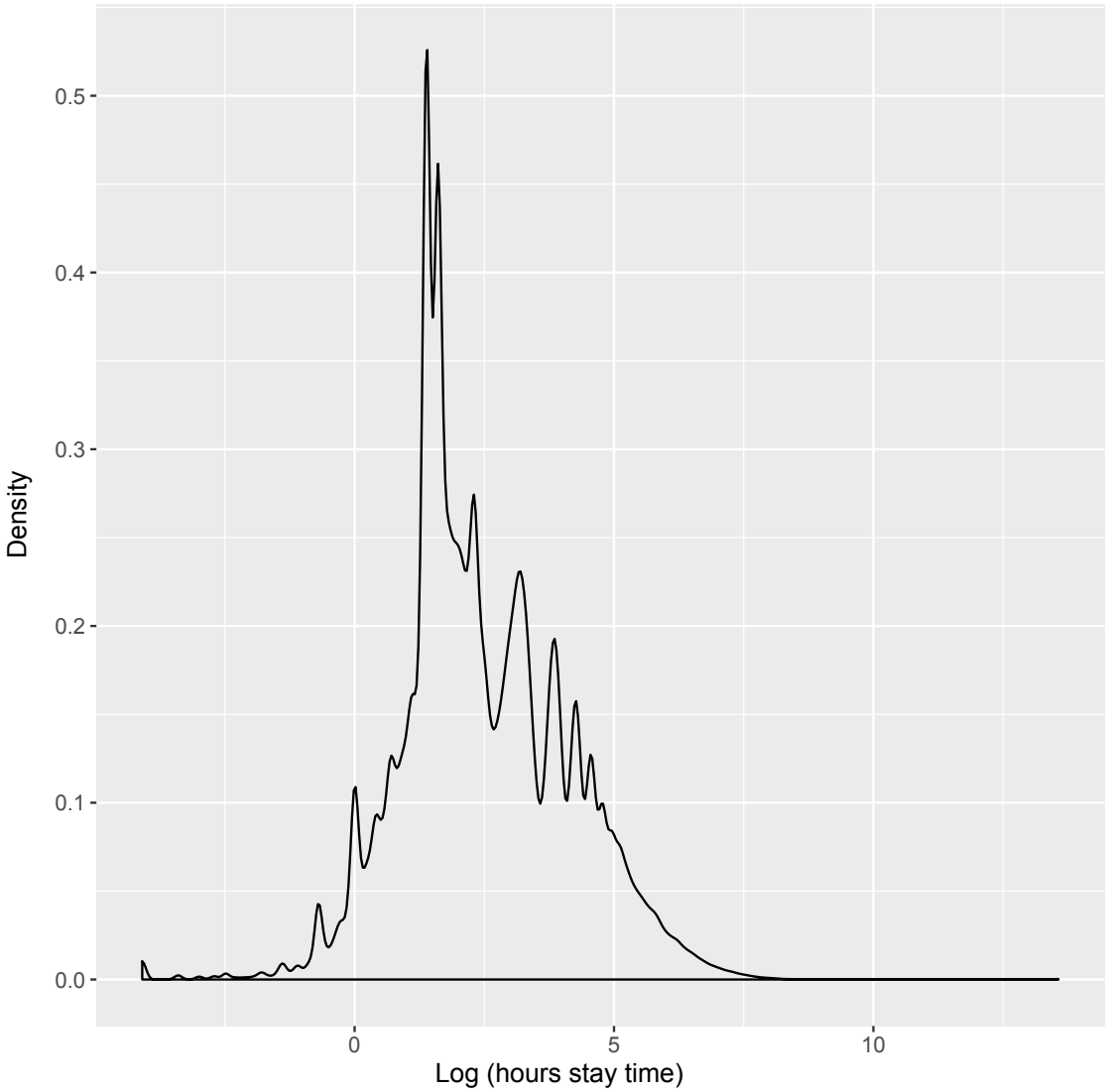

Figure 2.6: Distribution of hospital stay length times. Hospital stays are defined as the time between when a patient enters and leaves the health care system. Maximum peak occurs at 4.5 hours. Other peaks exist at 0.5, 1, 12, 24, 48, and 72 hours.

co-present with one another at any given time in a given ward. In certain situations, this needs to be controlled for, as co-presence in a ward with five spaces may be different than co-presence in a ward with 100.

Also important is the layout of each ward. Wards may be open, partially open, or composed of individual rooms. Co-presence in these cases may also be qualitatively different. From 2011 onwards, ward information includes more specificity, sometimes down to the bed. This allows for a more precise understanding of patient location within wards, which may relate to the social and biological processes I studied here.

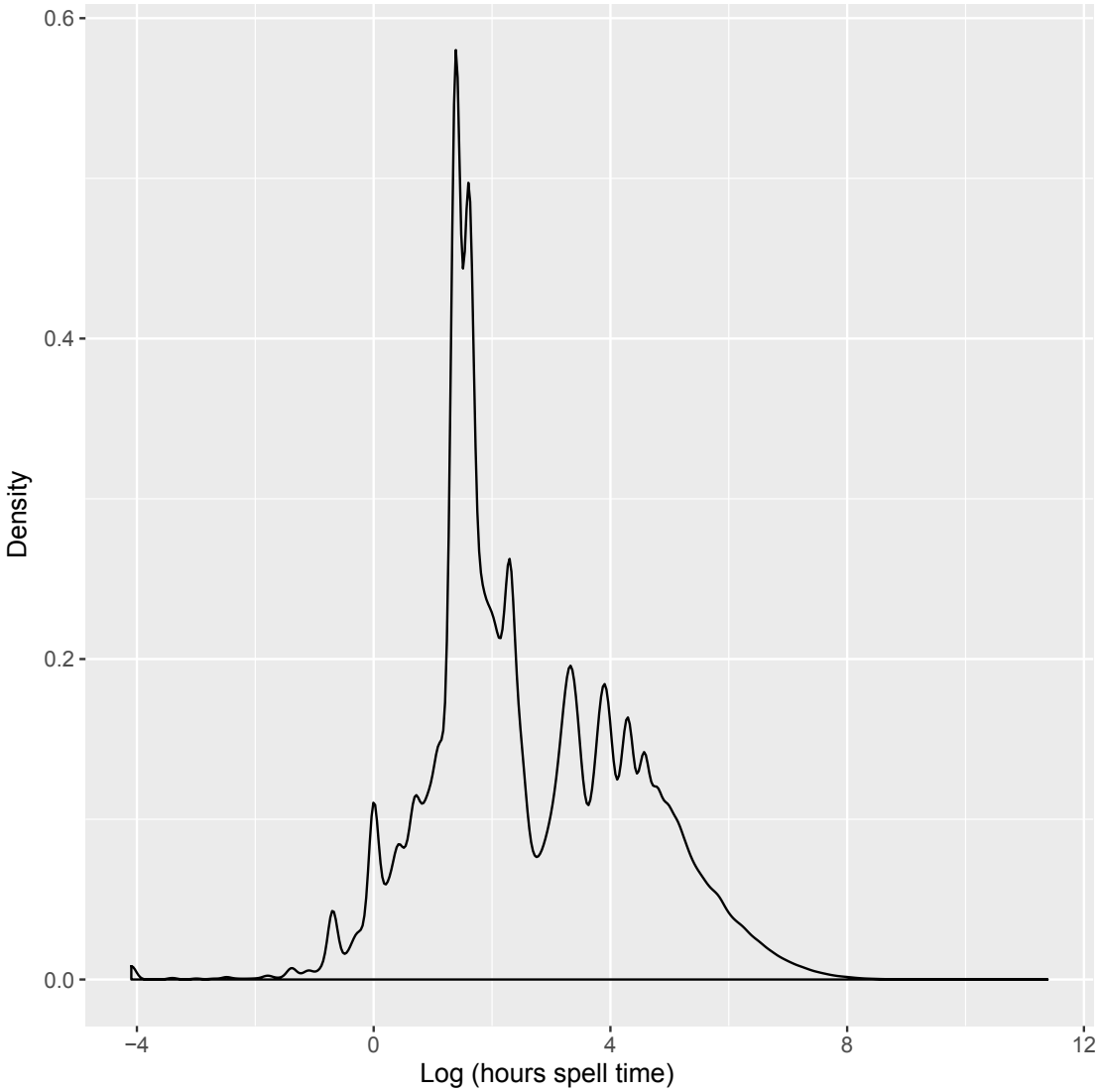

Figure 2.7: Distribution of hospital spell length times. Hospital spells are defined as the time between when a patient enters and leaves a single hospital ward. As such, each hospital stay comprises one or more hospital spells. Peaks exist at 0.5, 1, 3.5, 11, 20, and 44 hours.

Finally, some wards are in-patient, meaning patients check into the ward and stay for longer periods of time. Other wards are out-patient meaning a patient enters the ward and leaves the same day, often after a specific procedure is performed. 90 wards were in-patient and 55 wards were out-patient. I determined this by noting which wards had zero patients in them overnight, as out-patient wards operate on regular business hours and close overnight.

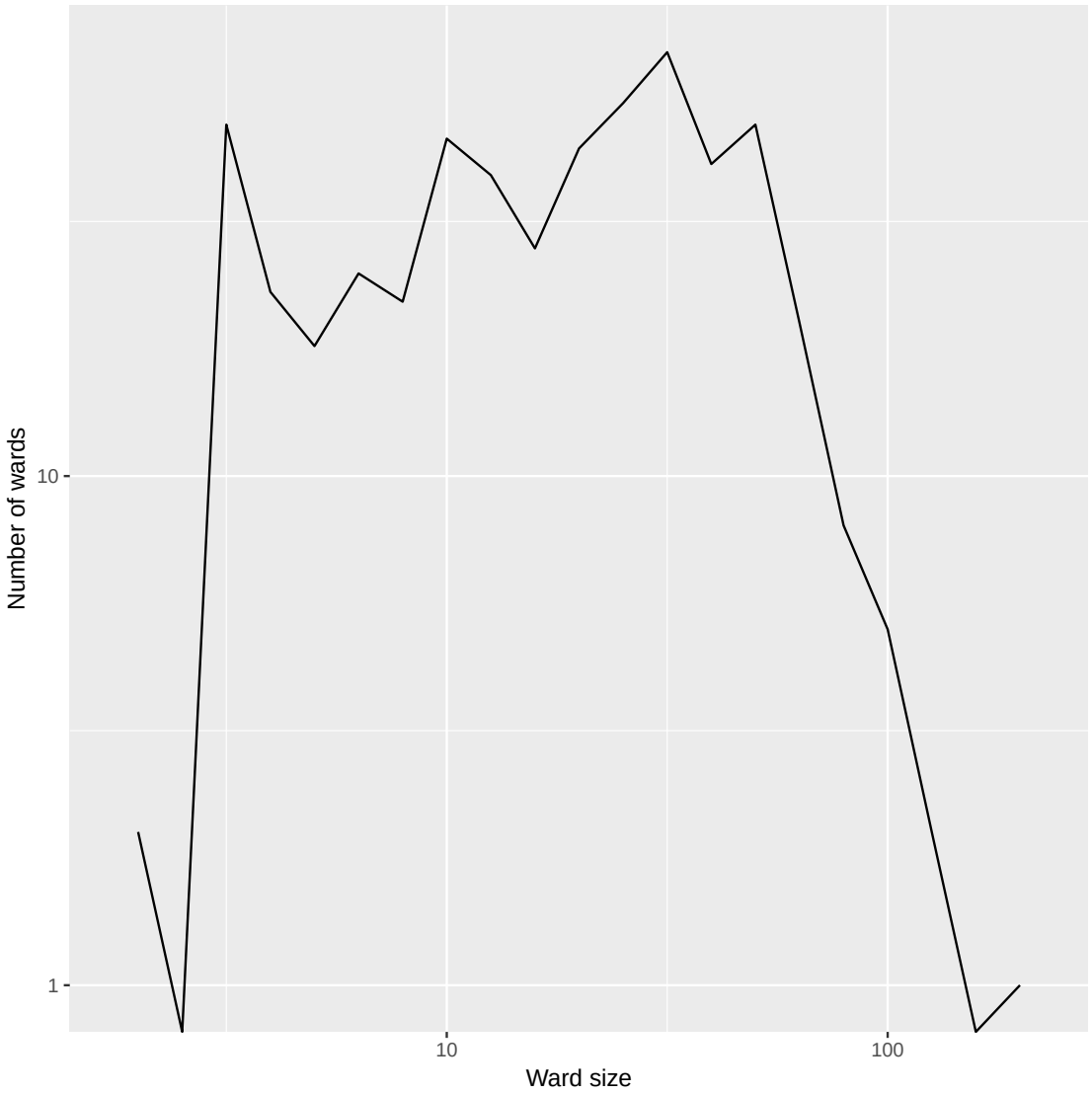

Figure 2.8: Distribution of maximum ward occupancy in 2010. Maximum occupancies were determined algorithmically by observing the number of patients concurrent in the ward at any given time, and taking the maximum number over the course of 2010. This method assumes each ward was at capacity at some point in 2010.

2.5 Networks from administrative data

One of the goals of this thesis is to show the ways in which administrative data can be used above and beyond its original vision of monitoring patients. One of these is the interconnectedness between observations. Entries in an electronic medical record are designed to connect all of a patient's stays and medical tests together such that all the relevant information on a patient can be accessed. However, they are rarely used to relate patients to one another, despite the many ways to

do so based on the richness of the data collected. Specifically, the data therein has many interconnected data points which can be used to construct networks. These networks typically take the form of bipartite networks: those where two types of nodes exist, and there are connections between different types of nodes, but not between nodes of the same type[120].[2]

The main network I will use for my analyses is the patient-patient co-presence network, which is the unipartite projection of the patient-ward bipartite network. From this network, the patients who are together in a ward or a ward bay at the same time are linked together. However, there are a number of other important networks that can be obtained from the data. For example, diseases can cluster in patients, patients can move from ward to ward, and doctors can share patients. In this section, I survey the types of relational and network data that can be constructed from the IORD. Although what I show below is specific to the IORD, much of it can be generally applied to electronic medical records and hospital administrative data. In addition to constructing the networks, I also reveal some basic insights that can be learned from this approach.

The networks I use are summarized in Table 2.4. These networks are the co-presence network derived from the patient-ward and the patient-ward bay bipartite networks, the disease network derived from the patient-diagnosis bipartite network, the physician network derived from the patient-physician bipartite network, and the ward-transfer network derived from the patient-ward bipartite network but using the alternative unipartite projection of the co-presence network. Importantly, all of these types of networks have been used in the literature to make important conclusions about health care and human health. The table also includes some basic network statistics for each network[53].

[2]For a brief discussion of bipartite networks, see Chapter 1.

Names	Node1	Number1	Node2	Number2	Density	Transitivity
Co-presence	Patients	3,004	Wards	132	0.007	0.674
Co-presence	Patients	3,004	Ward bays	205	0.003	0.739
Co-presence	In-patients	1,107	Ward bays	54	0.006	0.632
Co-presence	Out-patients	1,897	Ward bays	140	0.004	0.958
Ward transfer	Wards	127	Patients	17,076	0.044	0.336
Disease	Diagnoses	3,221	Patients	21,901	0.01	0.201
Physician	Physicians	385	Patients	21,901	0.007	0.183

Table 2.4: Five unipartite networks created from the electronic medical records and administrative data. All four networks are a unipartitite projection of a bipartite network constructed from the data. For all networks, “Node 1" is the type of node that is kept when the unipartite projection is made. “Node 2” is then the node that is removed, and common connection to that type of node forms the basis between nodes of the first type.

2.5.1 Co-presence network

Importantly, there are few cases of the ward co-presence network in the literature, and it therefore represents one of the major novelties of this work. Some studies examining the trajectory of infections in a nosocomial outbreak implicitly use the ward co-presence network when examining which patients were at risk of infection[121]. Some studies have examined co-presence as measured by devices sensing one another[67,75]. Other studies use a hospital ward as a focal point or inclusion criterion for entry into the study, but do not examine any of the patterns between patients once in the ward. However, no study to my knowledge has yet used EMR and HAD specifically to create a co-presence network.

First, as it is the most central to the thesis, I show a general co-presence network taken from just one month of the data, January 2010 (Figure 2.9). I briefly discuss some general points of this network here; for a more thorough discussion of these types of networks, see Chapters 3, 4, and 5. Importantly, I create this network based on both ward co-presence, and ward bay co-presence. A ward bay is a subunit of a ward, and allows for more meaningful co-presence information. As stated in Chapter 1, co-presence largely serves as a proxy for social interaction or for the transmission of biological vectors. Although some wards are a single, open room where patients can interact (e.g. the chemotherapy ward), some wards contain multiple individual rooms. In these cases, co-presence based on just the ward may not actually indicate that patients were in view of one another. Ward bays make this assumption more plausible. Based on this, there is a clear difference in network density (Table 2.4), as the ward bay removes may edges that are present in the ward-level co-presence network. As can be seen when comparing Figure 2.9 A and B, the more than double density of B obscures the ability of one to observe meaningful edges with the naked eye, whereas they are observable in the ward bay co-presence network.

There is also clustering based on age. Nodes of similar sizes are more likely to be connected to one-another than to patients of disparate ages. This is because many types of morbidities are specific to life-stage, and wards often contain patients with

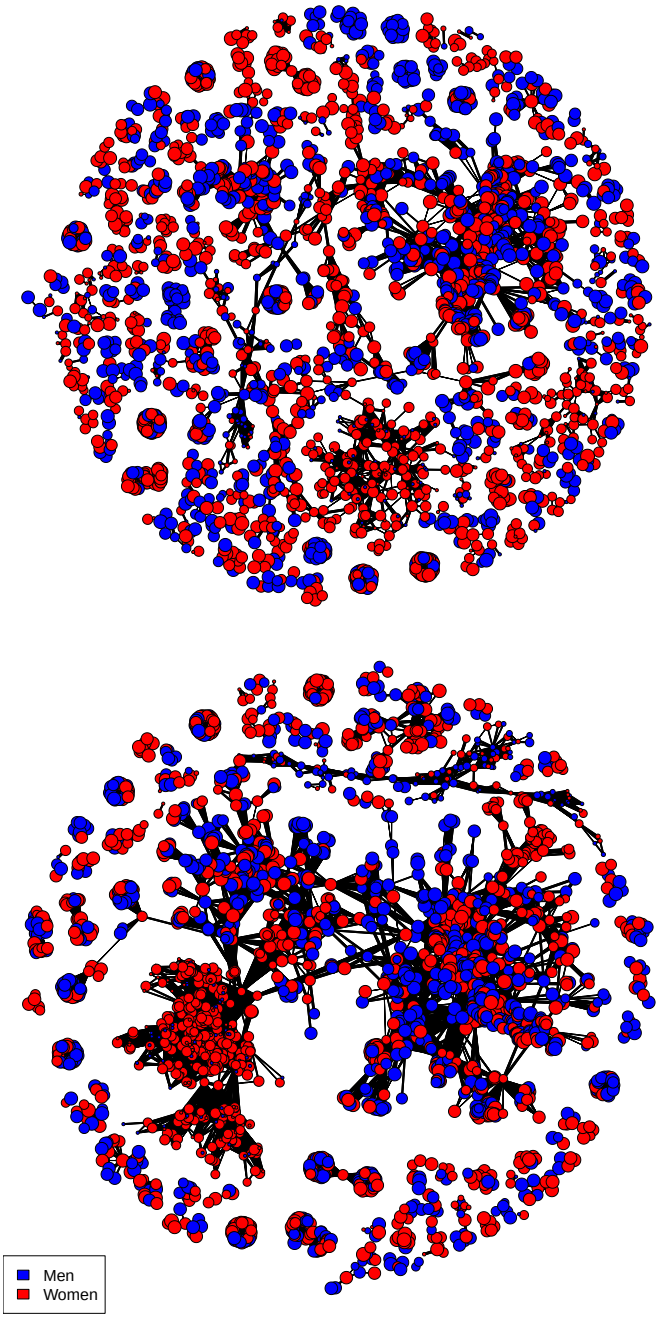

Figure 2.9: Patient-patient co-presence network in 2010 using a Fruchterman-Reingold layout. Nodes represent patients, and edges between nodes indicate that the two patients were co-present in A) a ward bay or B) a ward, for at least one hour. Edge opacity is proportional to how much time patients were co-present. Node size is proportional to how long the patient was in a hospital over the year. Node color is based on the most coarse grouping of ICD-10 codes representing the patient's first primary diagnosis during the year.

similar morbidities. For example, obstetric wards contain women of child-bearing age. This, combined with sex segregation in some ward bays, also explains the sex homophily observed in the network.

The transitivity for both of these networks (as well as the latter networks) is relatively high. This is largely a result of the unipartite projection of the bipartite network[120]. Because any set of n patients simultaneously in a ward automatically forms an n-clique (a subgraph where all n nodes are connected to all other $n-1$ nodes) when the unipartite graph is created, this high transitivity and creation of triangles inflates both of these terms. Here, using the ward bay instead of the ward only lowers the transitivity slightly, despite the much-reduced density relative to using the ward level.

I also separate the co-presence networks of in-patients (Figure 2.10) and out-patients (Figure 2.11). An in-patient is defined as any patient who was in the hospital for at least one night, and an out-patient if not. This distinction is important because these wards are separate, and patients in them are subject to different protocols and exposures. In later chapters, I often subset to only either out-patient (Chapter 3) or in-patient (Chapters 4 and 5) wards. In-patient status was defined solely from the HAD. Importantly, the number of ward bays in the two networks do not sum to the number of ward bays (Table 2.4) in the total co-presence network because some in-patients begin in an out-patient ward, but based on their diagnosis may be rapidly moved to an in-patient ward.

The in-patient network (Figure 2.10) shows patients who remained in the hospital for at least one night. This constituted 36.8% of the patients in the system over the period examined for the figures, and this stays relatively constant over the entire period of the data. One can observe many distinct ward bays in this network based on the clusters of patients who are tightly connected to one-another, but have no connections to other patients. Despite these highly-clustered wards, the network as a whole is more dense but less cohesive than the overall co-presence network. This is likely because other in-patients are transferred between wards

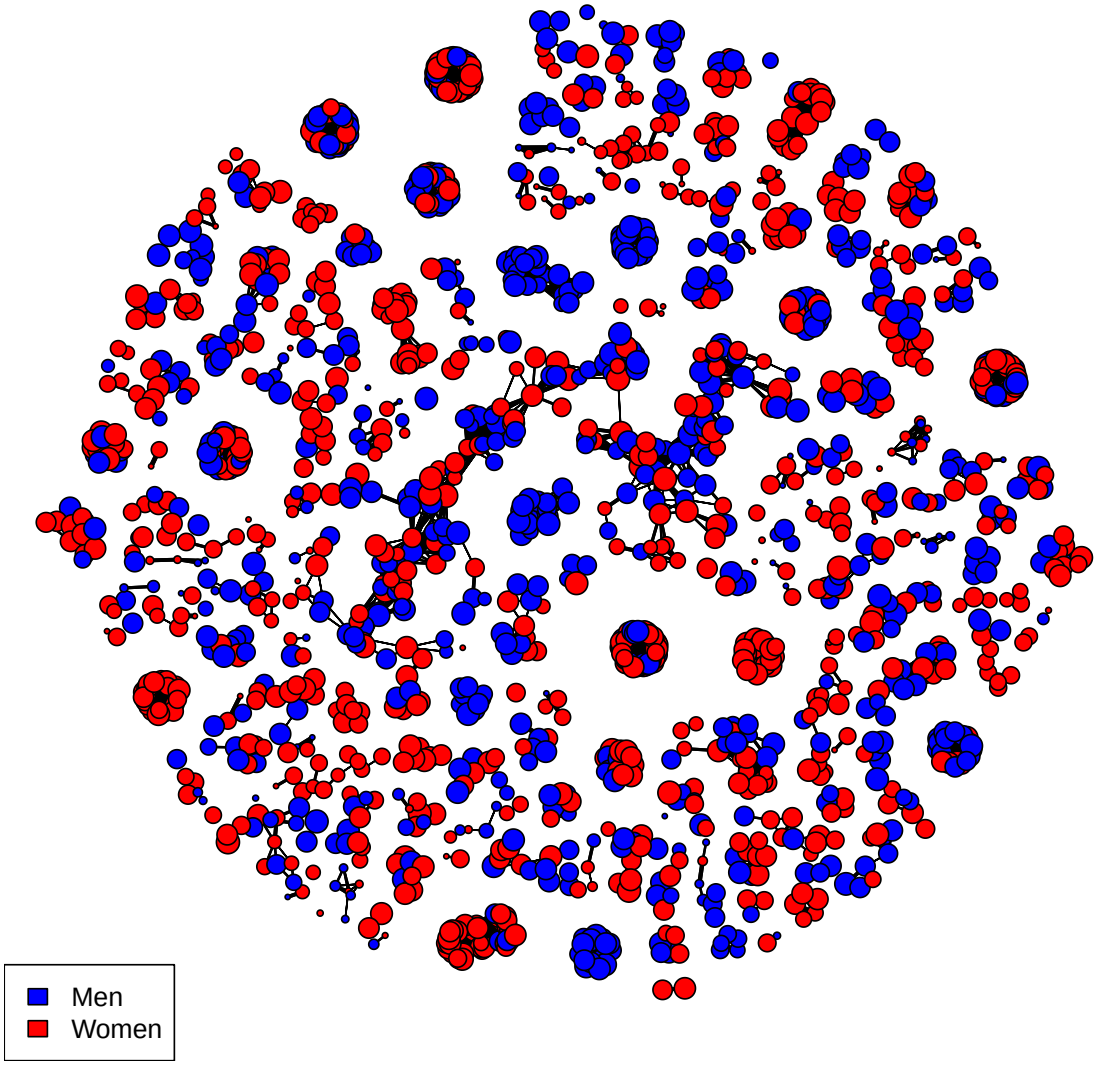

Figure 2.10: Patient-patient co-presence network in 2010 for in-patients only. Nodes represent patients, and edges between nodes indicate that the two patients were co-present in a ward bay for at least one hour. Edge width is proportional to how much time patients were co-present. Node size is proportional to how long the patient was in a hospital over the year. Node color is based on patient sex, and node size is proportional to patient age.

to receive different procedures, or as their condition changes, which can lead to connections between patients that are not transitive.

The out-patient network (Figure 2.11) shows patients who were in the hospital for a maximum of one working day (63.2% of patients). Interestingly, one can immediately distinguish the maternity ward from this network, as it shows high age heterogeneity via the large and small dots, and by the sex homogeneity of large

nodes indicating the delivering mothers. This goes to show the potential benefits of visualizing these types of networks before even considering the quantitative aspects of the network. Although the density of this network was slightly higher than the overall network, the transitivity was extremely high, at 0.96. This is largely because out-patients do not experience many ward transfers, they enter the hospital into a single ward and leave from that same ward. Therefore the only way transitivity does not occur is if two other patients overlap with the first patient, but not with one-another. If one assumes patients enter the ward at a uniform rate, and are there for equal amounts of time, a set of three patients would have a 78% chance of being transitive[3]. However, when one factors in the beginning and the end of the work-day, this would be increased substantially, particularly as the average stay length increases. This would also be affected by the fact that HAD is not precisely continuous due to the reasons described previously.

2.5.2 Ward transfer network

Also important to understanding co-presence is the propensity for patients to move between wards. A ward transfer is an important event in a patient's health care trajectory, and often represents either a serious complication, or overcrowding[122]. Because of this, the ward transfer network reflects both information regarding the hospital's policies and external pressures such as an increase in disease rates and information regarding individual health status. The ward transfer network comprises wards which are connected if a patient is transfered directly from one ward to another. If a hospital is envisaged as a hierarchical pyramid with the hospital as the top-level, wards as a middle level, and patients ensconced within wards at the lower level, this information therefore represents both top-down and bottom-up influences. These can be difficult to tease apart at the level of the static network.

Intra-hospital ward transfer networks have not often been used in research. Rather, much research focuses on the more macro-level network of inter-hospital

[3]Each patient is $p_i \sim U(0,1)$, with a stay length of 0.33. If one conditions on the first patient (p_1) being in the ward from (0.33,0.67) and on both other patients being co-present with p_1, then all three patients are transitively co-present *iff* $|p_3 - p_2| < 0.33$. The difference of two uniformly-distributed random variables takes a triangular distribution. Therefore, $P(|p_3 - p_2| < 0.33) = 0.78$.

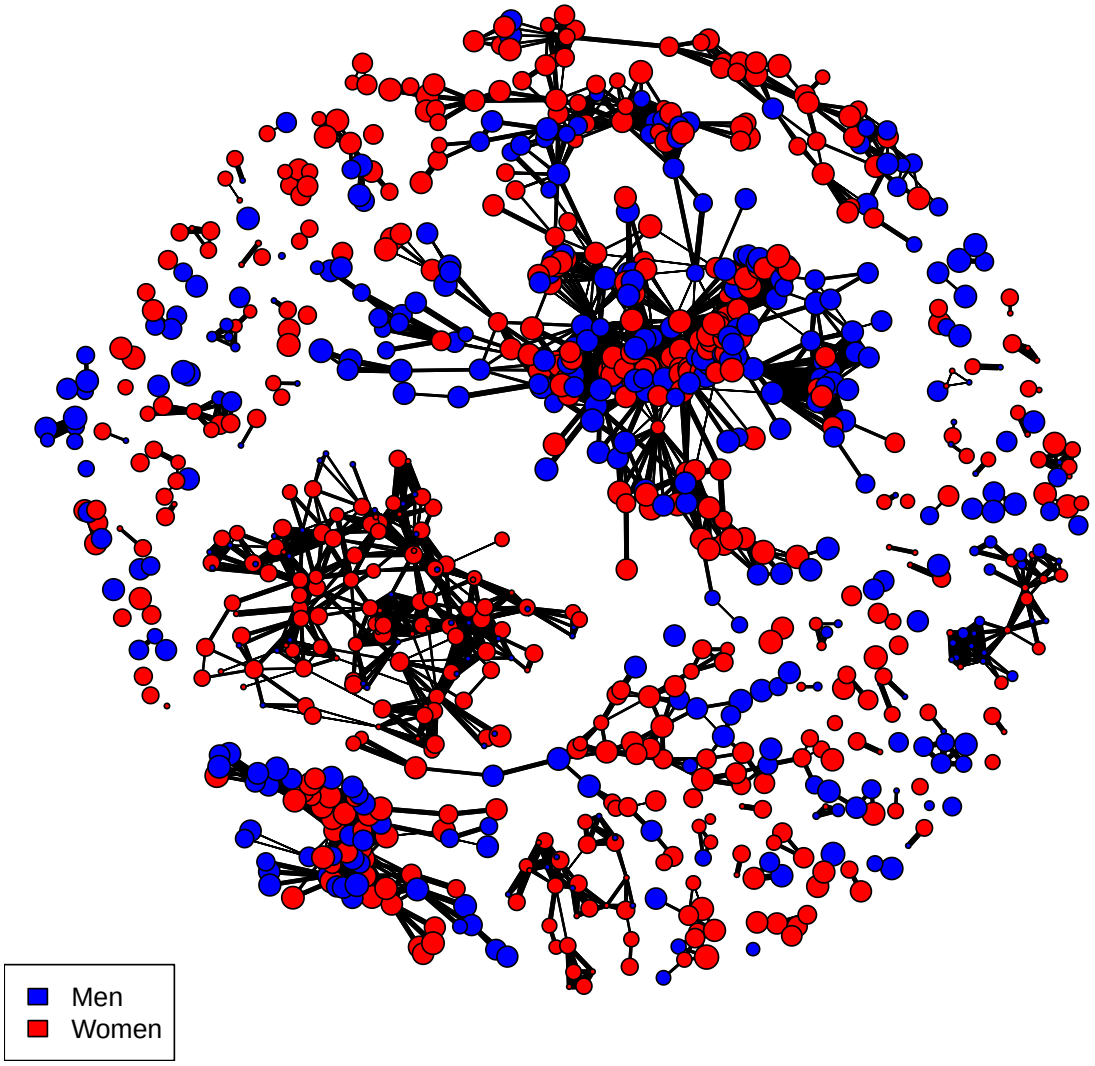

Figure 2.11: Patient-patient co-presence network in 2010 for out-patients only. Nodes represent patients, and edges between nodes indicate that the two patients were co-present in a ward bay for at least one hour. Edge width is proportional to how much time patients were co-present. Node size is proportional to how long the patient was in a hospital over the year. Node color is based on patient sex, and node size is proportional to patient age.

transfers. These networks often give key insights about patient referral mechanisms, and the organization of a particular network of hospitals (e.g. hub-based or regional-based)[123,124]. The lack of ward transfer networks in the literature is likely because each hospital has its own policies about ward transfers, and these decisions can also be influenced by insurance companies, particularly in the US. This makes separating noise from signal in US ward transfer networks difficult. However, in the UK this is less of an issue, as the systems across the NHS are relatively well-integrated.

I show the ward transfer network in Figure 2.12, with nodes colored according to in-patient/out-patient or hospital. In both cases, I observe significant clustering between wards of similar types. From this, one can see that wards do not equally send patients to all other wards; there are patterns in how patients move from ward to ward. For instance, the Accident & Emergency ward (A&E), sends many more patients out than it receives. Additionally, 16 wards neither send nor receive patients to or from other wards during the year.

2.5.3 Disease network

The disease network connects ICD-10 codes when they co-occur in a patient. ICD-10 codes are a strong proxy for diagnoses and underlying morbidity[102]. Goh et al.[125] created one of the first uses of a disease network in the literature by searching curated databases for genes which were mutually implicated in a single disease, thereby linking the two diseases. In doing so, the authors were able to discover clusters of diseases with related genetic causes.

The main downside of the above paper was the time needed to create the network. Databases had to be manually searched and vetted for each edge. Using hospital administrative data allows for the clustering of morbidities without additional curation. I show the disease network based on the hospital administrative data below (Figure 2.13). Due to the sheer number of ICD-10 codes, not much can be determined visually. For comparison, the main network in Goh et al.[125] included a few hundred nodes. To reduce a similar network constructed from ICD-10 codes down to a few hundred nodes, many related ICD-10 codes would need to be grouped

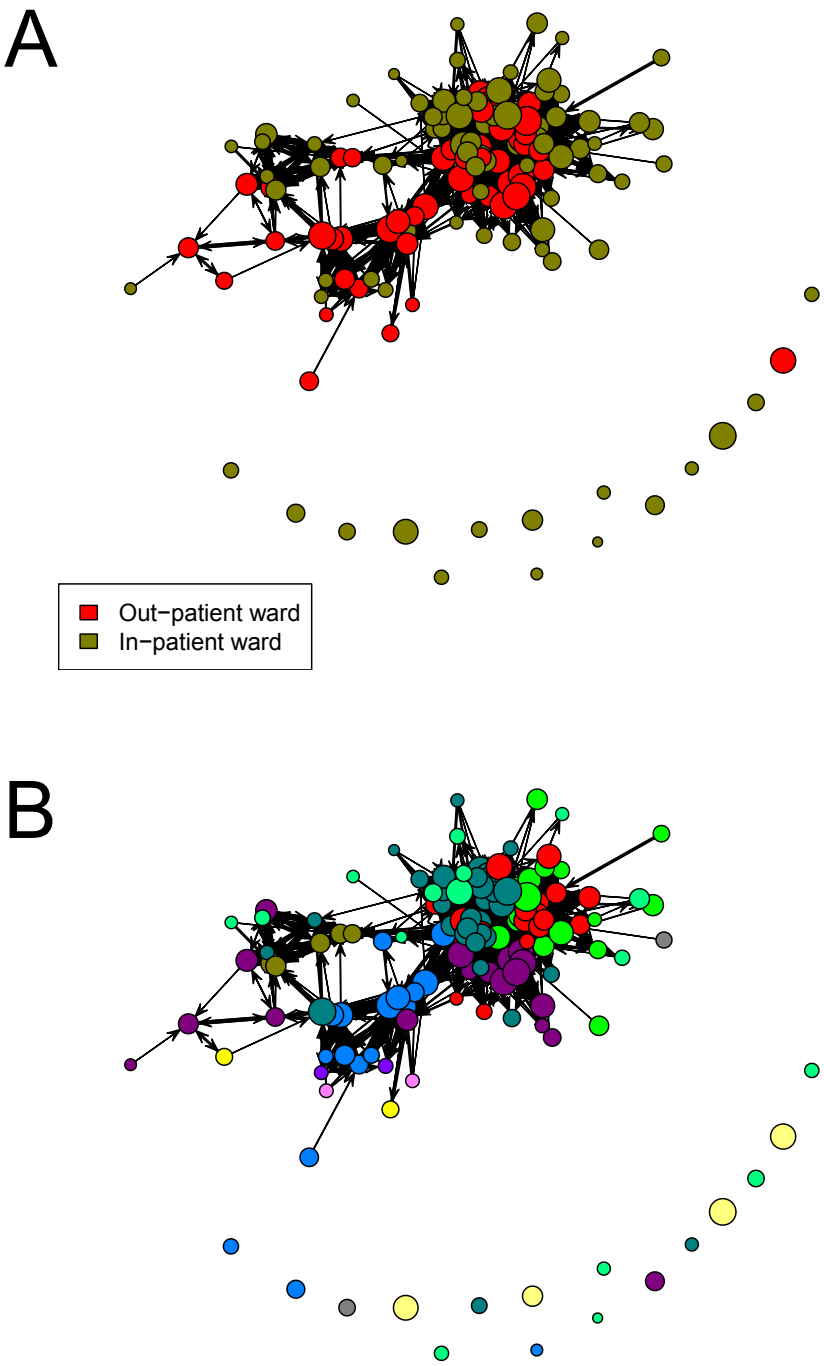

Figure 2.12: Inter-ward transfer network of the IORD in 2010. Each node represents a ward, and each edge represents patient flow from one node to another. Nodes sizes are proportional to the number of beds in the ward. Edge widths are proportional to the logarithm of the number of patients going from one node to another. In A), nodes are colored based on whether they are inpatient or outpatient wards. In B), nodes are colored based on the hospital in which they are located. Hospitals are not labeled due to potential lack of anonymization that might result.

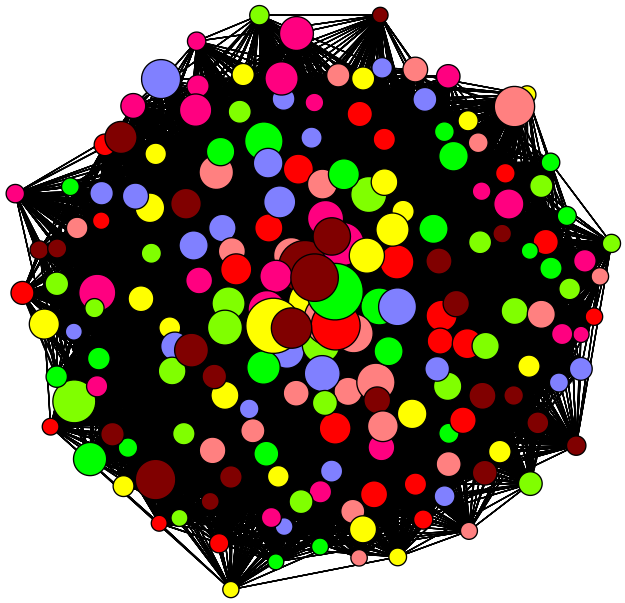

Figure 2.13: The patient disease network 60-core in 2010. Nodes represent individual ICD-10 codes, serving as a proxy for diagnoses (n=188). The network was reduced in size form the full 3,221 nodes because nothing could be gleaned visually from the full network. Nodes are colored according to the highest-level indicator of the ICD-10 code, often referring to the broad system affected, or type of morbidity. Nodes are sized in proportion to the number of patients receiving the code. Edges connect nodes when at least one patient has both diagnoses, with edge width proportional to the number of patients sharing those two diagnoses. The nodes for "End-stage renal disease" and "Preparation for dialysis" had their sizes capped, as they appeared more than an order of magnitude more often than the next most common ICD10 code.

together. To be directly comparable, the ICD-10 codes would have to be manually recategorized, as the tiers of specificity within the ICD-10 coding system do not directly correspond to those used by Goh and colleagues.

For the observed network, the density is relatively low, and the transitivity is relatively low, especially for a unipartite projection from a bipartite network. This is largely a result of hospital procedure, in that most patients only have a primary and potentially a secondary diagnosis; transitive closure would most often only

occur when a patient had at least three diagnoses [4].

In the network visualization, 60% of the nodes are of one of three ICD-10 categories: circulatory system diseases, abnormal laboratory findings, and factors influencing health/contact with health services. This indicates that circulatory system diseases stem from a multifaceted etiology (factors influencing health), and have a wide variety of lab factors associated with them. Additionally, the same circulatory system diseases are often co-morbid with one another. This kind of visualization can give some insight into constellations of diseases that co-occur, and how they interact with both the health care system, and how their etiology is notated in EMR and HAD.

Importantly, the NHS data here is inclusive of the connections from the human disease network based on shared genetic causation. However, it also includes connections between diseases that arise from a common environmental cause rather than genetics. Therefore, this network constructed from administrative data, could, in theory, be used in conjunction with the above work to separate out the genetic and environmental causes of disease.

2.5.4 Physician network

Finally, the physician network represents which physicians within a hospital share patients. This is important for a number of reasons. Within the context of this thesis, it is important because physicians can often serve as a method of transmission for infectious vectors[126]. Therefore, even when patients are not connected based on ward co-presence, they may be connected by their physicians. Although this is not something I explore in this thesis, it is something that may affect the results. However, for reasons I explain in Chapters 4 and 5, I do not think this is likely. In brief, many physicians see patients only in a single ward, so the co-presence network is inclusive of ties in the physician network.

[4] Although it is also possible for three different patients to form a complete triad, this requires three patients to each have one pair from the triad, and none of them to also have the third diagnosis. This is unlikely.

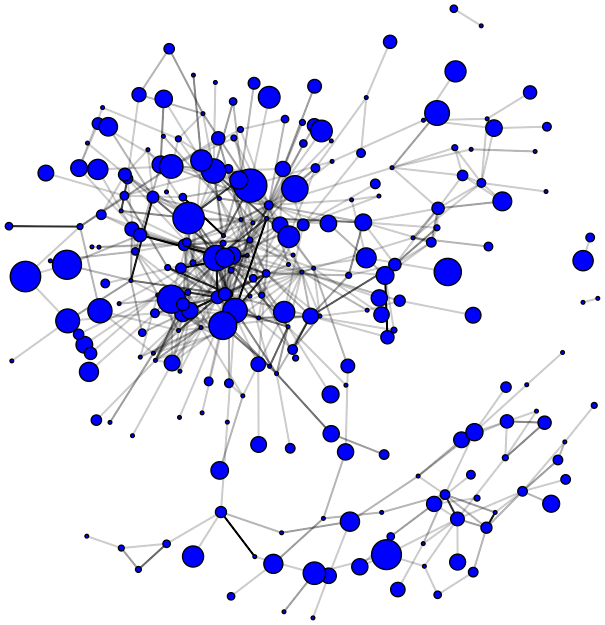

Figure 2.14: The patient physician network in 2010. Nodes are admitting physicians in the data (N=263), with node size proportional to the number of patients. Edges are formed when two physicians share at least one patient, with edge width proportional to the number of shared patients.

I show the physician network in Figure 2.14. Although each individual patient stay only has a single primary physician, a patient entering the hospital for multiple stays over a period of time could have different physicians, thereby connecting the physicians through shared patients. However, only 1,054 patients have multiple physicians within a year, and so the network is relatively sparse. The transitivity is even lower for this network than from the disease network, and largely for the same reasons. As it is rare for a patient to have more than one physician, it is even rarer for them to have more than two physicians in a short period of time (n=64). Therefore, transitivity between physicians is again relatively unlikely.

Unlike the diagnosis network, there are few enough physicians that structure can be observed in the network. There is a largest connected component which

includes 239 of the physicians. Over 100 physicians do not share patients with any other physicians. The total number of patients a physician has is highly correlated to whether or not a physician has any connections ($p<0.001$). As a physician has more patients, their likelihood of sharing patients with other physicians increases (OR=1.02 per patient). This is important, as it must be considered when controlling for patients sharing a physician; the number of patients seen by a given physician should be added as a covariate to adequately adjust for this. This is an additional example of an insight that can be gained from examining a specific type of network from HAD that can have an impact on other analyses also using EMR and HAD.

2.6 Study populations

For each study comprising this thesis, I use a subset of the overall population to answer specific research questions. The rationale for the particular subset of patients is described in detail in each corresponding chapter. Below, I show a general schematic for the population for each study (Figure 2.7).

In the first empirical study, I examine social influence in a chemotherapy ward (Chapter 3). This study shows proof-of-concept of using co-presence on a restricted dataset (only 4,700 of the 780k+ patients), and takes place in an environment where co-presence is likely to be most impactful.

The next two empirical studies (Chapters 4 & 5) use the same initial population of 245,709 patient in the dataset from 2011 to 2015. However, they differentially use subsets of this population based on different exclusion criteria (Chapter 5 additionally excludes patients with no biomarker data). In these studies, I examine the utility of co-presence as a predictor of infection (Chapter 4), and in concert with biomarkers of subclinical infection (Chapter 5).

In addition to the empirical work, I also make methodological advancements both within the empirical studies (Chapter 3 where I devise a method to detect consistent co-presence) and in the appendices (Appendix A where I describe a computational method for the colored triad census). I apply this latter method in

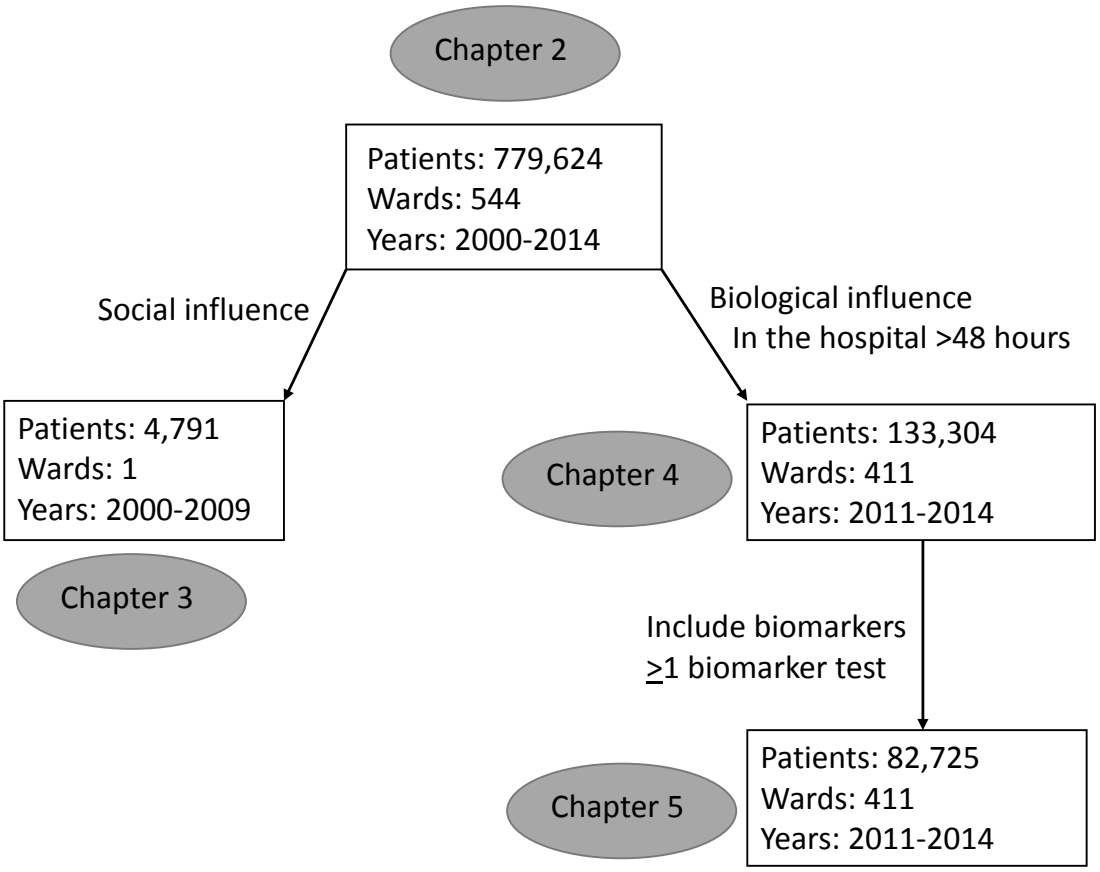

Figure 2.15: Study diagram indicating the subsample of patients in each of the following studies comprising the thesis. Each empirical chapter uses a different subset of patients based on the time, ward, and other inclusion/exclusion criteria. Year-ranges are inclusive.

Chapter 5 to understand the network characteristics of how subclinical patients impact nosocomial spread.

3

Social influence on 5-year survival in a longitudinal chemotherapy ward co-presence network

Contents

3.1 Abstract

Chemotherapy is often administered in openly-designed hospital wards, where there is the possibility of social influence on health between patients. Previous research has found evidence that cancer patients' health is impacted by social relationships; however, social influence has not been examined between patients receiving treatment in a chemotherapy ward. In the current paper, I investigate the influence of co-presence on five-year survival in a chemotherapy ward. Using data on 4,691 cancer patients undergoing chemotherapy in Oxfordshire, UK, I construct a network of patients where edges between patients are based on whether they are co-present more often than expected by chance. Patients averaged 59.8 years of age, and 44% were male. I count the total edges to focal patients' immediate neighbors or those two nodes away who finish their chemotherapy cycle and survive 5 years or die within 5 years. Generalized Estimating Equations were used to evaluate the effect of neighbors' outcomes on focal patient's 5-year mortality. Being consistently co-present with no other patients increased one's odds of death by 58.9% (CI: 36.6%,84.8%). Each additional edge to a patient dying within 5 years increases a patient's mortality odds by 6.6% (CI: 2.9%,10.4%). Each edge to a patient surviving 5 years reduces a patient's odds of dying by 10.2% (CI: 14.9%,5.4%). The results suggest that social influence occurs in chemotherapy wards, which may need to be taken into account in chemotherapy delivery.

3.2 Introduction

Cancer is a leading cause of death in the United Kingdom (UK), with one in four people dying of cancer (UK, 2014). Cancer patient outcomes, particularly the gold-standard 5-year survival, have been robustly associated with a number of individual characteristics such as treatment protocol, age, sex, and cancer severity[127,128,129]. A patient's social sphere may also impact patient outcomes: there is evidence, for example, that social ties (e.g. increased social network size and interactions) are associated with both reduced all-cause mortality and cancer-specific mortality,

and that this relationship may be stronger for women[130,131,132]. However, there is limited research considering the effect of the social context of cancer treatment itself on patient survival. Indeed, chemotherapy—one of the most common forms of treatment—is often administered in out-patient group settings, representing one important social context for further inquiry. To this end, I investigate the impact of network members' co-presence, survivorship, and death on individual chemotherapy patients' survival.

3.2.1 Stress-mediated effect of social influence on health

Social influence may be an important interpersonal mechanism impacting health outcomes of cancer patients receiving treatment in the chemotherapy ward through patients' stress response. Patients entering chemotherapy generally have three concomitant threats to their health: the physical disease of the cancer, the neuro-endocrine stress based on uncertainty related to the course of the physical disease, and the cellular response to chemotherapy[133,134]. Although the cancer itself is the major cause of mortality, the effects of stress on health are also important; reduced stress can significantly reduce 5-year mortality in chemotherapy patients[135]. Therefore, if stress is altered by some social influence process, then such processes can impact overall health and survival of cancer patients. A spectrum of social influence mechanisms exists, each distinguished by variation in the type of interaction.

Each has different implications for how social influence may impact the patient's stress response. The strongest effects of social influence are in the context of direct interactions with close social ties[136,131]. This work generally considers the exchange of social resources, including informational, emotional, and instrumental support[137,138]. There are a number of robust associations between the structure and function of patients' social support networks, patients' stress response, and ultimately their health outcomes. These include the strength of ties[11,139], received social support[140], and perceived social support[139]. Such support resources have been shown to influence how patients with a health stressor manage that stress[141]. For example, Croyle and Hunt[142] showed that patients experienced different responses

to a medical test result based on whether or not they interacted with a confederate receiving the same test result; those receiving similar results experienced lower stress levels than those that differed. However, not all social interactions result in positive outcomes. For instance, patients can also support one another via an exchange of information[143], but if the information is poor, one's perception of the sharer's knowledge is negative, or the relationship between network members itself is negative, the health outcomes of this interaction can be deleterious[144,145].

Social influence may also occur in the absence of direct interaction. A patient's physical appearance, for example, may convey information about their true health status which other patients observe; such information can impact, in turn, the observer's health. Previous research has shown that the mere presence of others engaged in the same task impacts physiological arousal and performance, in what is known as social facilitation[20]. Also, individuals will alter their behavior based on the behavior of those around them, in what is known as modeling[56]. While there is not a specific behavior or task in chemotherapy, one still observes others responding well or poorly to treatment over time. These observed responses could then effect arousal or stress pathways, resulting in indirect influence dependent on the outcomes of the observed patients. Both social facilitation and modeling can therefore result in differences in behavior and stress response that ultimately impact health.

Finally, it is important to note that these social processes may result in differential stress responses based on whether the relationship in question is with a stranger or a familiar individual. For example, one study found that people had slower latent cognitive reaction times based on fMRI when presented with a familiar face when compared to an unfamiliar face. Such responses likely indicate increased arousal for participants when observing familiar faces[136], a result previously found in animal models[146]. Thus, observing familiar people elicits different reactions than observing strangers. In the setting of chemotherapy, patients almost always begin as strangers, but may become more familiar over time (whether through direct interaction or observation and mitigated through a process of consistently

being co-present together), and those who do become familiar may exert stronger influence on one another.

3.2.2 Social influence in other social contexts for chemotherapy patients

There is strong evidence to suggest that social influence can impact stress and therefore health outcomes through a variety of interpersonal processes. Much of the research investigating social influence processes in cancer patients has involved patients in settings outside the chemotherapy ward. For example, couples support one another when one member has developed cancer. Although only one member is affected by cancer, both experience stress due to the diagnosis and both are likely to engage adaptive changes to the stress by supporting each other[141]. In this context, couples have a pre-existing relationship in which members intimately know and support each other prior to the cancer diagnosis.

There is also a large body of evidence that members of cancer social support groups benefit from exchange of social support resources[147]. While patients in cancer support groups likely do not know each other prior to joining the group, patients join such groups because they are seeking and in need of support from experientially similar others. Additionally, all patients entering into cancer support groups self-identify as persons with cancer, which can enhance the social interactions with other, like-minded individuals. This is particularly important, as those with similar experiences are ideally situated to influence others' stress buffering via emotional sustenance, assistance in active coping, or role modeling[148]. Such support groups represent a context in which patients form a rapport de novo, followed by subsequent strengthening or weakening of their relationship.

In contrast, patients within the chemotherapy ward may or may not be actively seeking social support from others receiving chemotherapy; their primary purpose is cancer treatment. Any social exchange occurring while receiving chemotherapy is secondary to treatment. However, the chemotherapy ward is an important social context to investigate whether social influence naturally occurs, and if so, what

impact such influence has on patients' health outcomes. I hypothesize that social influence can be both positive and negative. For example, I would expect that positive patient outcomes, such as cancer survivorship, would positively impact patients receiving treatment together. However, negative patient outcomes, such as physical decline and death, may negatively impact the survivorship of patients receiving treatment together. The latter proposition is based on evidence indicating that the death of a close network member, such as a spouse, likely accelerates one's own mortality[149,150]. More generally, previous research has shown that disruption in one's social network can adversely affect psychological distress which can lead to adverse health outcomes[151,152]. Noticing that familiar patients are no longer in the chemotherapy ward (due to death or successfully finishing chemotherapy) may have similar effects if one considers those others as members of their social network. Thus, the chemotherapy setting represents an ideal context to investigate the influence of patient outcomes and network disruption on survivorship and whether such influence can occur when neither pre-existing social ties nor an explicit desire for social support are present.

3.2.3 Social networks

Past research has alluded to the role of social network resources in cancer patients' health trajectories, but has not made use of sociometric data, primarily due to the cost and difficulty associated with its procurement. These studies have predominantly focused on very small networks of couples[153], been primarily qualitative (Wolf, 2015), or have used data from a subset of the people represented in the network[154]. However, all of these studies build the evidence base that the connections between people within a network have important implications for social influence processes that facilitate cancer patients' adaptation to their diagnosis and treatment. Moreover, there is a breadth of research showing that social influence is impacted by network structure for a variety of health outcomes, and this is robust across age, race, and socioeconomic status[38,155,156,157]. Indeed, this body of research focuses primarily on established social relationships or support

group settings and shows that influence does not depend solely on individual, independent dyads or a direct social connection, but that the overall structure of the network is important. There are methodological issues when examining influence processes within such contexts—differentiating between selection or influence is generally difficult—limiting the types of inferences one can make[158]. However, the chemotherapy ward is not a context within which one self-selects or chooses partners, providing a unique opportunity to evaluate influence processes with a limited possibility for social selection bias.

3.2.4 Research questions and hypotheses

In the current paper, I investigate the influence of cancer patient health outcomes on individual five-year survival in a longitudinal chemotherapy treatment co-presence network. I do so in the context of the chemotherapy ward, where patients are unlikely to have previous social ties with others in the ward, and who are not present primarily to seek social support. Despite this, there is ample opportunity for social influence to occur either directly via the exchange of social support resources or indirectly by consistent observation of other patients' health changes over time. As such, health outcomes of those with whom cancer patients are consistently co-present can potentially influence individual patient outcomes. To this end, I investigate whether co-presence between chemotherapy patients is associated with survival and whether such influence mechanisms vary by sex[132].

3.3 Data and Methods

The data come from the Infections in Oxfordshire Research Database, which is composed of standard National Health Service (NHS) administrative records. This data set was originally established to monitor infectious diseases (e.g. recording full genotyping of infectious agents), and also contains complete individual health records. The data for the analysis comprises all 4,691 patients in the hospital's single outpatient chemotherapy ward from Jan 1, 2000 to Jan 1, 2009. I exclude 49 patients who received chemotherapy for conditions unrelated to cancer (e.g.,

for multiple sclerosis). The data also include 109 left-censored cases (representing 2% of cases) who were already receiving chemotherapy at the time data collection began. Each patient's medical history is broken into individual visitation spells (n=43,898) in the chemotherapy ward with time stamps for both entry and exit. A chemotherapy spell is defined as every occurrence with unique entry and exit times of any patient into the chemotherapy ward. The date of death is also recorded for each individual through June, 2015.

3.3.1 Chemotherapy ward and process of treatment

As an out-patient setting, the chemotherapy ward was open between 8 a.m. and 8 p.m. from Monday through Friday. The ward was split into two treatment rooms, containing 10 beds and 6 chairs, respectively, arranged in a circle (Figure 3.1). The beds were fitted with a screen that could be drawn when privacy was desired. Other than that, all patients in the same treatment room were in view of one another for the duration of treatment. Upon arriving to the ward, patients began in the waiting room and underwent bloodwork to ensure eligibility for chemotherapy. This was typically done on the day of treatment, but could be done the day before. Depending on the results of the blood test, chemotherapy could commence, be postponed, or canceled.

Chemotherapy regimens are stringent in the timing of doses. The timing of a patient's initial dose is based on the importance of immediate treatment, availability in the ward and patient preference. Once the first treatment is scheduled, the rest of the treatments follow a standard schedule, with a few hours of variation on any given day depending on patient availability. Therefore, patients with whom one overlaps are primarily determined by who is in the ward at the time of the initial dose, and the prescribed timing of chemotherapy. This informs pairs of individuals for which I consider influence possible.

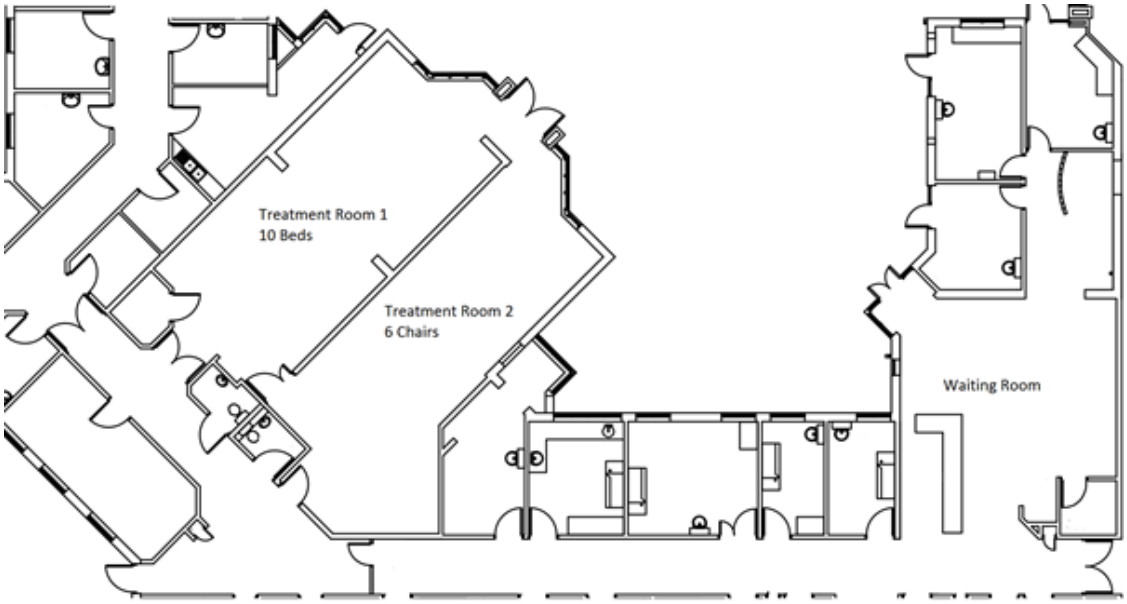

Figure 3.1: Layout of the chemotherapy ward. Treatment rooms 1 and 2 comprise 8 and 6 patient spaces, respectively. Patients begin spells in the waiting room, and are taken to either treatment room 1 or 2 depending on a number of factors.

3.3.2 Consistent co-presence network construction

The network of interest represents patient co-presence in the chemotherapy ward. Because I want the connections to represent quantities of co-presence with the potential for social influence, and no measure perfectly captures all the dimensions I think are important for social influence, I use two different methods for determining co-presence. Although a standard measure such as the Jaccard index provides a continuous, undirected measure that is both straightforward to compute and whose properties are well-known, it results in very high values for individuals who are only in the ward for short periods of time. Additionally, I observe significant heaping at certain values of the Jaccard index (Figure 3.2). This heaping likely represents underlying hospital policies that result in patterns of patient-patient overlap which are not controlled for in the Jaccard index. Moreover, the Jaccard index makes the assumption that social influence can occur after a moment of co-presence, which I believe is somewhat unrealistic since previous work has shown that the likelihood of influence is positively correlated with the strength of a relationship[159]. The primary measure of co-presence therefore defines patients as connected when they

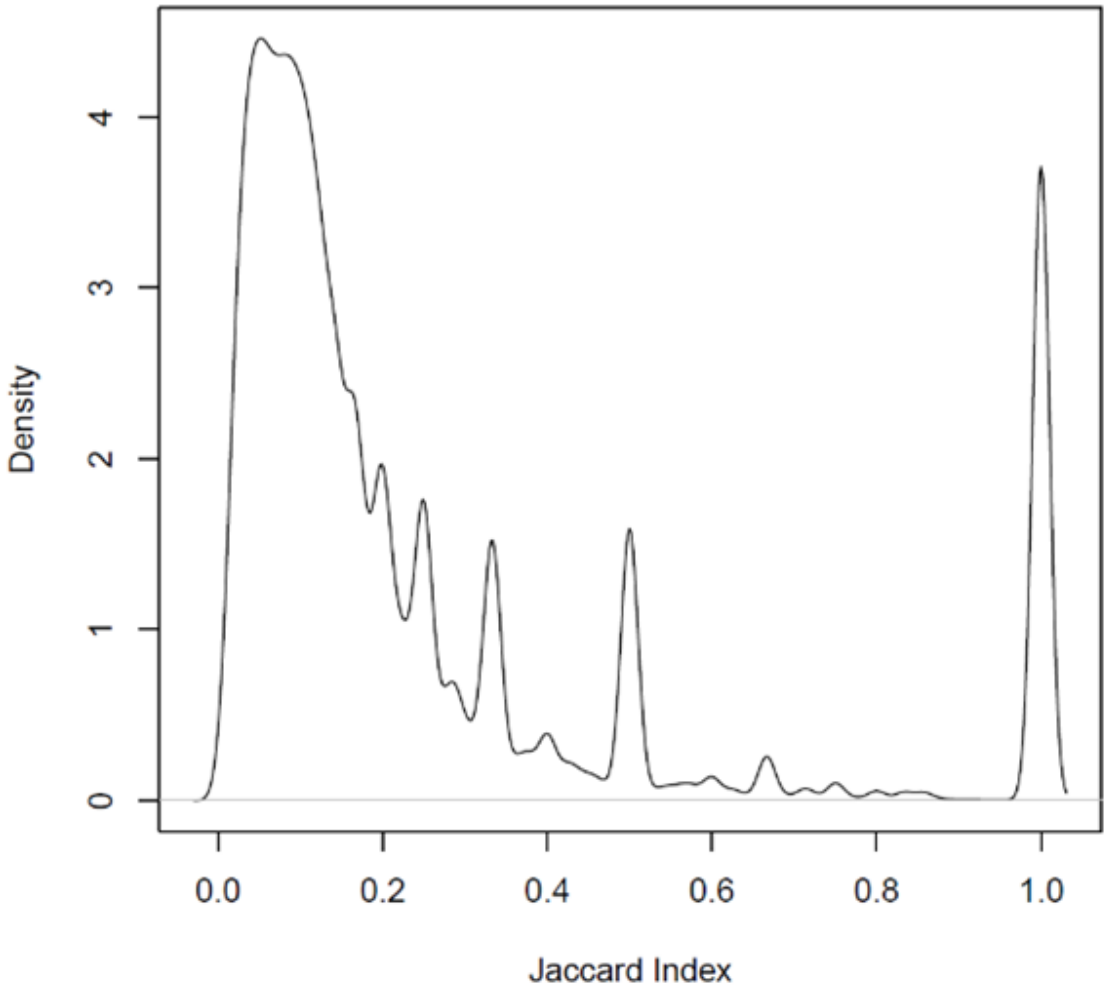

Figure 3.2: Kernel density-smoothed function of Jaccard indices. I observe heightened frequency of Jaccard index values at 1, 1/2, 1/3, 1/4, and 1/5, which indicates some sort of endogenous underlying process influencing patient ward spells and therefore overlap not accounted for by the Jaccard index.

are consistently co-present in the ward. This allows us to check the robustness of, and supplement the primary analysis.

I posit that two patients are consistently co-present (CCP) if their chemotherapy treatments overlap more often than would be expected by chance. To derive this variable, I first define a meaningful cutoff in terms of time co-present. As previously stated, a patient's chemotherapy visits are primarily determined by their first visit and the standard schedule for their prescription. I assume a patient's first visit could vary plus or minus one day based on bed availability and patient preference. For example, assume a patient's first treatment spell is on March 15. The window

for overlap is therefore March 14-16. All patients who had a spell within this window represent the corresponding risk set of patients with whom the focal patient could have overlapped. The window defining the risk set excluded weekends; thus, the risk set window for a patient scheduled to receive treatment on a Monday included the previous Friday and the following Tuesday. Based on these assumptions, I determine how often each patient would have overlapped with others conditional on when chemotherapy began. I observe with whom overlap and for how long they overlap based on a random sample from the risk set, assuming the periodicity of chemotherapy holds. By repeating this procedure 1000 times for each patient, I create a patient-specific empirical distribution of overlap times, and draw a cutoff at the 99th percentile of this distribution, forming an edge between the focal actor and the other patient. Formally, this can be written as:

$$A_{ij} = \begin{cases} 1 \text{ if } |H(i) \cap H(j)| > Q_{99}(|H(i) \cap H(j)|) \forall k \\ 0 \text{ otherwise} \end{cases}$$

Where H(i) is the set of hours spent in the ward by patient i and Q_{99} is the 99_{th} percentile of that patient's co-presence times with other patients based on the empirical distribution. For a visual explanation of this method, see Figure 3.3.

Thus, an edge in this co-presence network is drawn between two patients when the amount of time spent co-present in the ward is greater than the time at least one of the patients in question spends with 99% of the risk set of patients randomly sampled. I will refer to the edges in this network as an indicator of patients who are "consistently co-present" (CCP). Unlike the Jaccard-weighted person-hours, $A_{ij} = 0$ does not imply that there was 0 overlap between i and j, only that I do not consider such overlap significant, that is, i and j are not consistently co-present. While this method results in a directed network (patient i's overlap with patient j may be significant for patient i but not patient j), nearly all ties were mutual (>99%). Given that empirically nearly all ties are mutual and theoretically co-presence is symmetric, I treat the network as undirected. Thus, edges in the network have the convenient interpretation that two patients are connected when

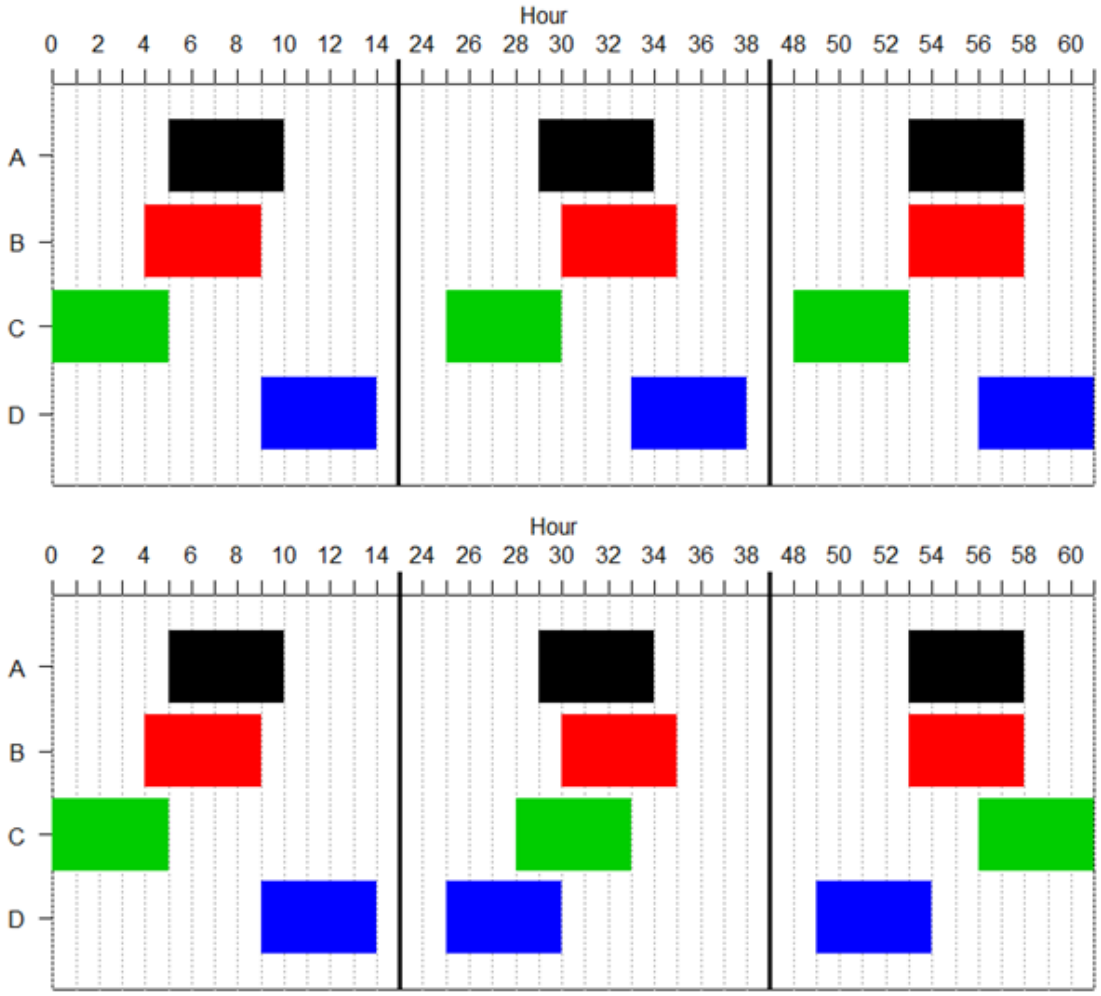

Figure 3.3: Heuristic vignette of what constitutes consistent co-presence. Colored blocks indicate the hours each of 4 patients are present in the ward over three different days. A) shows a case where patient A overlaps with patient B for all three spells. However, had A's first spell been 5 hours earlier or later, they would have overlapped with C or D, respectively just as much as they overlapped with B, so their overlap with B is not more than expected by chance due to random variation in the first spell, and therefore A and B are not consistently co-present. B) shows a case where A and B would be considered consistently co-present. Here, A still overlaps with B during all three spells. Had A's first spell been moved earlier or later, A would not have greatly overlapped with any other patients, so A's overlap with B is greater than that expected by chance. This therefore controls for the underlying scheduling possibilities not accounted for by the Jaccard-weighted person-hours.

at least one of them was consistently co-present with the other. The resulting network is shown in Figure 3.4.

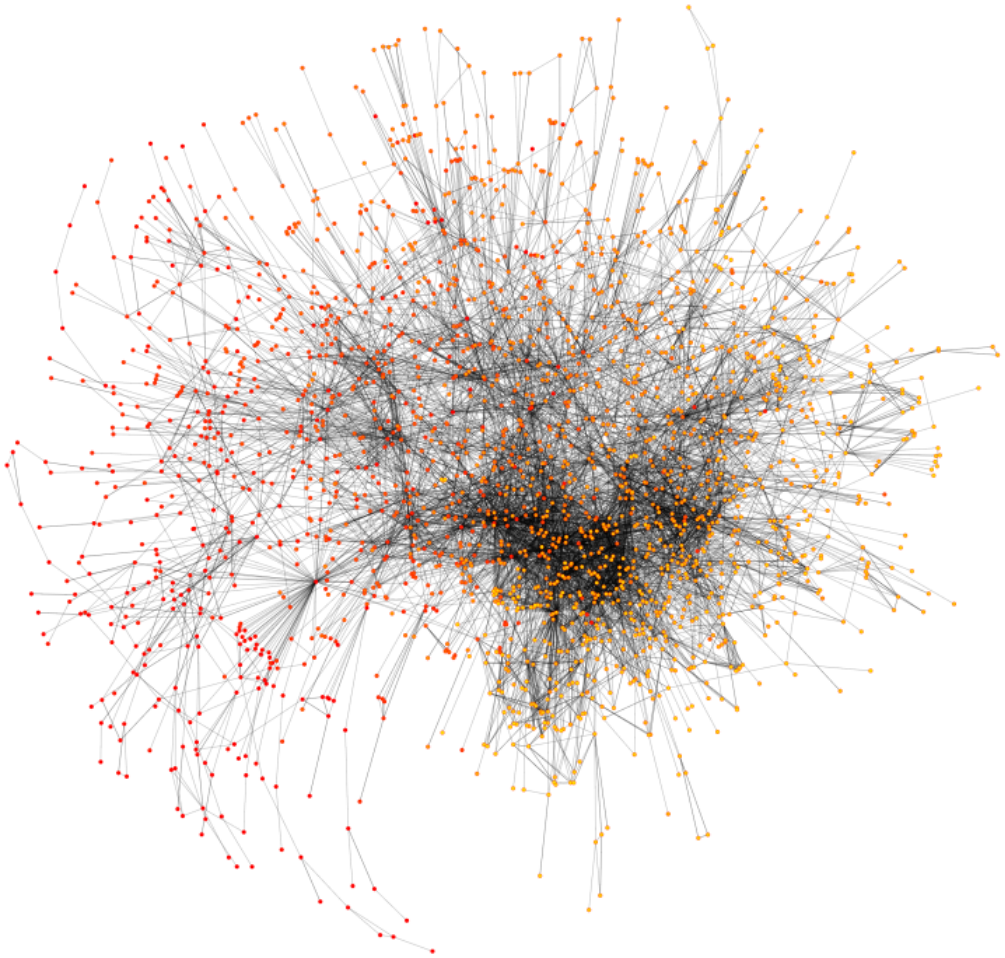

Figure 3.4: Exemplary largest connected component of network overlap among chemotherapy patients from 2000 to 2009 (n=2,228). An edge exists between two patients if they were co-present in the chemotherapy ward more than expected ($p<0.01$). Node color ranging from white to red indicates the week at which each patient began chemotherapy, representing the temporal nature of this network (with white values corresponding to January 1st, 1998). The edge color value indicates the amount of time the two connected patients spent together in the ward (darker edges representing more time co-present in the ward).

3.3.3 Dependent variable

The primary outcome is a patient's 5-year mortality, which is the gold-standard in cancer survivorship research and practice[127]. Patients' outcomes were measured at the end of treatment by recording the last chemotherapy session followed by a 6-month period with no visit to the chemotherapy ward. The outcome was recorded 1=death if they died within 5-years of their end of treatment date and had a diagnosis of cancer at the hospital spell most temporally proximate to their date of death (i.e. count as censored patients who die of causes unrelated to their cancer), and 0=survival otherwise. Because the chemotherapy data ends in 2009, and I have death data through mid-2015, there is no administrative censoring for 5-year survival—I treat de facto right-censoring after 2015 simply as 5-year survival.

3.3.4 Independent variables

Any consistent co-presence: I construct a dichotomous variable for whether a patient has any edges in the consistent co-presence network.

1-path Influence: I count the number of connections a patient has in the network.

$$CCP(i) = \sum A_{i\cdot} \tag{3.1}$$

Furthermore, I posit that the effect of consistent co-presence may differ based on the health status of the patient with whom one is consistently co-present. I therefore divide these counts by the 5-year survival status of the alters.

$$CCP_{S1P}(i) = \sum_{j}(A_{ij} * S(j)), j \neq i \tag{3.2}$$

$$CCP_{D1P}(i) = \sum_{j}(A_{ij} * (1 - S(j))), j \neq i \tag{3.3}$$

Where $S(j)$ is equal to one if patient j survived at least 5 years following their chemotherapy and zero otherwise.

2-path Influence: Although I *a priori* expect the influence to be between directly-connected patients only, I include 2-path influence variables representing the presence or absence of 2-paths in the consistent co-presence network as a negative control.

While it is possible that a patient two steps away may influence a directly-connected patient, I believe there would be no influence independent of the mediating patient, and such an effect should be non-significant. Additionally, because they both overlap with similar patients, any latent characteristics leading to co-presence would also likely be similar. Although a significant result for these covariates could be due to either latent similarity between patients two steps away or due to influence over two steps in the network, the lack of a significant result indicates that neither of these mechanisms are likely at work *ceteris paribus*. I therefore construct the variables as total weight of open two-paths between a focal patient and another patient based on whether the other patient survived (CCP_{S2P}) or died (CCP_{D2P}). Weights of the open 2-paths were equal to the product of the two individual edge weights on the path. These can be written as:

$$CCP_{S2P}(i) = \sum_j ((A \times A)_{ij} * S(j)), j \neq i \tag{3.4}$$

$$CCP_{D2P}(i) = \sum_j ((A \times A)_{ij} * (1 - S(j))), j \neq i \tag{3.5}$$

Co-presence and Sex Interaction: Finally, I add two terms to the model for the interaction between patient sex and the direct path consistent co-presence terms (CCP_{S1P} & CCP_{D1P}) to determine if there are differences in observed social influence based on the sex of the focal patient (as opposed to those with whom they are consistently co-present).

3.3.5 Covariates

A number of covariates are included in the fitted models as controls for other sources of heterogeneity. First, I include patient sex and the age of the patient at the start of chemotherapy. In addition, I control for variables related to the chemotherapy treatment itself. These include the number of visits to the ward during the chemotherapy cycle and the total time in the ward over all ward visits. As well, I include the timing with respect to the 10 years of observation of when each patient began their chemotherapy, allowing me to control for any exogenous

changes to survival as treatment improved over time. I also control for the total person-hours of overlap each patient had with all others in the ward.

To account for the effect of disease severity on 5-year survival I include variables derived from the ICD-10 codes (C00-C99) on patients' health records[101]. Previous studies largely focused on a single type of cancer and adjust for the disease stage[160]. No studies could be found that adjusted for the severity of cancer across primary cancer sites, giving no basis from past literature for which to adjust for disease severity. Thus, I generalized past approaches by including all primary cancer types as a series of dummy variables, one for each observed cancer type, for a total of 20 variables, with pancreatic cancer as the arbitrary baseline. Empirical 5-year survival in these data ranged from 7% of patients for brain cancer to 94% for prostate cancer, indicating a large variance in basal prognosis with respect to the location of the primary cancer[116]. Patients that were recorded as having non-specific, ill-defined, secondary, or miscellaneous multiple sites (i.e., from ICD codes C76, C77, C78, C79, C80, C97) were given their own dummy variable as these cancers are typically more rare and more difficult to treat[161].

Although the ICD-10 code does not explicitly differentiate between stages of cancer, I can distinguish Stage IV, the most severe cases, where the neoplasm has aggressively spread to a second tissue with a secondary cancer diagnosis (metastasized)[162]. I include an indicator variable for any secondary cancer diagnoses during a patient's treatment as a proxy for metastasis. Left-censored patients (n=145) had this variable imputed based on the full data set as a function of a patient's covariates. Model checking diagnostics revealed that the results were robust to this imputation.

Finally, I control for the admitting consultant physician. The admitting consultant physician can induce latent homophily among patient outcomes if their patients are placed together in the ward and survival outcomes are similar due to either shared physician treatment decision strategies or because patients have similar cancer types. For every spell, an admitting consultant physician is assigned – these physicians generally specialize in a given cancer type. Over the 9-year period

of the study, there were a total of 73 admitting consultant physicians. However, only 24 of them saw at least 10 different patients. To retain model parsimony and avoid degeneracy, I included 24 indicator variables for these physicians. The referent group is therefore those patients who saw one of the 49 admitting physicians with less than 10 chemotherapy patients. For parsimony and space I show only the physicians with the largest positive and negative significant effects.

3.3.6 Analysis

To evaluate the hypotheses that social influence in a chemotherapy ward impacts patient mortality, I fit a series of Generalized Estimating Equations (GEE) to account for the repeated measures of individuals with multiple chemotherapy cycles. I use a binomial variance with a logit link function to estimate the probability that an individual dies within 5 years of their last treatment. I use an exchangeable covariance matrix to model the correlation between patients' successive chemotherapy treatments. I fit several models using a step-wise blocked variable design. In the first block (Model 1), the outcome is modeled as a function of focal actor age, gender, time of chemotherapy cycle, number of treatment spells, the prognosis rank of cancer with which the patient was diagnosed, whether an individual had multiple tumor diagnoses (proxy for metastasis), total hours in the ward, and total person-hours of overlap with other patients. I then added the direct influence effects (Model 2) and then the indirect influence effects, or open two-paths (Model 3). Finally, as previously stated, patient sex may moderate the relationship between social contact[132]. I therefore add interaction terms between the influence effects and whether the focal patient is male or female (Model 4).

3.3.7 Sensitivity analyses

Based on the limitations and assumptions of the model and data, I perform a number of sensitivity analyses to assess how robust the results are to these limitations and results. These include an alternative assessment of co-presence, the effect of total independence between observations, the effect of nurses on health outcomes, a

survival analysis instead of a dichotomous outcome, and treating cancer diagnosis as a continuous variable rather than categorical.

Jaccard Index as an alternative measure of co-presence

Although the consistent co-presence measure was designed specifically to address the question at hand, I recognize that its novelty makes assessing and interpreting it more difficult. As a result, I use an alternative measure for co-presence which is more well-understood. The secondary network measure I use is the Jaccard index, the matrix of which is defined as the intersection of the two patients' treatment times divided by the union of these patients' treatment times. In other words:

$$J_{ij} = \frac{|H(i) \cap H(j)|}{|H(i) \cup H(j)|} \tag{3.6}$$

where $H(i)$ is a function that returns the set of hours patient i spent in the chemotherapy ward during the course of their treatment, and its magnitude is the total number of hours patient i spent in the ward. I subset the denominator based on the time patients i and j could have spent together; i.e. when they are both alive and undergoing chemotherapy. This gives a quantitative measure of how often i and j were together relative to how often they could have been together, resulting in a weighted and undirected network such that $J_{ij} = J_{ji}$, and $J_{ij} = 0$ indicates 0 hours of overlap between i and j.

However, the Jaccard index has some limitations, hence my use of it as a secondary measure. One such limitation is its lack of sensitivity to total patient chemotherapy hours; when two patients overlap and are only in the ward for a single spell, their Jaccard index is very high and indistinguishable from two patients who are repeatedly in the ward together over time. This is evident in the large peak in the empirical density of the Jaccard index at one (Figure 3.2). Of those, 1,762 Jaccard indices stem from single-visit overlaps. To address this, I up-weight the Jaccard index by the hours spent in the ward by the focal patient.

Using this metric, I construct 1-path and 2-path weights similarly to the consistent co-presence network. However, due to the limitations of the Jaccard

index, and because it is continuous rather than dichotomous, I make modifications. As previously noted, complete overlap of treatment, whether over 1 or 100 hours, will result in equivalent Jaccard indices. I therefore weight the Jaccard index by the number of hours the focal patient spent in the ward. I call this the Jaccard-weighted person-hours, which differentiates overlap of only a few hours from that of many hours while still adjusting for the relative potential for overlap by any two patients. The Jaccard-weighted person-hours for a focal patient can be written as:

$$JW(i) = |H(i)| \sum_j J_{ij}, j \neq i \tag{3.7}$$

This total count is then partitioned into the Jaccard-weighted person-hours with directly-connected patients who survived at least 5 years following chemotherapy (JW_{S1P}) and with those who died within 5 years following chemotherapy (JW_{D1P}), allowing us to separate their relevant effects and understand more about underlying influence mechanisms. These can be written as:

$$JW_{S1P}(i) = |H(i)| \sum_j (J_{ij} * S(j)), j \neq i \tag{3.8}$$

$$JW_{D1P}(i) = |H(i)| \sum_j (J_{ij} * (1 - S(j))), j \neq i \tag{3.9}$$

Where $S(j)$ is equal to one if patient j survived at least 5 years following their chemotherapy and zero otherwise.

I also include 2-path influence variables representing the sum of Jaccard-weighted person-hours for paths to non-adjacent patients two steps away from the focal patient based on their outcomes (survival or death) as a negative control. I therefore construct the variables as total weights of open two-paths between a focal patient and another patient based on whether the other patient survived (JW_{S2P}) or died (JW_{D2P}). Weights of the open 2-paths were equal to the product of the two individual edge weights on the path. These can be written as:

$$JW_{S2P}(i) = |H(i)| \sum_j \sum_k (J_{ij} * J_{jk} * S(k) * M_{ik}), j \neq i, k \tag{3.10}$$

$$JW_{D2P}(i) = |H(i)| \sum_{j} \sum_{k} (J_{ij} * J_{jk} * (1 - S(k)) * M_{ik}), j \neq i, k \tag{3.11}$$

Where M_{ik} equals 1 when i and k are not directly connected in the co-presence network (i.e. they are never co-present), and zero when they are.

Cancer severity as a continuous variable

Although I treat cancer severity as a series of dummy variables for the primary cancer diagnosis, I recognize this might not be the most parsimonious way to do so. Additionally, this method averages out the correlation with mortality across the study period. If the prognosis changes (as I observe a significant decrease in mortality the later during the study period one begins chemotherapy), the method I use will not account for this. However, if I treat the cancer type as a rank-ordered variable based on the predicted-year mortality, I save degrees of freedom in the model. Additionally, although the prognoses of the types of cancer may change during the study period, the rank-order should be less susceptible to change.

Correlations between patients

It should be noted that the GEE does not adjust for the correlation between patients. Patients within two steps of each other may have correlated survival/death counts (e.g. if patients A and B are both connected to patient C, and C survives, A and B's count are correlated) and may also have other unmeasured latent factors which are correlated between them. To examine whether this had an effect on the model inferences, I reran the GEE with random samples of patients who are all at least two steps away from one another and should, therefore, be largely independent.

Survival Analysis

Although the outcome of interest is 5-year survival, the data include precise patient survival times. As such, I also fitted a Cox proportional hazard model to evaluate the robustness of the inferences[163]. This allows assessment of the hazard of death accounting for survival time instead of just the probability of death.

Sensitivity Analysis for Nurses

I recognize that nurse heterogeneity could affect the health outcomes of patients and also be correlated to patient-patient co-presence, which could explain the results. I was unable to obtain data on nursing staff, but I present here a sensitivity analysis. First, I create n nurses, ranging from 5 to 95 and randomly assign each a quality of care parameter. This parameter takes a normal distribution with mean=0 and sd=0.65, the distribution of physician parameters from Table 3.2, Model B, meaning I assume that the effects of nurse heterogeneity are approximately the same magnitude and distribution as physician heterogeneity (which I observe). Each patient is assigned a primary nurse either based on assortative mixing with the nurses of their neighbors or their own health outcome. The probability of whether the nurse is based on assortative mixing or patient outcome was chosen to range from 0.05 to 0.95. If the nurse is assigned based on the neighbor(s)' nurse(s) one nurse is chosen at random among neighbors set with probability proportional to the Jaccard index with those neighbors, which is written as:

$$P(N_i = N_j) \propto J_{ij} \tag{3.12}$$

Where N_i is the nurse assigned to patient i.

If the nurse is chosen based on the patient's outcome, then a nurse is chosen with probability proportional to the probability of survival if the patient survived, or proportional to the probability of death if the patient died. I assume the nurse heterogeneity parameter relates linearly to the log-odds of survival (because the physician parameters do), and so calculate the probability straightforwardly. For this probability, I assume all patients have the population mean probability of survival (33%) as their baseline which is modified only by the nurse heterogeneity parameter. This is written as:

$$P(N_A = n) \propto \begin{cases} \frac{exp(log(0.5)+HN(n))}{1+exp(log(0.5)+HN(n))} \text{ if } S(A) = 1 \\ 1 - \frac{exp(log(0.5)+HN(n))}{1+exp(log(0.5)+HN(n))} \text{ if } S(A) = 0 \end{cases}$$

Where HN(n) is the heterogeneity parameter for nurse n.

This results in 95% of probabilities of survival ranging from 12.0% to 64.7%. Assuming the patient survived, a nurse with a heterogeneity parameter at the 97.5th percentile of the distribution would be chosen 5.4 times more often than a nurse with a heterogeneity parameter at the 2.5th percentile. This forced association between nurse parameter and patient outcome is not meant to precisely reflect how nurses are assigned in the ward, but rather to fulfill the necessary condition that nurse assignment is correlated with outcome to induce confounding. Each combination of number of nurses and assortativity probability was run 100 times, and the proportion of significant ($p<0.05$) direct effects were recorded.

Ties concurrent with pre-existing social ties

Given that the study population is drawn from a relatively small catchment area, it is possible that my belief that patients in the chemotherapy ward do not know one another prior to initiating chemotherapy is incorrect. However, I believe ties of this sort are very unlikely, as they would stem from the confluence of a number of unlikely events. First, both patients in a dyad would need to be diagnosed with cancer, the lifetime risk of which in the UK is 50% (UK, 2014). Second, both patients would need to know one another. Given a population size at risk of cancer in Oxfordshire is 80,000 and that each person knows on average 600 individuals[164], each pair of individuals has an 0.75% chance of knowing one another under a random mixing model. Finally, the individuals must be diagnosed around the same time, have similar availability schedules, and go to the same clinic, to be on chemotherapy concurrently. Given that most cancers occur between the ages of 50 and 80, this occurs one time in thirty if I non-conservatively assume that a year qualifies as "around the same time". If I assume these events are independent then the probability of two patients knowing one another and being on chemotherapy around the same time is 0.0000625. Given that each patient sees 115 patients on average in chemotherapy for at least one hour, this only nets out to an average of about 0.007 such alters per patient stemming from pre-existing

social ties. I therefore do not believe that such preexisting ties would be a large enough presence to affect the results.

3.4 Results

Table 3.1 reports the descriptive statistics of the sample. The 4,691 chemotherapy patients from Jan. 1, 2000 to Jan. 1, 2009 had a mean age of 59.8 (SD=13.1), and 44% (n=2,094) were male. The patients underwent a total of 43,898 chemotherapy spells, which consisted of an average of 8.5 (SD=10.9) visits to the chemotherapy ward per spell with each visit lasting, on average, 4.0 hours (SD=5.3). Two thirds of the patients had a single diagnosis of cancer (N=3,122), 994 had 2 diagnoses, and one patient had 9 diagnoses. With respect to the affected organ or organ system, 1,108 patients were diagnosed with breast cancer – the most common cancer diagnosis – treated in the chemotherapy ward with a 5-year survival of 91% (based on UK statistics). Approximately 500 patients were diagnosed with lung cancer; lung cancer is the second most severe cancer type with a 5-year survival of only 18% in the UK. A total of 850 people had a diagnosis of unspecified or multiple cancers that could not be classified to a single location. Finally, the 145 patients who were left-censored and therefore had no recorded cancer diagnoses had imputed values between 1.21 and 1.54 cancer diagnoses.

Results based on the GEEs using consistent co-presence are presented in Table 3.2. Being older, male, or having more severe cancer was associated with increased likelihood of death across all models (Model 1). This is consistent with previous literature and trends in cancer survival. The number of ward visits during a chemotherapy cycle and the total time of the cycle were not significant predictors of death. Additionally, the later in the study period a person began chemotherapy, the better their chance of survival, indicating a trend towards better treatment over time.

Variable	**Mean (SD) or N (%)**
Age	59.79 (13.00)
Male	2094 (44%)
Number of ward visits during cycle	8.51 (10.94)
Time of chemotherapy cycle (years)	0.32 (0.48)
Average time in ward per spell (hours)	3.95 (5.32)
Number of cancer diagnoses	1.30 (0.63)
Primary cancer diagnosis	
Breast	1108 (24%)
Lung	443 (9%)
Pancreas	125 (3%)
Unspecified	850 (18%)
Other	2165 (46%)
Number of patients co-present with	113.91 (122.18)
Total person-hours of co-presence	1012.66 (1,997,599)
No CCP	2,712 (50%)
CCP_{S1P}	0.93 (2.86)
CCP_{D1P}	1.50 (4.77)
CCP_{S2P}	13.96 (36.73)
CCP_{D2P}	23.09 (49.43)

Table 3.1: Demographic characteristics of the 4,691 patients receiving chemotherapy at any time from January 1, 2000 to Jan 1, 2009.

Variable	**Model 1** Estimate (95% CI) Residuals^2 = 817.7	**Model 2** Estimate (95% CI) Residuals^2 = 859.6	**Model 3** Estimate (95% CI) Residuals^2 = 857.6	**Model 4** Estimate (95% CI) Residuals^2 = 854.8
Intercept	1.166 (0.166,2.167)	1.257 (0.251,2.263)	1.28 (0.275,2.268)	1.266 (0.259,2.272)
Age (years)	0.038 (0.032,0.044)	0.038 (0.032,0.044)	0.038 (0.032,0.044)	0.038 (0.032,0.044)
Sex (male)	0.216 (0.029,0.404)	0.22 (0.032,0.408)	0.219(0.031,0.408)	0.28 (0.083,0.478)
Time of cycle (years)	-0.316 (-0.561,-0.071)	-0.351 (-0.608,-0.095)	-0.315 (-0.579,-0.051)	-0.359 (-0.626,-0.092)
Number of visits in course	0.003 (-0.009,0.015)	-0.006 (-0.0019,0.007)	-0.01 (-0.024,0.005)	-0.009)-0.023,0.006)
Years after 2000 patient begins chemotherapy	-0.123 (-0.16,0.086)	-0.122 (-0.159,-0.083)	-0.121 (-0.159,-0.083)	-0.121 (-0.159,-0.083)
Total person-hours of overlap	0.001 (-0.001,0.002)	0.001 (-0.001,0.002)	0.001 (-0.001,0.002)	0.001 (-0.001,0.002)
More than one cancer diagnosis	1.177 (0.528,1.826)	1.19 (0.535,1.845)	1.197 (0.542,1.851)	1.192 (0.536,1.848)
No CCP	0.463 (0.312,0.614)	0.401 (0.246,0.556)	0.420 (0.265,0.575)	0.410 (0.216,0.514)
CCP_{S1P}[1](per 1,000 person-hours)		-0.104 (-0.157,-0.051)	-0.108 (-0.161,-0.055)	-0.243 (-0.148,-0.338)
CCP_{D1P}[2](per 1,000 person-hours)		0.086 (0.051,0.121)	0.064 (0.029,0.099)	0.186 (0.169,0.203)
CCP_{S2P}[3](per 1,000 person-hours)			-0.006 (-0.026,0.014)	-0.006 (-0.026,0.014)
CCP_{D2P}[4](per 1,000 person-hours)			0.006 (0.0003,0.012)	0.006 (0.0003,0.012)
CCP_{S1P} X Sex (per 1,000 person-hours)				0.168 (-0.259,0.595)
CCP_{S2P} X Sex (per 1,000 person-hours)				-0.184 (-0.478,0.109)
Has most severe cancer (Brain)[5]	1.07 (-0.012,2.151)	1.028 (-0.054,2.109)	1.008 (-0.074,2.09)	1.015 (-0.067,2.098)
Has least severe cancer (Prostate)[5]	-3.965 (-5.133,-2.797)	-4.012 (-5.181,-2.843)	-4.015 (-5.184,-2.846)	-4.057 (-5.227,-2.887)
Seen by oncologist with best average outcomes[6]	-2.025 (-4.344,0.293)	-2.093 (-4.41,0.224)	-2.09 (-4.41,0.23)	-2.121 (-4.448,0.206)
Seen by oncologist with worst average outcomes[6]	0.913 (0.074,1.752)	0.903 (0.062,1.744)	0.919 (0.076,1.761)	0.918 (0.076,1.76)

Table 3.2: Results of Generalized Estimating Equations modeling influence via consistent co-presence. The model outcome is death within five years of ending chemotherapy. I used a binomial variance with logistic link function, and an unstructured covariance matrix for repeated outcomes on individuals.

[1] Consistent co-presence with directly-connected patients who survive at least five years
[2] Consistent co-presence with directly-connected patients who die within five years
[3] 2-path consistent co-presence with patients who survive at least five years
[4] 2-path consistent co-presence with patients who die within five years
[5] Also adjusted for 18 other primary cancer types, including unspecified as one type
[6] Also adjusted for 22 other physicians with at least 10 spells as the admission consultant

In the consistent co-presence network GEE I fit a term for whether an individual had any significant co-presence (No CCP). In Model 1, the term for whether an individual is an isolate (i.e., no CCP) indicates that isolates were more likely to die within five years than those co-present with other patients (0.420, CI: 0.265,0.575). Thus, patients benefit from significant consistent co-presence with at least one other patient in the ward, irrespective of alters' outcomes.

In Model 2, the 1-path influence parameters show a beneficial effect of neighboring patients having survived for 5-years (-0.104 CI: -0.157,-0.051) (CCP_{S1P}), and an adverse effect of neighboring patients having died within 5-years (0.086 CI: 0.051,0.121) (CCP_{D1P}). Thus, there appears to be both positive and negative direct influence based on the health status of those one spends time with. When considering 2-path influence, the effects of non-transitive two paths to patients surviving at least 5 years survival is not significant (CCP_{S2P}), but the result for non-transitive two paths with patients dying within 5 years was significant in Model 3 (0.006 CI: 0.0003,0.012) (CCP_{D2P}). This latter result indicates some evidence of influence through open 2-paths. Finally, there is no significant moderating effect of sex (Model 4).

To illustrate the influence of co-presence in the chemotherapy ward, I present the predicted probabilities from Model 2, for three hypothetical patients at varying levels of risk (Figure 3.5). The low, medium, and high-risk patients had predicted probabilities of death of approximately 23%, 69% and 91%, respectively. For low- and medium-risk patients, I observe approximately a 2% change in predicted survival when comparing patients with no co-presence with those co-present with only patients having one type of outcome (survival or death). However, when patients are co-present with a mix of patients who die and patients who survive, the net effect, on average, is a minor decrease in predicted probability of survival for the patient. For the high-risk patient, smaller changes in predicted probability are observed due to the high baseline risk of death.

The robustness analyses are, on average, consistent with the main findings.

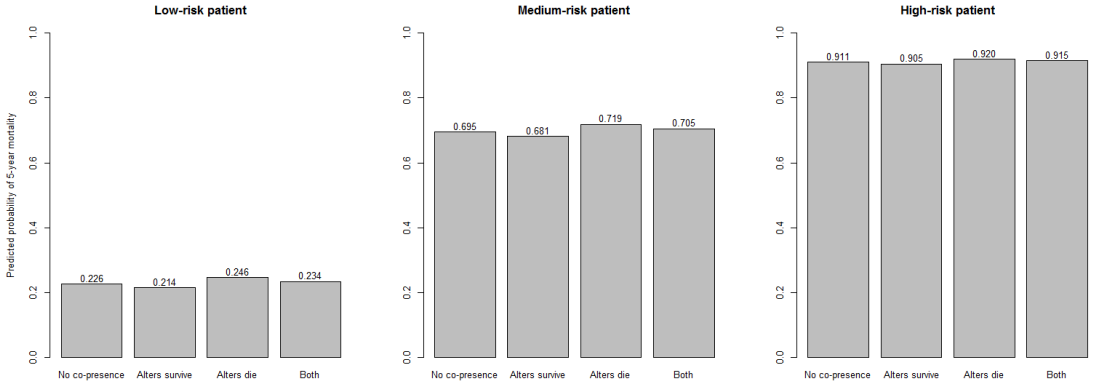

Figure 3.5: Predicted probability of 5-year mortality for patients with varying risk profiles and potential for social influence. Across panels, the first bar represents the predicted probability from model 4 with 0 for all influence terms. The average patient was one who had the median values for all covariates (rounded for dichotomous and categorical variables). This equates to a 69 year old female whose chemotherapy lasted 9 visits over 3 months and spent 30 hours in the ward starting in 2005, with a single diagnosis of a tumor of the ovaries. The low-risk and high risk patients had values based on the first and third quartile of the covariates depending on whether the relationship between 5-year mortality and the covariate was negative or positive, respectively. The low-risk patient was a 61 year-old female who visited the ward 9 times over the course of a month and spent 30 hours in the ward starting in 2007, with a single tumor of the breast. The high-risk patient was a 79 year-old male whose chemotherapy included 2 visits to the ward over 4 months and spent 30 hours in the ward starting in 2003, whose primary diagnosis was cancer of the stomach, but had multiple cancer diagnoses. It is important to stress that these patients are not necessarily observed in these exact combinations of covariates; they are chosen in the way they were to demonstrate heterogeneity of the predicted probability of survival. Within each panel, influence terms were given the rounded mean value for the variable in question (refer to Table 3.2). No influence means the patient was co-present with no-one (never actually observed but gives a baseline probability). "Alters survive" means a patient was only co-present with patients surviving at least 5 years, and "alters die" means a patient was only co-present with patients dying within 5 years. "Both" means a patient was co-present with both types of patients.

There are a few exceptions (Table 3.3 and Figure 3.3). Notably, the direct effect of co-presence with patients surviving 5 years is not significant when using an outcome of survival time instead of dichotomous survival. However, this analysis does not adequately adjust for intra-patient correlations. Additionally, the sensitivity analysis for nurse heterogeneity parallels the reduced robustness of the direct effect of co-presence with patients who survive at least 5 years evident from the survival analysis, assuming the simulation model adequately captures nurse-induced heterogeneity.

Variable	A) Jaccard index Estimate (95% CI)	B) Continuous cancer severity Estimate (95% CI)	C) Independent patients Median estimate (Min,Max)	D) Survival analysis Log(hazard ratio) (95% CI)
No CCP	NA	0.294 (0.181,0.407)	0.186 (0.094,0.290)	0.345 (0.191,0.536)
CCP_{S1P}	-0.117 (-0.175,-0.059)	-0.344 (-0.538,-0.149)	-0.290 (-0.905,0.823)	-0.598 (-1.249,-0.053)
CCP_{D1P}	0.103 (0.064,0.142)	0.357 (0.204,0.510)	0.851 (-0.301,1.143)	0.761 (0.326,1.197)

Table 3.3: Results of sensitivity analyses. The first 3 models are GEEs constructed in the same way as the primary analysis but with specific changes. A) Uses the person-weighted Jaccard indices (per 1000 person-hours). Because all patients were co-present with at least one other patient, the variable for any co-presence was not included. B) Treats cancer severity as a categorical variable. C) Is based on a sample of patients who were not co-present with one another to remove correlation between variables. Results are based on 100 trials of sampling a subset of patients in this way. D) Instead of a dichotomous 5-year survival outcome, I treat survival time as the outcome of interest, using a Cox proportional hazards model. In addition to main findings shown, all models also adjusted for the same covariates as in Table 3.2.

When examining cancer severity, Model B shows the results of treating cancer severity as a continuous variable rather than categorical (Table 3.3). The results are parallel to the main results measure in that both co-presence with surviving and co-presence with dying patients are significant.

For the potential lack of independence between observations, I remove a random sample of patients such that the only patients remaining are not directly connected. Model C shows results that are generally in the same direction and approximately the same magnitude as the models with all the data included (Table 3.3). Generally, the spread of coefficients is greater than the variance of the coefficients from the full model. Importantly, consistent co-presence with patients dying within 5 years had a median value very similar to the point estimate from the full model, indicating that the finding for that variable was reasonably reliable. The median value for the consistent co-presence with patients surviving 5 years also had a median value very similar to the point estimate from the full model.

Because a dichotomous outcome allowed us to use a GEE framework, I focused on that outcome rather than a survival analysis, despite the potential for censored observations. I therefore examined a survival analysis (Table 3.3, Model D). I observe that the results are generally similar to the main results, except there is a non-significant relationship between consistent co-presence with patients surviving for 5 years and patient survival. This may in part be due to the lack of control for intra-patient correlations as in the GEE. It may also be in part due to the assumption of proportional hazards in the Cox model - if the assumption is violated the model would fit poorly.

Finally, I examine the potential for a nurse's ability to induce correlations between patients' outcomes, accounting for these results. For each combination of number of nurses and assortativity, I show the proportion of models in which the main results remain significant (Figure 3.6).

Overall, I see that the significance of the main findings is relatively robust to the

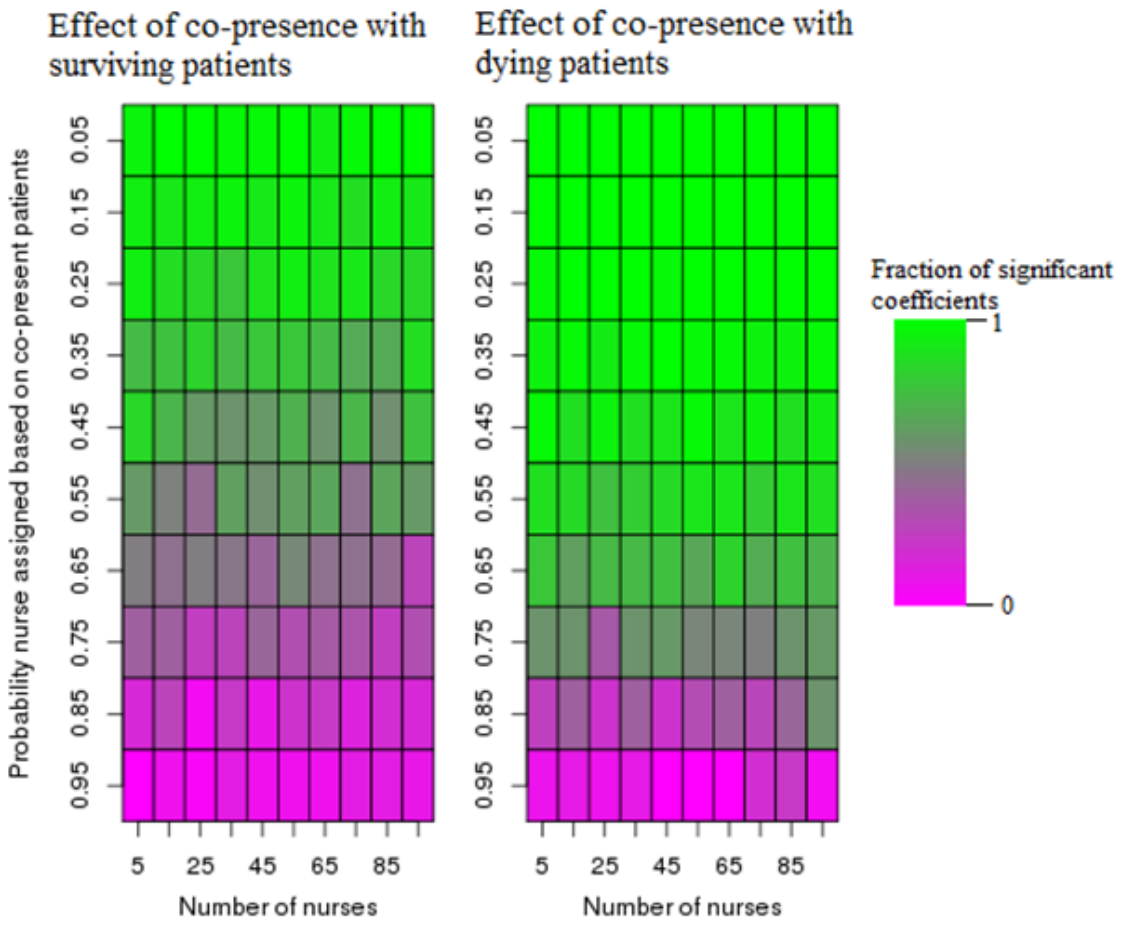

Figure 3.6: Heat map represents results of sensitivity analysis for effects of nurse heterogeneity.

number of nurses. I observe that the fraction of significance falls below 50% when the probability of assigning nurse based on adjacent patients' nurses falls between 0.55 and 0.65 for the effect of co-presence with patients surviving at least 5 years, and between 0.75 and 0.85 for the effect of co-presence with patients dying within 5 years. These sensitivity analyses show that the results are relatively robust to nursing effects on the order of physician effect, since at any given time there are multiple nurses in the chemotherapy ward, and nurses generally take on patients as the nurses are available. Because of this, many patients who are co-present will have different nurses, resulting in a relatively low probability that co-present patients will have the same nurse. Therefore, the high values of nurses being assigned to

co-present patients are likely greater than those actually occurring, meaning the results would likely not be reduced by the effects of having a common nurse.

3.5 Discussion

In the current paper, I investigated whether cancer patient survival is associated with the survival of those with whom they are co-present during chemotherapy treatment. The results suggest that co-presence matters. I find that a connected patient's death increases the likelihood of the focal actor dying, and a connected patient's survival decreases the focal actor's chance of death. These results are approximately symmetric; being connected to a single survivor is similarly protective as being connected to a single non-survivor is deleterious to patient survival. There are no significant results for interactions between consistent co-presence and patient's sex, which is not what I expected given previous findings[132]. However, the study data come from previously unexplored social environment, so different mechanisms may be at play. Placing the results of this study in the context of cancer treatment, the magnitude of these results (Figure 3.5) is less than that of chemotherapy clinical trials but still clinically meaningful. Here I observe a survival differential of 2% for patients at low to moderate risk when comparing no co-presence vs only co-presence with surviving patients, whereas effective chemotherapy clinical trials report survival differences of around 8%[165,166]. In effect, chemotherapy patient survival may be modified by one quarter the quantity of the effect of the choice of chemotherapy.

Regarding the 2-path variables, I observe a non-significant result for the consistent co-presence with surviving patients two steps away. At the same time, I found a significant effect of consistent co-presence with patients dying within 5 years 2 steps away, albeit marginally. As such, it would therefore only take a minimal amount of latent confounding between patients to remove that effect while leaving the main effect significant. Given the small magnitude of this effect and

its significance level, even a small amount of measurement error could result in this finding. Therefore, I believe that the 2-path influence variables reinforce the main findings. However, future research should aim to more thoroughly test if this, and other network effects not measured here are in fact at play.

Based on the above results, the mechanism by which social influence occurs becomes clearer. I observe no significant association between total person-hours of overlap and one's outcome, indicating that solely being around more people for more time on its own does not affect one's health. In this context, just being around others receiving treatment with similar stressors does not seem to impart any health effects suggesting that social facilitation and social support are not the underlying influence mechanism. Previous research suggests that only 2/3rds of chemotherapy patients in the UK indicated receiving adequate emotional support from hospital staff (NHS, 2014). Thus, many patients may be actively seeking but not receiving support from others, particularly other patients, during treatment. Future research should therefore focus on whether and how support networks emerge in chemotherapy wards to elucidate the content of interactions we've detected via co-presence here.

The influence effects between connected patients, on the other hand, likely reflect mechanisms where one's outcome is related to the outcome of others, where either co-present patients form social relationships, or they observe others' health trajectories. The mere observation of other cancer patients' health changes over time may influence the observer's own stress regarding their own cancer prognosis and subsequent health[142]. This process is akin to social modeling, however one may not consciously be altering how they respond to treatment. Rather, patients may see others doing better or worse which might decrease or increase their stress, respectively, which in turn can impact their health.

Although network disruption was possible, it is unlikely that it is the mechanism given the results. If patients being removed from chemotherapy caused disruption to other patients' networks in the ward, then I would expect to see adverse effects

of focal actors' neighbors finishing chemotherapy regardless of their neighbors' outcomes. Instead I observe connected patient' outcomes are positively correlated, indicating network disruption is not the mechanism underlying the findings.

Since the data used herein are observational, the results may stem from unmeasured confounding. I address the two forms I believe have the greatest potential to explain the results, but this is not an exhaustive list. First, patients may know one another prior to entering the ward, and any "social influence" observed here is the result of social interaction and influence outside of the chemotherapy ward due to preexisting social ties. If true, I detect social influence outside the chemotherapy ward via the measure of co-presence and in the presence of large amounts of noise. However, as I show in the Methods, I believe that the number of ties due to pre-existing relationships is likely minimal, limiting the effect this could have on the results. Second, nurse heterogeneity may explain the results if related both to patient outcomes and co-presence between patients. All of the nurses in the ward are specially trained as chemotherapy nurses which would ideally limit heterogeneity across nurses[167]. Furthermore, given that I assume nurse effects are on the same order of magnitude as physician effects in the sensitivity analysis, I observe that the results are robust to nurse heterogeneity. Alternatively, if nursing effects (due to either nurse heterogeneity or other endogenous factors such as understaffing) do in fact explain the results, then I have detected meaningful nurse effects on patient survival previously not reported in the literature. Such potential effects should be investigated in the future. However, it is still possible that these confounders are not exhaustive. Outside of the approaches taken here, one can also employ sensitivity analyses which can place bounds on the size of an unmeasured confounder which are sufficient to remove the significance of the findings[168].

With the above limitations in mind, this research has potentially important implications for the study of social networks. Whatever the underlying mechanism, I detect influence effects based solely on co-presence data from administrative records,

which is much more efficient to gather than detailed relationship data. Although one may ask whether the ties used here really represent meaningful social ties, a recent survey found that 78% of patients in Britain said they would prefer to be treated in a communal setting indicating that patients feel they benefit from in close proximity to other patients[169]. Thus, consistent co-presence represents the opportunity for patients to develop meaningful social ties during treatment. While it is unknown whether such ties developed amongst the patients studied here, it is clear that there is evidence of social influence among those who are consistently co-present. Moreover, I have employed a variety of novel approaches in my effort to rule out alternative explanations of the main findings. Future researchers may find the use of rank-order cancer diagnosis, cancer severity, physician effects, and holistic robustness checks valuable in their own work.

The results imply that co-presence relates to health in the context of the chemotherapy ward. This is particularly important, as some chemotherapy wards are moving towards individual rooms for patients. However, given the observed negative effects and possibility of unmeasured confounding, implementing changes to the ward and patient scheduling to benefit from this knowledge is difficult. If the observed social influence operates via changes to stress, then reducing patient stress in the ward without changing patient scheduling may be able to positively impact all patients. Oncologists can consider whether social influence may be at play in the wards to which they admit patients for chemotherapy, and whether scheduling to maximize co-presence of patients would *sui generis* be therapeutic. Given evidence that cancer support groups improve survival, patients could also be encouraged to engage in social support while in the chemotherapy ward, which could reduce stress[137,133]. Altering chemotherapy in this way can mitigate the deleterious effects of co-presence with patients experiencing negative outcomes and strengthen the positive effects of co-presence with surviving patients.

With these future directions and applications in mind, the findings in this

paper are an important first step in describing the possibility of social influence occurring among patients co-present in a chemotherapy ward; a setting primarily for biological treatment, not treatment through social support and influence processes. Importantly, because I focus on mere co-presence, any findings not due to unmeasured confounding are likely to under-estimate the effect of social forces on health outcomes as co-presence represents the minimally necessary condition for influence. I hypothesize that the mechanism of this influence is mediated by stress response to co-presence with familiar others. Future research should focus on measurement of individual coping and stress processes in these settings to test this hypothesis directly.

4

Co-presence with infected patients predicts nosocomial infection

Contents

4.1 Abstract

Nosocomial infections are a significant burden on the health care system. One potential avenue of research aimed at decreasing nosocomial spread is to detect infection earlier. I examine to what extent patient-patient co-presence, defined by patients concurrently residing in the same ward bay, as assessed by electronic medical records, serves as a screening test for infection. Although assessed retrospectively, this method is designed for prospective use in earlier identification of infected individuals in hospitals. Also, rather than examining only those with confirmed infections (as in the case of contact tracing), I count co-presence with those *suspected* of infection as well. I examine all 133,304 patients in a single UK county's NHS trust from 2011-2015 who were in the health care system for at least 48 consecutive hours. I count the number of hours each patient is co-present with those who received an infectious disease test. I treat this count as the index diagnostic test for subsequent infection, and examine its efficacy across five infections from the following organisms: *E. coli*, MRSA, *C. difficile*, *P. aeruginosa*, and norovirus. I compare this to the reference test of a positive microbiological test or diagnosis. I calculate ROC curves and their corresponding AUC, as well as sensitivities and specificities at optimal cut-points. Finally, I determine how many hours earlier or later each patient would receive a positive result from the index test relative to the reference test. Across the five infectious diseases, measures of Area Under the Curve (AUC) ranged from 0.92 to 0.99. The optimal cut-point ranged between 25 and 59 hours of co-presence. If the index test had been used real-time in this population, true positives could have been detected an average of one day earlier.These findings show that configuring electronic health data to monitor co-presence with individuals tested for infection would help predict subsequent nosocomial infection, and would

do so earlier than current standard of care.

4.2 Introduction

Nosocomial, or hospital-borne, infections are a burden on the health care system. Despite advancing medical technology and standards of care, the attributable costs of each case of nosocomial infection across the globe range from $2,992 to $29,000[170,171]. In the UK, each patient with a nosocomial infection costs an additional £3,154, for an estimated total of £930.62 million per year[172]. As well, nosocomial infections adversely impact patient health outcomes such as length of stay[173] and mortality[174].

Because the problem of nosocomial infection is well-known, mitigation has been approached from many different angles. These include preventing infection[175], reducing the time to identify an infection[176,177], and stopping subsequent outbreak once an infection is identified[178]. Preventing infection is generally preferable to treatment after infection[179]. However, preventing infection relies on changing norms and factors often out of the hospital's control since many infections occur within the community and are brought into the hospital. These community-acquired infection cannot be prevented within the hospital; only subsequent infections can be prevented. Stopping the spread of infectious disease relies on timely identification of an infection. Furthermore, there are rising concerns about the overuse of antibiotics and their subsequent effect on antibiotic resistance. Thus, this work focuses on early identification strategies that will provide opportunity for containing an infection and thus preventing its spread.

Current approaches to effective infection control in hospitals are limited by the time it takes to run microbial tests, and also the costs associated with these tests. Depending on hospital resources, most nosocomial infections are tested via a microbiological culture, (q-rt)PCR or BacLite *Rapid* MRSA[180]. All of these methods have downsides; bloodstream concentration of the vector can be below the detection

threshold, and microbiological culture in particular may require multiple days before a result is confirmed. Both of these limitations make controlling infectious outbreaks difficult. As a result, researchers have focused on other diagnostics that may indicate infection, such as biomarkers of immune activity or inflammation[181,182]. However, tests based on biomarkers also suffer from multiple downsides - they are costly, invasive, and have sensitivities and specificities of around 80%.

Other, non-biological methods exist for detecting infection. For example, contact tracing has been used in the past to find those most at risk of acquiring an infection. Contact tracing identifies those persons who have come in close physical proximity, that is co-presence, with others who have been infected[183,184]. This has been used in previous nosocomial outbreaks to determine an infection's source, and the path by which it spread through the hospital population[185]. Contact tracing is typically done retrospectively once an infection is confirmed[186]. Despite this use, the patterns of patients being near infected individuals has not been examined prospectively as a screening test *sui generis*.

To do so in real-time requires a fast method for tracking and linking patients, such as hospital administrative and electronic medical records and guidelines on when co-presence has reached a critical point such that infection is likely. Researchers have advocated for such novel uses of large administrative datasets, or "Big Data"[187]. These data sources are already used in health care for applications such as general infectious disease monitoring[188], population health surveillance[189], and hospital transfer networks[190]. EMR and HAD have the advantages of providing near-real-time information. Additionally, once the infrastructure is in place, the costs are minimal, both in terms of time and financial resources.

One way to address this challenge is to push for methods that allow more rapid testing of biomarkers. An entirely different approach, presented in this paper, seeks to use administrative and patient data generated as a byproduct of managing patient care. I show that an approach using patient administrative data can both be more

cost effective and speed up the time to diagnosis, thus decreasing the associated health burdens on hospitals. In this paper, I consider the interpersonal measure of co-presence with an infected patient as an indicator of infection risk. Specifically, I use the amount of time a patient is in the same bay of a hospital ward as patients who received a microbiological test as a screening test for nosocomial infection to identify thresholds after which infection is likely to occur. I use this for two reasons: 1) this is the information that would be available in real-time if this test were to be used prospectively, and 2) testing for infection is often indicative of a suspicion that the patient is infected. I will refer to this co-presence time as the "index test"[191]. Although similar to contact and link tracing, this approach is distinct in that it is ideally applied prospectively to predict infection, rather than retrospectively to follow the spread of an infection (although retrospective uses are also valid).

As an index test, co-presence with tested or diagnosed individuals would have the intended use of surveillance. As many hospitals already have administrative data and electronic medical records that could be monitored for patient-patient co-presence, implementing this index test would likely be inexpensive and efficient. The clinical role of this index test would be as screening; a result indicating likely infection would lead to additional tests or increased monitoring of those patients during the incubation period of the vector. Here I present evidence that the number of hours of co-presence with patients either receiving a microbiological test or diagnosis of infection is a strong screening test for nosocomial infection within UK hospitals.

4.3 Methods

4.3.1 Study design and population

The study population comprised all 133,304 patients with NHS hospital stays of at least 48 hours in a single county in the UK from 1 January, 2011 to 1 January, 2015.

I subset to 48 hour stays because I define a nosocomial infection as one occurring more than 48 hours after a patient enters the hospital[192]. These patients comprise a consecutive series, where the index test was assessed retrospectively. I assess the index test on the following infectious diseases: MRSA, *Clostridium difficile*, *Escheria coli*, *Pseudomonas aeruginosa*, and norovirus. This study follows STARD guidelines for the reporting of a new diagnostic[191]. Ethics committee approval was gained from the University of Oxford IRB.

4.3.2 Test methods

The reference test was either a diagnosis of the infection in question as assessed by ICD-10 code[101], or a positive microbiological test based on the test used by the NHS at the time for the disease in question, which was pre-specified. The hospital administrative data was not configured such that physicians or lab technicians could examine co-presence, and as such the index test results were not available to the performers of the reference test. There were no missing or indeterminate reference test results, and the data do not contain any reference to adverse events due to the reference test.

The index test was the number of hours a patient spent co-present with tested or diagnosed patients assessed at the time of microbiological testing or diagnosis, whichever came first. Co-presence was defined as the time both patients were in the same hospital ward bay. Clinical information and reference test results were available to performers of the index test. There were no missing or indeterminate index test results due to the administrative nature of the data; every patient's co-presence was precisely quantifiable. As the index test was done entirely *in silico*, there were no adverse events due to conducting the index test.

The index test of time co-present was based on tested or diagnosed patients rather than patients with confirmed infection to more accurately reproduce the

knowledge that would be available were this a prospective study. Because the test was assessed retrospectively, the data would allow us to perfectly calculate the hours of co-presence at the time of observation, but this scenario would not occur in practice. Tests for the presence of infectious vectors take time to return results, particularly microbiological cultures, where previous studies have shown the time for optimal results is five days[193]. Therefore, I reduced the information available to us when calculating the index test to accurately reflect what could be done in practice.

Index and reference test times were determined in the following manner: for patients with a confirmed infection, co-presence time was computed based on the time when the microbial test was collected or the diagnosis recorded in the EMR, whichever came first. For patients who were neither tested nor diagnosed, no corresponding time existed. For these patients, their time of assessment was chosen randomly from their hospital stay such that the distribution of times for patients with a negative reference test matched the distribution of times for those who had a positive reference test. This ensured that no systematic differences existed between those with positive or negative reference tests. To ensure I only assessed nosocomial infections, I exclude those whose test or time of assessment was within the first 48 hours of hospitalization[192].

Finally, the exact infectious period of tested or diagnosed persons was unknown. Co-presence is most meaningful if there is the potential for transmission of the infection, which only occurs during the infectious period. Therefore, I applied deterministic infection periods to model when patient-patient co-presence had the potential to transmit infection[194]. Patient's diagnosis or microbiological test time was considered the midpoint of their infectious period, with the length of their infectious period equal to literature values for the mode of the infectious period length (Table 4.1)[195,196,197,198,199]. The times surrounding a diagnosis or test were therefore what was considered salient for the purpose of patient-patient co-presence.

Disease	Mode Infectious period (hours)	Infectious period (range)
C. difficile[198]	60	24-96
E. coli[196]	120	80-160
MRSA[195]	72	48-96
Norovirus[199]	44	12-72
P. aeruginosa[197]	48	12-72

Table 4.1: Values for the infectious period for each infectious disease or control condition used. Microbiological tests were assumed to occur at the midpoint of the infectious period.

4.3.3 Analysis

To compare the measures of diagnostic accuracy, I use the Receiver-Operator Characteristic (ROC) curve and the area under the curve (AUC). I calculated the 95% confidence intervals of the AUC using bootstrapping. This is a conservative approach which generally increases the width of the confidence interval to better-estimate out-of-sample performance. To determine the optimal cut-point for these curves, I assume that the clinical costs of false positives, false negatives, true positives and true negatives are all equal. Following this, the optimal cut-point is the point closest to a perfect test (100% sensitivity and specificity) in Euclidian space. I assess the sensitivity and specificity at optimal cut-points.

To quantify the effectiveness of the index test, I calculated the number of hours between the time a patient's microbiological test was administered and when they first crossed the cut-point of co-presence during their hospital stay. I also calculated the number of hours between when a patient's infectious period began and when they first crossed the cut-point of co-presence during their hospital stay. I take the minimum of these two numbers, which represents the number of hours earlier the infectious disease could be detected if this method were implemented relative to the current standard operating procedure.

Finally, because of the number of assumptions made in assessing the index test of co-presence, I conduct sensitivity analyses on the data. To determine if the results generalize to hospitals with only ward-level co-presence data (rather

than bay), I rerun the analysis at the ward level. To determine whether the use of the mode for infectious period length unduly affected the results, I also ran the analysis using the literature values for minimum and maximum infectious period lengths. Finally, it is possible that any significant predictive power of co-presence is because of the presence of a quarantine ward. If patients are moved into a quarantine ward due to suspected infection, then hours of co-presence with infected or tested individuals would be artificially high. I test this by also removing the quarantine ward from the analysis.

4.4 Results

The patients included in the study, their reasons for exclusion, and their reference tests can be seen in Figure 4.1. I observe that most patients were excluded due to being inpatients or being tested for microbiological vectors within the first 48 hours.

This left the study population of patients who could contract nosocomial infection. The demographics of these patients are shown in Table 4.2. Patients were on average 56 years old, and 45% were male. On average, these patients spent 13 days in the hospital, and 5.40% of them died while in the hospital. In total, 8,684 (6.51%) patients were infected with one of the five nosocomial infections studied.

After applying the index test to this set of patients, I observed very distinct distributions of co-presence time with diagnosed or tested patients for those whose reference test was negative compared to those whose reference test was positive (Figure 4.2). Irrespective of a specific cutpoint, the distributions of co-presence times were strongly differentiated based on whether or not a patient had a positive reference test (infectious disease test or diagnosis). For all five infections, patients with a negative reference test had a distribution of co-presence times (index test results) much lower than for patients with a positive reference test.

To quantify the performance of the index test, I created receiver operator

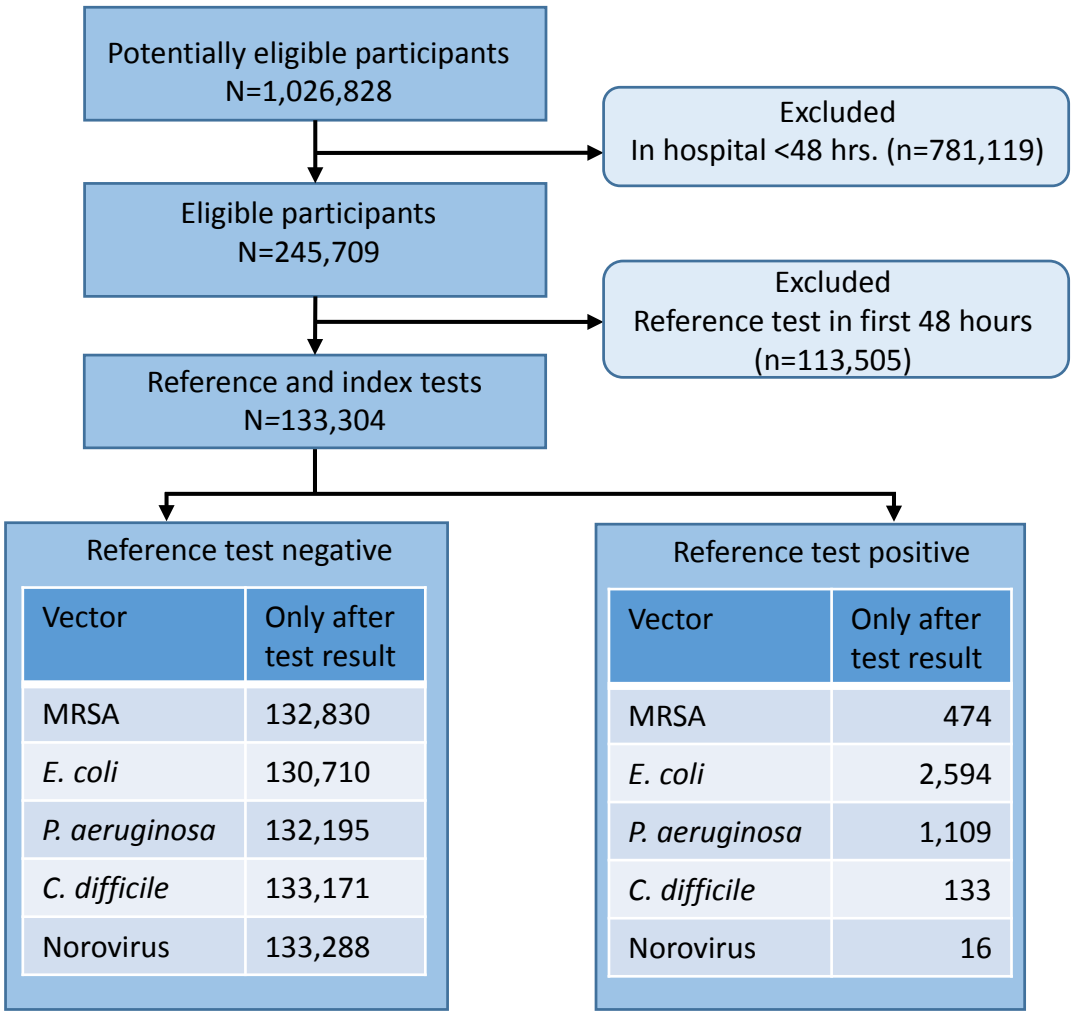

Figure 4.1: Patient flow diagram. Number of excluded eligible patients differs by infectious disease because different numbers of patients had their reference test within the first 48 hours of their stay, and therefore were likely not nosocomial infections. The number of eligible participants excluded differs between infectious diseases because different sets of patients had their reference and index tests within the first 48 hours of their hospital stay. Importantly, the different populations for each infectious disease are not exclusive; each patient is in all five population, and only their results on the reference and index tests change.

Variable **Mean (SD) or N (%)**	**Mean (SD) or N (%)**
Age (years)	56.4 (27.8)
Sex (male)	59,988 (44.80%)
Stay length (hours)	319.5 (568.8)
Died in hospital	7,180 (5.40%)
Infected with *C. difficle*	3,192 (2.39%)
Infected with *E. coli*	3,501 (2.63%)
Infected with MRSA	665 (0.50%)
Infected with norovirus	22 (0.02%)
Infected with *P. aeruginosa*	1,304 (0.98%)

Table 4.2: Baseline demographics and clinical characteristics of patients based on the set of patients who received both a reference and an index test for the nosocomial infection in question.

characteristic (ROC) curves based on multiple cutoffs of hours co-presence with infected individuals. The ROC curves have areas under the curve (AUC) ranging from 0.92 to 0.99 (Table 4.3). The optimal cutpoint ranges from 29 to 59 hours depending on the infection in question. This means patients must spend over 24 hours co-present with infected patients before the number of false negatives is minimized. Of note is that these hours can be accrued concurrently as a patient can be co-present with multiple tested or diagnosed patients simultaneously. Sensitivities and specificities at these cutpoints are also shown in Table 4.3. Sensitivities are all at least 0.95, and sensitivities are all at least 0.90. These results are robust to analyses conducted at the ward level rather than the bay level, decreasing by a maximum of 10%. Moreover, the results are largely unchanged when the quarantine ward is removed from the analysis, as tests or diagnoses are often received prior to transfer to the quarantine ward. The hours of co-presence leading up to the receipt of test or diagnosis are therefore not affected by the co-presence in the quarantine ward which follows.

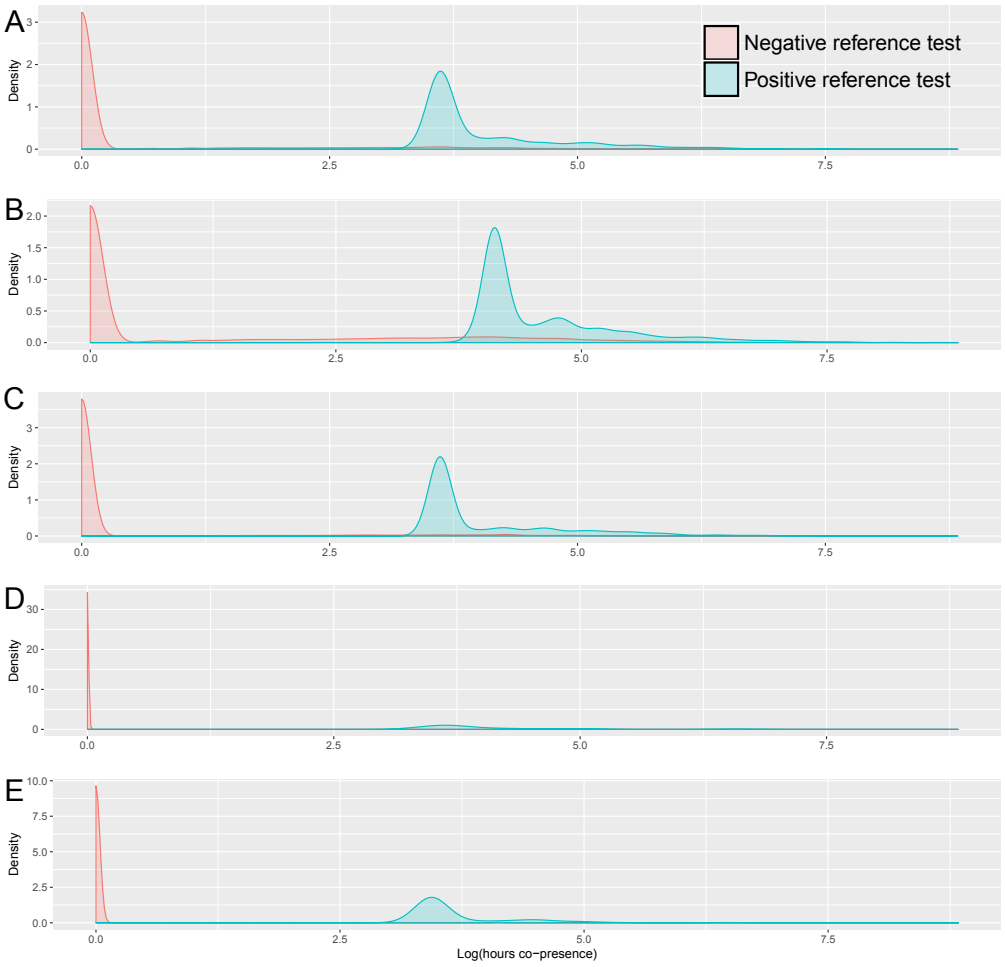

Figure 4.2: Empirical probability density functions of hours of co-presence with infected individuals (index test) stratified by the presence of a diagnosis or positive microbiological test (reference test). Each panel represents one of the communicable diseases tested: A) *C. difficile*, B) *E. coli*, C) MRSA, D) *P. aeruginosa*, and E) Norovirus.

Disease	AUC (95% CI)	Threshold (hours)	Sensitivity	Specificity	True positives	Average hours saved per patient (range)
C. difficile	0.993 (0.992,0.994)	29.00	1.00	0.99	133	6.36 (4.08,10.14)
E. coli	0.966 (0.965,0.967)	59.00	0.95	0.90	2,472	8.35 (4.91,13.61)
MRSA	0.962 (0.96,0.964)	35.00	1.00	0.95	474	10.93 (7.05,16.18)
Norovirus	1 (1,1)	34.00	1.00	1.00	16	8.19 (5.12,10.31)
P. aeruginosa	0.925 (0.923,0.927)	35.00	1.00	0.95	1,107	21.97 (13.98,32.96)

Table 4.3: Index test statistics for all five diseases. The threshold, or optimal cutpoint, for each test was the number of hours of co-presence that gave sensitivities and specificities which were closest in Euclidian space to the optimal test. Sensitivities, specificities, and the number of true positives were taken at these optimal cutpoints. Finally, hours saved is the difference in time between when a patient first crosses the threshold of the index test and when they were actually tested for or diagnosed with the infection. This number then represents how much earlier a patient may be screened for infection when using the index test than when using the reference test. Ranges indicate the minimum and maximum numbers when infectious period lengths were stochastic rather than deterministic.

For patients with a positive reference test and a positive index test (true positives), I examined how many hours earlier they would have been tested if the reference test was administered immediately upon crossing the threshold of co-presence (Table 4.3). I observe that on average, the amount of time saved ranges from 6 hours for *C. difficile* to 22 hours for *P. aeruginosa*. The ranges show that even when the infectious period is changed, the results remain qualitatively the same.

4.5 Discussion

In this paper, I have shown that the number of hours of co-presence with tested or diagnosed individuals serves as a strong predictor of infection. Further, I show that the optimal cut-point for all four diseases tested is greater than 24 hours. This is important, as previous methods, such as contact tracing, typically uses any contact vs none (or no co-presence vs. any) as a cutoff for potential infection. Instead, I are able to use a data-driven approach to determine the quantity of hours co-present with a patient suspected of infection which maximally predicts infection. Additionally, I show that if co-presence time is used as a screening test, patients' infections may be detected as much as 22 hours earlier, on average. Importantly, I do so using only electronic medical records which are constantly updated in near real-time within hospitals using such systems. In quantifying this test, I hope that it will potentially be used as a screening tool in hospitals, and that it will foster the use of electronic medical records for other purposes. Thus, informatic systems can be designed based on this work to identify those patients who may be at risk of infection due to time co-present with those suspected of infection.

The index test of patient-patient co-presence would ideally be used as a screening test; once a patient was co-present with a tested or diagnosed patient for at least the cutoff time of 29-59 hours, they would then be tested for the microbiological agent and subsequently monitored for signs of infection. I have shown that if

this were done, infections could be identified earlier than on current standard operating procedure. This earlier detection may lead to reduced infectious periods for patients, as treatment could be administered sooner. Further, if there were downstream effects, the values I estimate in Table 4.3 may be underestimates. For instance, if earlier detection of a patient's infection via this index test resulted in earlier quarantine for that patient, then this would prevent other patients being infected via co-presence with the original patient. These downstream effects are not captured in the calculations of the person-hours of infection potentially saved, and would require simulation studies to evaluate this counterfactual, which is beyond the scope of this paper.

Although I omitted quantifying the cost of various outcomes, I recognize that this is an important factor in the design and implementation of new tests, both for understanding the impact the test may have, but also for deciding the ideal cut-point. Here, I used the point closest in Euclidian space to the perfect test, as a starting point for optimal cut-points. If the cost of a false positive based on co-presence is too high (i.e. the cost of time, money, and space for microbiological specimens), the optimal cut-point can be adjusted to better reflect the costs of false positives and false negatives. This will also influence whether the use of this screening test is cost effective in practice - given the large number of false positives, testing all patients who are co-present with tested individuals may incur large costs, making this test inefficient for rare infections. To clarify, although the specificities across infectious diseases are very high, the very large number of patients without infection means even these high sensitivities translates to hundreds or thousands of false positives. In addition to false positives, diseases like MRSA which are closely-surveilled in hospitals may not see dramatic increases in the early detection rate. Even though the index test correctly identifies almost 500 MRSA-infected patients, only eight hours per patient are saved. However, for common infections that are not as closely surveilled, this test would have myriad benefits both in terms

of health outcomes and costs, even with the cost of false positives.

As previously stated, this test has many advantages. First, many hospitals already use some form of electronic medical record, so adapting them to monitor co-presence with infected and tested individuals should carry minimal effort. These forms of data can also be passively monitored for relevant amounts of co-presence, meaning the cost to perform the index test is inexpensive once the passive monitoring is enabled. This test is also fast; a result is returned immediately when a patient crosses the threshold of co-presence. Finally, this test is specific to the infection being examined. A positive result of co-presence with patients tested for *E. coli* only strongly predicts an *E. coli infection*. Because the co-presence is only counted with patients who have a specific disease, this test specifically identifies the infection in question. This is in contrast to some diagnostic tests which measure general indicators of infection, and must be used in concert with physician expertise to identify the specific infection[182]. All of these strengths indicate that this test would strongly supplement the tests currently available.

Because the index test is based on administrative data, there are inherent limitations that make this approach imperfect. First, this method cannot disentangle disease-specific modes of transmission, and does not perfectly capture all the methods by which infectious diseases can transmit. A vector may transfer directly from patient to patient, or may be transmitted between the two by a third non-patient party, such as a health care practitioner. There may be other methods of transmission not from person-to-person that would not be predicted via co-presence (e.g. residual vector on a surface). The co-presence test makes no assumption on the method of transmission, and instead leverages the increased risk of infection for patients who are co-present with an infected patient. These results therefore show the strength of association between co-presence with *other patients* and subsequent infection, and it is this association which can be leveraged for screening.

If this method were to capture transmission through third-parties, every patient

contact with said third party would need to be included, which is outside the realm of most hospital administrative datasets. Including such third-parties would limit the general applicability of this method. Further, these results are based on the relatively crude measure of co-presence, and therefore likely represent the minimum strength of the association; increased sophistication (e.g. bed-level information) can increase the association between co-presence and infection, and therefore the predictive power of this test. I also make assumptions in creating the index test, which could impact the results. These include assuming all infected patients receive the reference test at the midpoint of their infectious period, and that everyone has an equal-length infectious period. However, I apply the same assumptions to all the patients, so there should be no differential effect between infected and uninfected patients. Finally, this test was quantified using data from NHS hospitals. Standard operating procedure for infection control exists that may make the results here non-generalizable. However, future work should be done to examine whether similar results occur elsewhere.

In this paper, I have shown that using electronic medical record-based co-presence time with patients tested for a microbiological agent is a strong candidate as an indicator of nosocomial infection. Beyond the implications for nosocomial spread, this suggests that co-presence in hospitals matters: patients are not truly isolated and independent from one another, and this needs to be recognized, and leveraged for better health care. Although these results are a strong starting point for using co-presence as a screening test in the hospital setting, more work needs to be done to validate and strengthen these findings. Co-presence as a screening test should be evaluated prospectively in a hospital, which will then even more accurately reflect what health care practitioners see in real-time, rather than the retrospective data I use here. Additionally, the potential benefits and costs of using this as a screening test are not clear; the downstream benefits are complicated and may be stronger than indicated here. On the balance of the strengths and

potential caveats discussed herein, I have shown that co-presence is a powerful indicator of nosocomial infection, which merits further study and may have benefits towards reducing nosocomial spread in hospitals.

5

Using patient-patient co-presence to detect subclinical nosocomial infections

Contents

5.1 Abstract

Subclinical infections, those where the bacterial or viral load are below a test's detection threshold, likely have negative affects on patient outcomes, and affect the disease dynamics of nosocomial outbreaks. However, given the difficulty to detect subclinical infections, their presence and potential impact has gone unaddressed. I use a random forest model to separate infected and uninfected individuals based on electronic records and hospital administrative data. Patients whose classification is switched from uninfected to infected are considered subclinically-infected. I perform cross-validation on infected patients to determine the model's accuracy. Using regression models, I estimate the impact of subclinical infections on hospital stay length and likelihood of death. Finally, I determine the effect of subclinical infections on disease dynamics by observing both connections in the static network, and disease models on the temporal network. Of 82,711 patients, the model identifies 183 with subclinical MRSA, 136 with subclinical *C. difficile*, and two with subclinical norovirus. The model detects between 75% and 100% of known infections in cross-validation. Subclinical infections negatively impact both hospital stay length and likelihood of death during a hospital stay, but to a lesser extent than full infection. Subclinical infections of MRSA and *C. difficile* likely impact nosocomial dynamics by connecting groups of otherwise-unexposed patients to the infection in question. However, timely detection and quarantine of subclinical infections could reduce the impact of nosocomial outbreaks. I find strong evidence

that the model detects subclinical patients, and that these subclinical infections negatively impact both individual patients, and those with whom they share the hospital. Future studies should examine this prospectively to determine the efficacy of this model in real-time, and whether the simulated intervention would indeed reduce the impact of nosocomial outbreaks.

5.2 Introduction

Nosocomial infections burden the health care system, each costing an average of £3,154, for an estimated total of £930.62 million per year[172], and this impact has been increasing over time[200,201]. Because of this importance, reducing the impact of nosocomial infections has been an important research aim and clinical care focus, with a variety of approaches. One important aspect of reducing nosocomial infections is having effective diagnostic tests for the disease in question, as faster, more accurate diagnostic tests can lead to reduced infectious disease spread. Standard-of-care tests[202] and other methods, such as biomarker tests[203] do not detect all infections because either the timing of the test is incorrect (e.g. a patient is tested just before or after infection) or the test is not sensitive enough to detect a low viral or bacterial load. Therefore, patients with a minor infection which cannot be detected by diagnostic tests may go unnoticed by the health care system. I term such patients subclinically-infected, and my main aim in this paper is to identify subclinically-infected patients, evaluate whether those patients determined to be subclinically infected impact the health of other patients and, thus, contribute to nosocomial outbreak dynamics.

The magnitude of this problem is currently unclear. Because subclinical infections are by definition undetectable, there is no way of knowing the impact these infections may have on patient health and health care outcomes. However, it is likely a problem, as research in other fields such as infectious disease epidemiology

and immunology have shown. In studying bovine tuberculosis, researchers are aware that there is a point during the progression of the disease where a cow is infected but not yet test-sensitive, often building this stage into their disease models[204,205]. Importantly, including this pre-detection stage of infection in models of disease dynamics improves accuracy of observed infection rates, indicating that subclinical infections impact disease dynamics. Similarly, it is known that microbiological testing in nosocomial infection does not have a 100% sensitivity, and naming the stage at which someone is infected but does not yet test positive on an infectious disease test "subclinical infection" makes this phenomenon explicit and provides definitional clarity[206,207].

Subclinically-infected patients are distinct from colonized patients. Those with subclinical infections have a bloodstream infection, but their infection neither reaches detectable levels nor results in any symptoms. This may be due to increased innate immunity, receiving only a small initial amount of infectious vector, or some other reason. Colonized patients, on the other hand, have an infectious vector somewhere on their person that enables transmission and puts themselves at increased risk of infection, but they do not have infectious vector in their bloodstream[208]. These differences are important, as the results I present are specific to subclinically-infected patients rather than colonized patients.

Despite being distinct, there are a number of similarities between subclinically-infected and colonized patients that make understanding subclinical infections easier. For instance, patients who have been colonized by an infectious disease may transmit the infection to others without showing signs themselves[209,210]. Similarly, subclinically-infected patients may themselves be infectious, leading to infections in others. Subclinically-infected patients may also be able to transmit to more patients than those with overt infections, as they are not removed to a quarantine ward after having their infection confirmed[211]. This is because they, like colonized patients, are not easily detected by standard methods. Because of this, the presence

of subclinically-infected patients may alter the dynamics of nosocomial spread.

Recent improvements in our understanding of colonized patients have led to better protocols aimed at reducing the impact thereof, which can partially be applied to subclinical infections as well. One main method is active surveillance, where all patients are screened for colonization on entrance to the hospital, and can reduce nosocomial infection by up to 39%[177]. This allows physicians to catch sources of infection that do not otherwise manifest. However, active surveillance would only limit transmission of those entering the ward colonized, missing nosocomial subclinically-infected patients[185]. As patient who eventually develop subclinical infections do not enter the ward infected, and never reach detectable levels of infectious vector, they would be detected by neither active surveillance nor standard microbiological testing.

Although active surveillance itself cannot be used to identify subclinical patients, incorporating both HAD and EMR can lead to detection of nosocomial subclinical infections. HAD is any information collected on patients for the purposes of monitoring their progression through the health care system[212]. These data are commonly collected on patients, but are not often used for diagnostic purposes. EMR contain the information on patients specific to their health and treatment, and are often used for medical decision-making[213]. However, constantly monitoring these data sources on a patient while they remain in the health care system can shed light on the status of their immune functioning, and lead to the inference of infectious vector in their system despite an inability to directly detect it.

EMR data include biomarkers, which are often used in testing the status of the immune system. Infection has pronounced effects on many biomarkers, which can be leveraged to detect infection[203,214]. Biomarker tests are routinely done in patients even when infection is not suspected, particularly with the increasing utilization of active surveillance. Here I propose to use these data to detect potential subclinical infection.

There is additional information that can be used to predict infection in HAD. As I have shown previously, an important indicator that predicts infection is co-presence with patients tested for infection 4. Researchers have also used patient-patient co-presence and administrative data to to trace outbreaks through hospital populations[185]. Because co-presence strongly predicts subsequent infection, it also likely predicts subclinical infection, as subclinical infection is a necessary step towards overt infection. Co-presence then, is an additional index that can improve our ability to identify patients who are likely subclinically infected.

Therefore, using biomarkers, co-presence with infected patients, and additional information from the HAD and EMR, I can likely detect individuals with subclinical infections who would otherwise go unnoticed via standard disease surveillance. Individuals in the subclinically-infected phase will demonstrate low bacterial or viral load, but will likely have biomarker indicators that are elevated, placing them somewhere between those who are uninfected and those with confirmed infections. Thus, I hypothesize that subclinically-infected patients will exhibit biomarkers more similar to those with overt infection and will have time co-present with patients who have confirmed infections. My aims in this paper are therefore fourfold. I will a) use machine learning classification methods to determine patients with likely subclinical infectious diseases b) examine the validity of this method and the identified patients c) determine the individual health effects of having a subclinical infection, and d) understand how the presence of subclinically-infected patients may impact nosocomial disease dynamics.

5.3 Methods

5.3.1 Data source

The data come from the Infections in Oxfordshire Research Database, which comprises standard National Health Service (NHS) administrative data and electronic medical records. This data set was originally established to monitor infectious diseases, and also contains complete individual health records. The data for the analysis comprises all 82,725 patients in the health care system from Jan 1, 2011 to Jan 1, 2015 in the hospital for at least 48 hours, and with at least one biomarker test during their stay. I subset to at least 48 hours to ensure I only identify nosocomial infections, rather than community-acquired infections[192].

The data contain information on microbiological and biomarker tests. For microbiological tests, each test is timestamped and includes information about the tested vector, whether it was detected, and any antibiotic resistance detected. Biomarker tests include the timestamp and the corresponding value of the test (with units). Records without any biomarker tests (n=50,577) do not have sufficient meaningful information to be used for classification, and so are removed from the analysis. Each patient's medical history is broken into individual visitation stays in the hospital with time stamps for both entry and exit. A stay is defined as every occurrence with unique entry and exit times of any patient into a hospital. Stays are broken down into spells, which are every unique period of time in a given ward with entry and exit times. The date of death is recorded for each individual through 2015. Specific wards are also divided into ward bays, which provide additional spatial precision about patient locations. Additionally, the dataset includes information about diagnoses based on ICD-10 code and admitting physician[101]. The ICD-10 codes are used as a proxy for diagnoses, which has been shown to have good sensitivity and specificity[102].

In this paper, I examine three infectious diseases which cover the range of

nosocomial illnesses: *C. difficile*, MRSA, and norovirus. This includes bacterial and viral infections, infections that transfer through multiple mechanisms, and have different standard operating procedures for diagnosis, treatment, and containment. All of these factors may influence the performance of the method, so I gain additional understanding of the method and of subclinical infections in doing so. For more information on the dataset, see Chapter 2.

5.3.2 Random forest classification analysis

To identify patients who have a subclinical infection, I use random forest models. A random forest is a machine learning method for grouping observations based on patterns of covariates. Outside of neural networks, random forests are often one of the best machine learning methods that do not over-fit to the data at hand, yet retain significant predictive power[215]. They are also able to non-parametrically fit non-linearities without prespecification. Because the true status of patients isn't known (as there is no detection method for subclinical infection), the models cannot be supervised. However, I initially group patients according to their infection status, and in this way, the models are semi-supervised[216]. Models were fit allowing for 2 classes (initialized as uninfected and confirmed infected based on microbiological test or diagnosis).

The variables I chose to include in the models were those that would indicate a potential infection earlier than the markers used for infectious vector tests. Current tests measure either the bacterial/viral load directly or measure the specific antibodies to those infections. Both of these only occur relatively late into an infection (Figure 5.1). I therefore use variables relating to the cell-mediated immune response (CMI), which includes many markers of general infection or swelling. These are eosinophils, neutrophils, white blood cells, and C-reactive protein (CRP). Based on the work in Chapter 4, I include as a predictor the

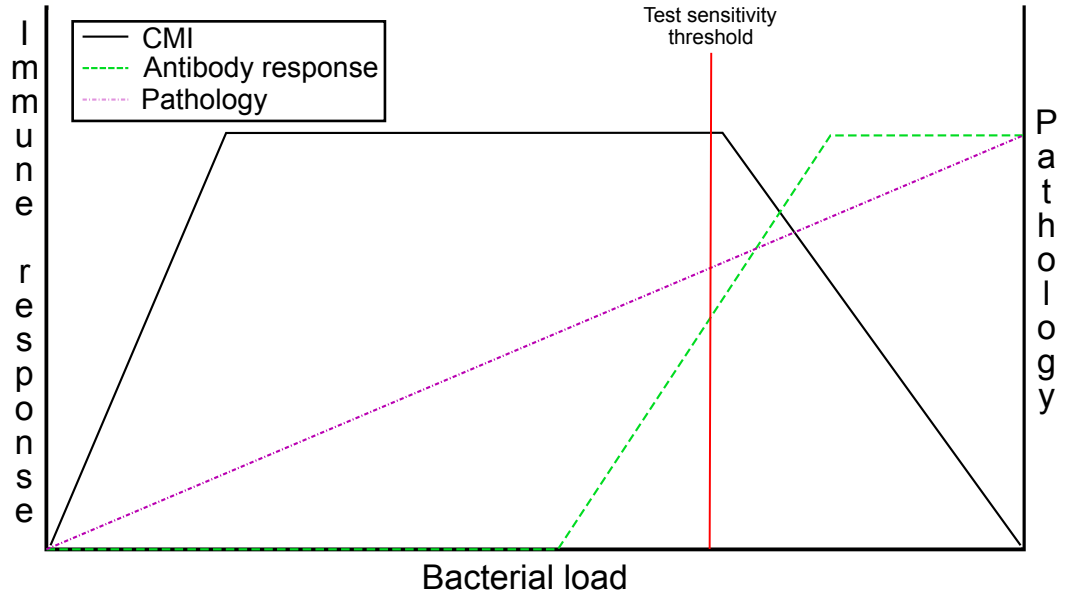

Figure 5.1: General progression of the immune response. Standard infectious disease tests either measure directly the presence of bacteria or virus, or specific antibodies to those vectors, which both occur relatively late in the infection. However, the cell-mediated immune response (CMI) occurs relatively early, allowing for early detection of subclinical infection. Figure adapted from Pollock and Neill [204].

amount of time a patient spent with other infected patients in the ward, as this gives an early indicator of subsequent infection. I also include whether a patient had any other infectious disease diagnoses during their stay. This will prevent patients who have some other infection from appearing subclinical via this method. Finally, I include demographic information including sex, age, hospital stay start time (to allow for time trends), primary diagnosis, and admitting physician. After models were run, some patients who were initially classified as uninfected were instead classified as infected. These are the patients classified as subclinically-infected, as their characteristics appear similar to those with overt infection.

5.3.3 Validation

In addition to being difficult to detect, the presence or absence of subclinical infection cannot be verified, especially retrospectively. Because there is no gold standard

detection method for subclinical infection, no identification method can conclusively determine a patient's subclinical infection status. Previous work has discussed the practice of classification models when no reference standard exists[217,218,219]. This research advocates examining external factors or outcomes that would be affected by the results of the model, but are not part of the model. It also stresses the importance of standard methods such as cross-validation. I use both approaches here.

To assess the internal validity of the findings, I first conducted a cross-validation. To do so, I randomly initially classified 10% of infected patients as uninfected. I then re-ran the random forest model, and observed how many of the known infected patients were recovered and properly classified as infected. I repeated this simulation 100 times to observe the variability in the process.

For external validation, I fit models for two health outcomes not used in the classification process: hospital stay time, and survival during the hospital stay. These are both outcomes in which infectious diseases result in worse outcomes[220,221]. Therefore, if the random forest model has accurately identified individuals with subclinical infection, models using the classification from the random forest model as a predictor of health outcomes will have improved fit indices relative to using only microbiological test and diagnosis information. For hospital stay time, I fit a linear regression model. For dying while in the hospital, I fit a logistic regression. In both cases, I compare the model fit (as measured by the BIC) when using only infection information from microbiological tests and the model fit when using the latent class membership. All models were adjusted for age, sex, time of hospital stay beginning, primary diagnosis and admission consultant. I also attempted to do the same for the time to readmission, conditional on the patient surviving through their hospital stay. However, due to the reduced sample sizes, the fitted accelerated failure time models did not converge.

5.3.4 Health outcomes

Although I group those I believe have subclinical infection with observed infected patients for the purpose of validation, I hypothesize that the outcomes for those in the subclinically infected phase will lie somewhere between those who are not infected and those with confirmed infections. Therefore, I fit the same models I fit previously (linear regression for hospital stay time and logistic regression for death during hospital stay), but instead of only using the two classes, I use a categorical variable with three values: uninfected, subclinical, and infected. All models were adjusted for age, sex, time of hospital stay beginning, primary diagnosis and admission physician. The baseline in all models is the uninfected class, and I expect that for each model, subclinical patients will have outcomes worse than uninfected patients, but better than infected patients.

5.3.5 Disease dynamics

In addition to understanding how subclinical infections may impact individual health, it is also important to understanding how they may impact the dynamics of nosocomial infection. To do this, I construct the patient-patient co-presence network, where each patient is represented as a node, and nodes are connected by edges if the patients are in the same ward bay for at least an hour together (see Chapter 2). Due to the small number of subclinical infections observed for norovirus, I subset this to only MRSA and *C. difficile*.

I start by dividing the observed infections into individual outbreaks. An outbreak was defined as all infections occurring with less than 96 hours between consecutive infections. This period is greater than the incubation phase for the majority of cases, so any exposed individuals would have likely transitioned to infected or subclinically infected within this time frame. In creating outbreaks in this manner, I separate out putatively-independent occurrences of infection, allowing me to

study each in isolation. This also gives a population of outbreaks, against which simulations can be compared, rather than to just an n of one situation if I examined the entirety of the data at once.

To understand how subclinical infections affect the spread of infection on this network I build statistical models for the occurrence of configurations of sets of three patients each of whom may be uninfected, infected, or subclinically infected. I can represent these configurations as colored triads, with three different colors corresponding to the respective infection status of each patient, and then apply a colored triad census to each recorded outbreak (Appendix A). I calculate the colored triad census for each outbreak individually. This census counts the number of uniquely colored and structured triads in the network. By coloring the nodes according to infection status (uninfected, infected, or subclinically infected), there are 56 unique colored triads.

I construct a baseline, or null, model by fitting an Exponential Random Graph Model (ERGM) to the network with terms for *edges*, *geometrically-weighted edgewise shared partners*, and *nodemix*. These terms model the number of edges, number of times a pair of connected nodes both connect to a common third node, and the number of times a node of a given attribute connects to a node with a given attribute (may be the same or different), respectively. In sum, these model the density, triadic structure, and attribute mixing of the network, all the components of the colored triad census. Any colored triad counts which significantly differ from the distribution implied by this model therefore indicate an interplay between these factors not modeled by the ERGM. Again, I do this to the network for each outbreak. I simulate 1,000 networks from each ERGM, and calculate the colored triad census for all of them, summing the counts across outbreaks, and comparing the observed counts of the colored triads to the null distribution. The percentile of the observed count is a pseudo-p-value. Any colored triads which are significantly over- or under-observed in this model indicate colored triads where the structure

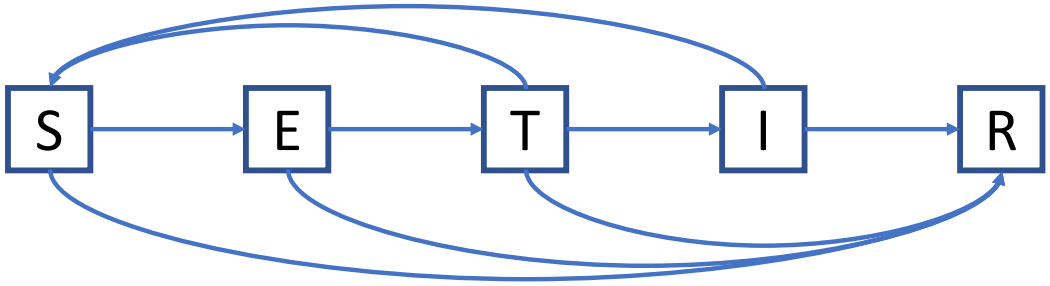

Figure 5.2: Dynamic model schematic. The five stages are Susceptible, Exposed, Subclinically-infected (Test insensitive), Infected, and Recovered. Arrows directly to and from the recovered category are based on the empirical data of patients entering and leaving the hospital rather than governed by any parameters.

and color of the nodes interacts to alter the propensity to connect.

Using knowledge gained from the colored triad census, I move to the temporal network. Although the static network leads to important insights, edges between patients are only temporarily extant, particularly with respect to the lifetime of the infectious vectors, and as such, a temporal network is the appropriate model for fuller analyses[60,222]. I therefore fit a dynamic infectious disease model to the temporal network. This model has five categories: uninfected, exposed, subclinically-infected, infected, and removed (Figure 5.2). Patients can move from uninfected to exposed based on their exposure to subclinically-infected and infected patients. Once exposed, they transition through to either subclinically-infected or infected at exponentially-distributed rates. Whether they become infected or subclinically-infected is determined by the observed fraction of infected and subclinically-infected patients in the data.

A patient can move to the recovered stage at any time, based on the underlying temporal network (e.g. when a patient's stay ends through discharge or death). Additionally, patients who are infected or subclinically-infected can also clear the infection while at the hospital, becoming susceptible again. I also model the hospital's quarantine policy by artificially increasing the recovery rate of infected individuals; with respect to infectiousness, they "recover" once they are moved to

the quarantine ward. For this reason, the duration of infectiousness is lower for infected than for subclinical, although the likelihood of transmitting infection in a given hour is much higher for infected patients. Becoming infected is proportional to the person-hours one spends with infected and subclinical patients. I reduce the average length of infection and the infectiousness over time as the hospital increases its response to the outbreak. This way, simulated outbreaks die out within the time of each observed outbreak, reflecting reality.

I first calibrate the model by using the observed data from the outbreaks, and I do so separately for the MRSA and the *C. difficile* models. Using the observed outbreaks without the subclinically-infected patients, I estimate the average time exposed, the average time of infection, and the probability of exposure given co-presence with an infected patient. I then add in the subclinically-infected patients, and recalibrate the model, adding in the parameters for average length of being subclinically-infected, and for the probability of exposure given co-presence with a subclinically-infected patient. I do these calibrations by simulating an outbreak on each outbreak network 100 times with a given set of parameters. The objective function to minimize was the Euclidian distance between the 3-dimensional set of median outbreak size, mean outbreak size, and Fano factor (variance divided by mean, which gives an estimate of the "burstiness" of the outbreak)[223]. I explored the parameter space via a Metropolis-Hastings algorithm[224].

With the calibrated models, I simulate an intervention on the subclinically-infected patients. I rerun the simulation after reducing the infectious period of subclinical patients so that it mirrors that of infected patients comparing this to the results of the simulations without the intervention imposed. If the intervention has an effect, then the intervention would ameliorate the effects of infection spread via subclinically-infected patients.

5.4 Results

The study population consisted of 82,711 individuals who were on average 57 years old, and 46% of them were male (Table 5.1). The average stay length was 389 hours, and the most common primary diagnosis was hypertension, with 2,396 cases (2.9%). There were 977 unique diagnoses. 911 individuals tested positive for MRSA, 1,920 patients tested positive for *C. difficile*, and 31 individuals tested positive for norovirus.

Variable	**Mean (SD) or N (%)** (n=82,711)
Age (years)	57.42 (25.17)
Sex (male)	1,579 (46.0%)
Stay length (hours)	388.88 (714.27)
Patients with MRSA	911 (1.10%)
Patients with *C. difficile*	1,920 (2.27%)
Patients with norovirus	31 (0.04%)

Table 5.1: Demographic information about the study population.

5.4.1 Identification and validation

Based on the random forest classification, there are a total of 183 patients subclinically infected with MRSA, 136 with *C. difficile*, and 2 with norovirus. The different covariates used had differential strength in classification, with co-presence leading to the strongest separation between the latent classes (Figure 5.3). Of the biomarkers, C-reactive protein and neutrophils best differentiated between uninfected and latent classes.

The cross-validation reliably picked up at least 80% of known infected cases across all three diseases 5.2. The recovery was best for *C. difficile*, and worst for norovirus. This was largely due to the small numbers involved in predicting norovirus cases.

The results showed that the models had better fit when considering the subclinical patients as infected rather than uninfected (Table 5.3). Additionally, patients with

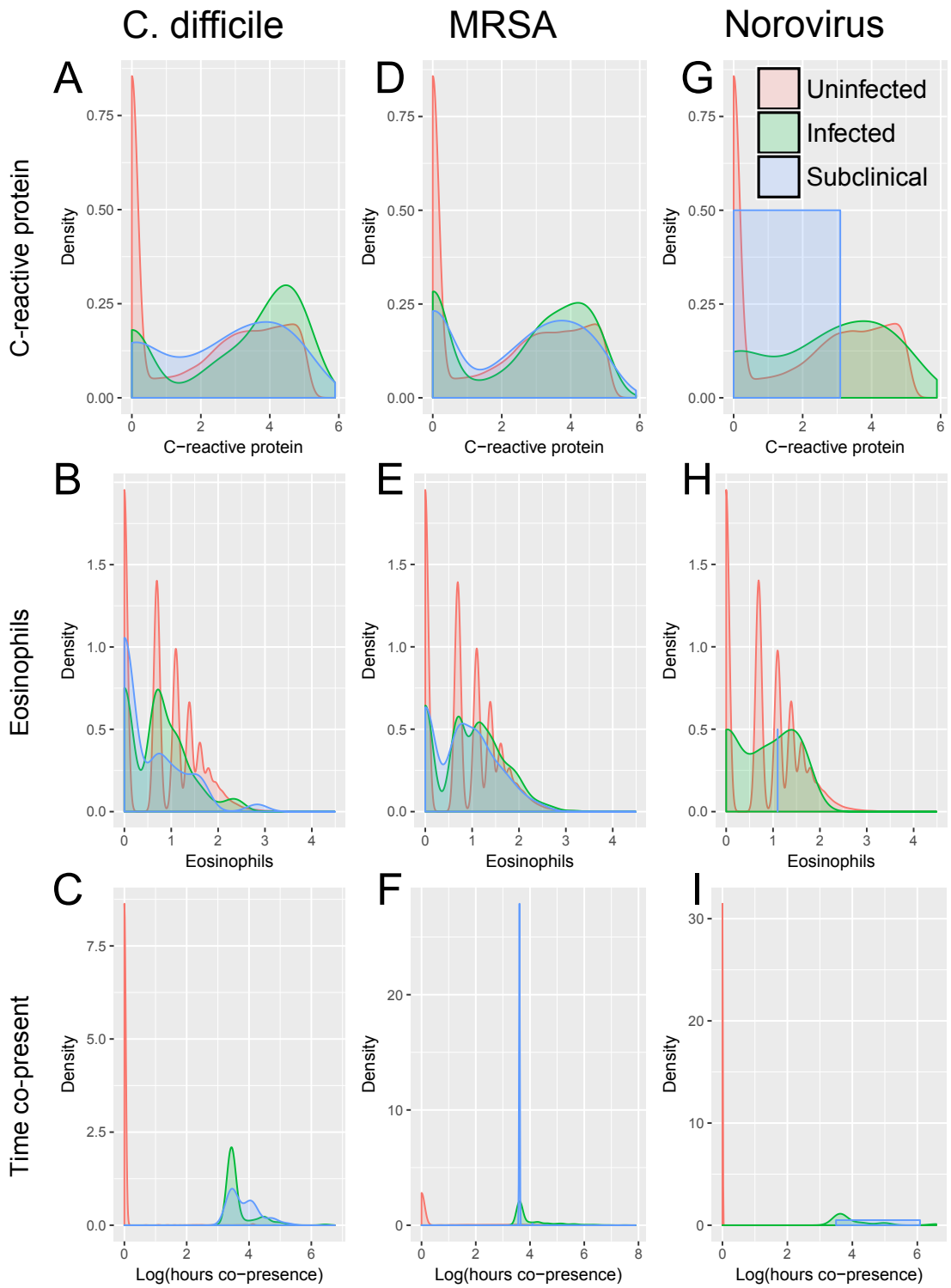

Figure 5.3: Probability density function plots of biomarkers and overlap-hours with infected patients divided by latent class status. A, B, and C show C-reactive protein, eosinophils, and time spent co-present with infected individuals, respectively for *C. difficile*. D, E, and F show the same for MRSA, and G, H, and I show the same for norovirus.

Disease	Median % correctly identified (Range)
MRSA	0.86 (0.80,0.95)
C. difficile	0.90 (0.84,1.00)
Norovirus	0.81 (0.75,0.89)

Table 5.2: Cross-validation results. For each of 1,000 trials, 10% of infected patients were randomly reclassified as uninfected. The percentage of these that were recaptured as infected by the random forest was recorded.

a subclinical infection on average had worse outcomes than uninfected patients, but better outcomes than infected patients. Although the models always had higher BICs using the latent class membership, the magnitude by which the fit increased was greatest for *C. difficile*.

Infection	Outcome	BIC Difference	Coefficient Subclinical (SD)	Coefficient Infection (SD)
MRSA	Stay length	6.41	108.40 (22.05)	634.47 (48.94)
	Death	2.99	0.47 (0.24)	0.61 (0.10)
C. difficile	Stay length	8.45	174.46 (40.64)	719.83 (102.93)
	Death	7.91	1.26 (0.39)	1.00 (0.15)
Norovirus	Stay length	10.95	284.53 (118.39)	562.39 (465.98)
	Death	-0.08	-8.57 (80.68)	0.19 (0.55)

Table 5.3: Model results for external outcomes. Models were fit using 1) only the diagnosis of infection as found in the EMR and 2) including a third category for subclinical infections. The BIC difference between these models is shown in the third column; a positive number indicates that the model with subclinical infections was a better fit. The models used for each of the three outcomes are as follows. Stay length was modeled via a linear regression controlling for age, sex, time to infectious disease test, primary diagnosis and consulting physician. Dying while in the hospital was modeled using a logistic regression controlling for everything in the stay length model as well as stay length. A positive coefficient indicates infection or subclinical infection predicted an increased likelihood of dying while in the hospital

5.4.2 Effects on individual health

I next examined the effect that subclinical infection had on patient health outcomes. Recall that I hypothesize that patients with subclinical infection would have outcomes midway between uninfected and infected individuals, since subclinical individuals have bacterial or viral loads that are below the sensitivity threshold of

the respective test, and bacterial/viral load is correlated to severity. To an extent, I observe that in the results (Table 5.2). The results for MRSA and *C. difficile* are in line with what I expect; health outcomes are negative and significant for infected and subclinical patients, but of lesser magnitude for subclinical patients. For norovirus, I observe the same trend for hospital stay length, but no significant effect on survival for either the infected or the subclinically-infected.

5.4.3 Effects on disease dynamics

In addition to understanding how subclinical infections affected individual patient outcomes, it is important to understand how their presence affected the dynamics of the infection in the population. There were a total of 106 outbreaks of MRSA, and 39 outbreaks of *C. difficile*. Outbreaks ranged in size from 1 to 66 patients. The total numbers of patients in the health care system during outbreaks ranged in size from 439 to 3,790 patients.

I calculated the colored triad census on the individual outbreak static networks to understand how subclinical infections related to the overall network structure, comparing the counts to those observed in simulations from the baseline ERGMs (Figure 5.4; Appendix A).

These results conform to a number of *a priori* expected features, showing that the colored triad census detects expected effects. The complete triad of infected patients is observed more often than expected by chance, due to the hospital's operating procedure of moving infected patients to a quarantine ward. The connected triad of subclinically-infected patients is observed less frequently than expected by chance, reflecting the fact that the random forest only includes co-presence with infected individuals, not with those later classified as subclinical. Therefore, co-presence with a subclinically-infected patient is not predictive of subclinical infection, leading to a lack of edges between subclinically-infected patients. Because I observe the expected

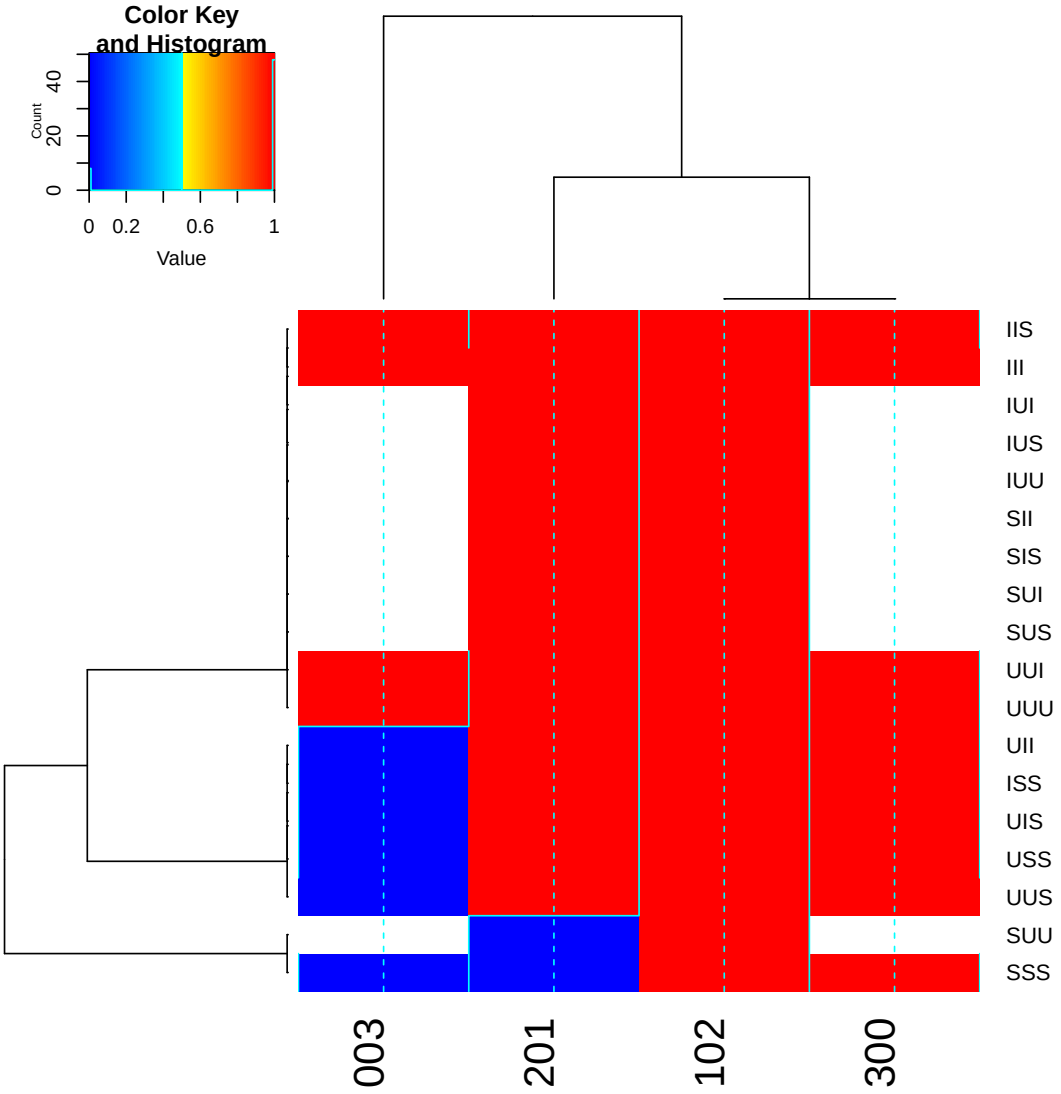

Figure 5.4: Heat map of colored triad census. The colored triad census was applied to the observed temporal network, and 1,000 networks simulated from the ERGM representing the null model. Each cell in the heat map represents the percentile of the empirical colored triad count in the distribution of colored triad counts. Red cells indicate colored triads observed more than expected by chance, and blue cells indicate colored triads observed less frequently than expected by chance. The rows of the heat map represent the structural configuration of the colored triad, while the columns represent colored triplets. Abbreviations on the rows are: "U" for uninfected, "I" for infected, and "S" for subclinically-infected. The values in the colored triplet correspond to the top node, the bottom-right node, and the bottom-left node, respectively. White cells indicate redundant colored triads.

effects, the unexpected results of the colored triad census are more likely true effects.

This approach also leads to conclusions about the role of subclinically-infected patients in disease dynamics. I find that the triad consisting of a subclinically-infected patient connecting two otherwise-unconnected infected patients is observed more frequently than expected by chance, and the triad consisting of a subclinically-infected patient connecting two otherwise-unconnected uninfected patients is observed less frequently than expected by chance. Together, these findings indicate that subclinical patients may connect infected patients who otherwise would not be connected. If the temporal ordering of infections in the triad is infected to subclinical to infected, then the subclinically infected patient may be responsible for transmitting the infection to the third patient.

To test the possibility that subclinically infected patients exacerbate nosocomial outbreaks, I use the temporal network to understand how the presence or absence of subclinically-infected patients affect the dynamics of the infections. The results from the SETIR model (infections including subclinical) are shown in Table 5.4. The SETIR model was calibrated on the observed data with subclinical infections, and as such, reflects the central tendency of the observed outbreak, but had higher variation (greater range and SD of outbreak size than the observed outbreaks).

Disease	Outcome	SETIR	SETIR with quarantine	Observed with subclinical	Observed without subclinical
MRSA	Median outbreak size (range)	5 (1,106)	3 (1,80)	5 (1,85)	4 (1,66)
	Mean outbreak size (SD)	11.05 (11.91)	6.43 (7.73)	10.21 (12.69)	7.45 (8.47)
	Fano Factor (SD)	10.08 (58.06)	10.78 (60.43)	10.44 (35.84)	10.75 (36.66)
C. difficile	Median outbreak size (range)	4 (1,84)	3 (1,65)	4 (1,31)	4 (1,30)
	Mean outbreak size (SD)	7.12 (7.05)	4.78 (6.12)	7.58 (7.70)	6.09 (6.31)
	Fano Factor (SD)	4.11 (3.75)	4.46 (4.31)	1.29 (3.69)	1.3 (3.65)

Table 5.4: Model results for disease dynamics. SETIR model was fit 100 times to each empirical outbreak network. The Fano Factor is defined as the ratio of the variance to the mean. Higher values indicate that the timing of infected patients was more "bursty", or occurred closely in time to one-another. I then rerun the simulation after increasing the recovery rate for subclinical patients to be in line with infected patients, reflecting potential quarantine of subclinical patients. The last two rows of the table are the observed counts across the outbreaks, first with the subclinical patients included, and then with them excluded.

I also examine how disease dynamics might be affected by early identification and quarantine of subclinical patients (Table 5.4, row 2). I assume that the early identification and quarantine of subclinical patients would reduce their infectious period to similar levels as the infected patients. When I do so, I see that outbreak size is reduced to below the levels of the outbreak not including subclinical infections (7.45 vs 6.43 for MRSA, 6.09 vs 4.78 for *C. difficile*). This indicates some of the infections are likely caused by exposure to subclinically-infected patients, and that intervening in this way could reduce outbreak size.

5.5 Discussion

In this study I show that I can detect patients with likely subclinical infection based on electronic medical records and hospital administrative data. I perform a number of analyses to assess the validity of the model, showing that it both recaptures known infected individuals and that identified subclinical patients have health outcomes between uninfected and infected patients, as expected. I find that subclinical patients help explain nosocomial outbreak dynamics by connecting otherwise-disconnected groups of infected patients. Finally, if subclinical patients could be identified and quarantined like infected patients, then nosocomial outbreaks could be reduced.

Importantly, the presence of subclinically-infected patients affects the disease dynamics of nosocomial outbreaks. Not only do I find that subclinical infections impact an individual's health, but that affected patients are themselves likely infectious. Even though the model indicated that subclinical patients are less infectious than infected patients, they go unnoticed due to their lack of symptoms, remaining a potential vector in the hospital for longer than patients with overt infection. This is particularly important if a patient is transfered between wards; if they are subclinically-infected, then they potentially expose entire additional groups of patients to an infection. Indeed, this likely occurs and is responsible

for some infections in an outbreak, as I observed from the $201 - SII$ triad being observed more often than chance. However, if subclinically-infected patients can be identified and quarantined at a similar rate as infected patients, nosocomial outbreaks can be dampened, and their eventual impact lessened.

The behavior of the algorithm identifying the subclinically-infected patients is important. As shown in Figure 5.3, the time co-present with infected patients was the strongest classifier, creating sharp divides between uninfected and infected individuals. None of the biomarkers were strongly predictive on their own, but in concert, the four biomarkers composing the CMI response (white blood cells, eosinophils, neutrophils, and CRP) helped to further differentiate infected from uninfected individuals. The inclusion of these factors was important, as the CMI response happens relatively early in infection, when bacterial and viral loads are still low[204].

The model has a number of inherent strengths as well, which make it well-suited to detect subclinical patients. Random forest models allow for complex non-linear relationships between the variables and latent class. This allows the model to maximally separate groups. The model can also capture a variety of potential transmission pathways in addition to direct patient-to-patient contact. For instance, the inclusion of a patient's consulting physician covers transmission via health care staff, which is often responsible for a substantial fraction of nosocomial infection[188]. For health care practitioners such as nurses who are often located within a single ward, the co-presence metric would also capture infection spread through these persons. I can also account for bacteria living on surfaces outside the human body by artificially extending a patient's time in the ward as contaminated surfaces often play a part in nosocomial infection spread[225].

Despite its general effectiveness, the model did behave less well with respect to norovirus. This is largely because of the very few cases of nosocomial norovirus in the hospitals examined. When a patient presents with norovirus, they are

very often immediately sent home[226]. Because of this, very few patients are exposed to norovirus in the hospital, and do not subsequently develop nosocomial norovirus. The model then observes almost no patients who were co-present with infected patients, leading to very few subclinical cases. Additionally, I observe better separation in some of the biomarkers for bacterial infections than for viral, which makes the remaining information less informative for separating infected and uninfected patients.[227,214]. This does not mean that subclinical cases of norovirus do not exist, but that if they do, this model is unable to capture them.

The assumption that the model identifies subclinical patients is a strong one, and is one that cannot be directly verified based on the definition of subclinical infections. However, the analyses I performed to assess the validity of the model all indicated that the algorithm was accurate at detecting subclinical infections. First, the models for stay length and survival in the hospital show better model fit when using the latent class membership rather than the positive microbiological test. If a subclinical infection triggers an immune response, particularly while a patient deals with the morbidity that brought them to the hospital, then negative outcomes are expected. The cross-validation also shows that the model can accurately recover known infected patients relatively well. If subclinically-infected patients do indeed appear similar to infected patients, then the model is likely similarly accurate in detecting subclinically-infected patients. My examination into the dynamics of the disease also reflect the likely accuracy of the model.

In conclusion, the findings demonstrate that electronic medical records and hospital administrative data can be used to screen for subclinical infection across a number of bacterial, but not viral, diseases. These subclinical infections impact both affected patients and can lead to infection in other patients, affecting disease dynamics. Additional studies should examine this prospectively, attempting to identify subclinically-infected patients in real-time, and whether interventions would reduce the impact of nosocomial outbreaks as shown here in simulation. This

work is therefore the first step towards showing that subclinical infections are an under-appreciated problem in the realm of nosocomial outbreaks. Only by recognizing that these infections likely exist and have adverse effects on patients can I move to pro-actively reduce their impact.

6

Conclusions

Contents

6.1 Summary of results

The results of the previous chapters can be simply stated thusly: patient-patient co-presence matters for health in hospital settings. It matters to patients' health, their trajectory of wards, and possibly even to their mental health. The evidence,

notably, is entirely based on an analysis that restricts itself to the use of electronic medical records and hospital administrative data.

In Chapter 2 I describe the dataset and the population comprising its catchment area. I also cover demographics and some other information about the structure of the data. Finally, I show how a variety of bipartite networks can be constructed from EMR and HAD, as well as some basic insights that can be gleaned from each source. Here, I also highlight how the patient-patient co-presence network is the least studied of these, and is part of why I choose to examine the effect of co-presence on patient health.

In Chapter 3, I show that the co-presence patterns of patients when in the chemotherapy ward matter with respect to 5-year survival following chemotherapy. Being around any patients *consistently* had a strong positive effect on survival outcomes, as predicted by sociological theories such as social facilitation[20]. Additionally, this strong effect was moderated based on whether patients responded positively or negatively to their treatment, in line with theories such as modeling[56]. This work makes strong contributions to the fields of health services research and sociology, both empirically and methodologically. With respect to health services research, the finding of ameliorated outcomes when consistently around other patients is at odds with the current trends of increasing privacy in chemotherapy wards in some countries[228], and of increasing use of oral chemotherapy which can be taken at home[229]. Sociologically, this work improves our understanding of how some types of interactions may occur *de novo*. The chemotherapy ward is a place where patients generally do not know one another prior to beginning treatment. The detected effects therefore arise from new social interactions, which is often difficult to observe. Furthermore, the method used to detect consistent co-presence was a novel data-driven method to determine a patient-specific cutoff, rather than an arbitrary cutoff, addressing a common criticism of dichotomization.

This chapter also served an important role in the construction of this thesis by

working as a proof-of-concept of the planned further analyses and chapters. Instead of using the entirety of the data, a small, manageable subset was chosen. This subset was somewhere I *a priori* most expected to be able to detect the effect of co-presence. If I was unable to detect it in this case, I would have been unlikely to detect it elsewhere, and would have likely needed to refocus my efforts. As well, it allowed me to slowly build towards optimization for larger analyses, as was needed in the following chapters. This was also an important point towards the construction of Appendix A, as it required previous knowledge of optimization techniques. Without the foundation gained in conducting this chapter, constructing the algorithm and gaining the insights from conducting it would likely not have occurred.

In Chapter 4, I show that the hours of co-presence with patients suspected of infection is a strong predictor of subsequent infection. This finding held across different infectious diseases with different modes of transmission, as well as both bacterial and viral infections. Furthermore, this test is disease-specific unlike many other predictive tests which test for general markers of infection. This builds on previous work where the co-presence network of patients was *implicitly* used to understand the path of transmission in a nosocomial outbreak, but is the first time it is explicitly used as a predictive test on its own. This is potentially an important advance in using EMR and HAD, as it shows how monitoring these data can lead to improved predictions of patient health.

In Chapter 5, I again show that co-presence is an important predictor of health outcomes. At a minimum, the results therein show that increasing co-presence with infected patients is correlated with worse outcomes, including length of stay, survival, and readmission rates. Beyond that, it shows that the data from EMR and HAD are more than the sum of their parts: health states outside of those measurable by standard methods may be discoverable through the rich information contained in the EMR and HAD. Again, this points to the import of EMR and AD in understanding and monitoring patient health. More pressingly, this work may

point to the presence and effect of patients with subclinical infections, which is not something being addressed by the medical community, despite others' knowledge that it occurs (e.g. in bovine tuberculosis[205]).

6.2 Strengths and limitations

Although I describe the strengths and limitations of each individual chapter within the confines of that respective chapter, there a number of generalizations I would like to highlight here.

As previously described, a number of strengths and limitations derive from the use of EMR and HAD here. The temporal precision and large number of data points on each patient were incredibly useful in understanding health at such a detailed level. This also allowed me to build a number of data-driven approaches that would have been much more noisy in smaller datasets. Using these data also changed the types of questions I was able to ask, particularly with public health implications in mind. Because these data came from EMR and HAD, I was very realistically able to ask how the results could be turned around and implemented in existing medical systems to actively monitor patient health. In other words, the transition from retrospective to prospective study in this setting is less of a gap than in many other frameworks.

One important limitation across all the studies was a lack of additional demographic data, including race and some proxy for SES. This precluded me from asking certain questions, particularly whether there was a difference in care or outcome across these attributes. Additionally, the sheer size of the data would have allowed for appropriately-powered analyses to detect these effects. For this reason, having access to race and SES data would benefit studies such as those presented in this thesis.

Beyond the general limitations of Big Data described in Chapter 1, these studies had a common limitation that bears further examination: being unable to infer

mechanism. Although all the results I show have potentially important consensuses for health care outcomes, the exact mechanism by which they occur is not explicitly defined through the methodological approach I took. For example, both social facilitation and social support could explain the outcomes observed in Chapter 3. Using only co-presence data cannot distinguish between them. In this way, other, more traditional methods such as surveys would be needed to supplement these results. The same goes for Chapters 4 and 5, where the data do not distinguish between method of vector transmission. Whether co-presence predicts infection because it is directly passed from patient to patient or because the two patients share a nurse who transmits the infection from one patient-to-another cannot be determined from this data alone.

6.3 Future work

The work shown herein is largely a first step towards understanding co-presence and health. This is because 1) these questions have generally not been asked due to the difficulty of gaining data to answer them and 2) even now that sufficient data was gathered, the underlying mechanisms couldn't be elucidated. Therefore, this work opens up a number of future directions to better understand the effect of co-presence on health. I intend to carry some of it out on my own as an immediate follow-up, but some of it is outside of my expertise and is likely best done by experts in their respective fields.

Importantly, Chapter 5 was the first chapter in which I explicitly made full use of the temporal nature of the data and the subsequent networks. Although the results from the temporal models largely conformed to what I had already found via the static network, they allowed me to address more-nuanced explanations of the initial findings. This approach also allowed me to model a variety of time-dependent interventions which would have likely been impossible to implement

via a static network. Using what I learned from this approach and applying it to Chapter 3 would likely yield new insights. Specifically, instead of using the static network and having entire chemotherapy regimens as the unit of observation, I can use the temporal network and have individual chemotherapy spells be the unit of analysis. In doing so, I can use more temporally-proximate outcomes, such as day-to-day health based on biomarkers. This will allow for a much more nuanced understanding of the effects initially observed in Chapter 3.

As briefly stated in Chapter 3, the results therein invite additional investigation using data and methods I have not used here. The data lack a number of variables which could partially or totally account for the observed associations. Future research may aim to replicate these analyses inclusive of additional data on these potential confounders. The data I used also lacked information allowing me to explicitly disentangle the potential mechanisms. Furthermore, it did not include any information about the subjective patient experience; it is unknown whether patients did interact with one another, and whether they were conscious of the health status of those around them. This kind of data will likely never be captured in HAD or EMR, and therefore other approaches, such as surveys or ethnography would be useful to establish this. This approach would be needed to fully understand the effects first observed in Chapter 3.

In addition to gleaning more insight from this data on the same effects previously observed, this work also opens the door to other approaches towards the same question. Now that I have established that social influence likely occurs in the chemotherapy ward, previously-uncertain efforts, such as surveying chemotherapy patients, may be seen as more likely to yield results, and therefore are more worth pursuing. This could lead to additional understanding of the mechanisms, which are not easily fully identifiable based on my approach.

The results in Chapters 4 and 5 lend themselves to further prospective research to understand how well the results would perform in practice. Using EMR and

HAD to predict subsequent infection could be highly-impactful if shown to still be highly predictive in a prospective setting. Although I did all I could to have my retrospective data best resemble the analogous prospective data, there may be eventualities I did not foresee. These would become apparent in a prospective study.

Similarly, the results from Chapter 5 only show that it is likely that subclinical infections exist in the health care system. A prospective study that uses the algorithm I developed could identify potential subclinical patients in real-time. These patients could be monitored more closely, which would catch "subclinical" infections that were not truly subclinical, but were missed due to a lack of scrutiny of the patient. They could also be pro-actively quarantined without the confirmed presence of an infection. If subsequent infections then decrease, this would be additional evidence that these patients are truly harboring infection.

In Chapter 1, I note that research has shown that co-presence can increase immune activity. At the same time, co-presence is a strong predictor of infection, often a necessary component (See Chapter 4). These two effects of co-presence are opposed, but have not been studied in concert. Disentangling these effects could therefore lead to increased understanding of each individual pathway. For instance, the studies examining the effect of co-presence on immune activity often look at aggregate social interaction whereas studies examining the risk of infection based on co-presence look at much more temporally-proximate co-presence. These disparities in temporal scale make disentangling these effects difficult. A single study examining both on a similar time scale would likely provide additional insight into these phenomena.

Finally, the algorithm I developed in Appendix A invites future use in understanding networks where both structure and node attributes matter. My approach of defining a null hypothesis and sampling therein yields analyses that combine many traditional analyses at once. The algorithm itself is also designed to be efficient and work on networks up to 1000's of nodes, allowing it to be applied

to a wide variety of empirical networks.

6.4 Summary

In sum, I have used EMR and HAD to show that the patients with whom one is co-present while in a health care system impacts one's health. This occurs through both psychosocial and biological mechanisms. These findings are important to a number of fields, and advance both the methodology and the empirical findings forwards. Although not without limitations, they are important first steps towards understanding how this relatively-unexplored phenomenon occurs, and how we can use this knowledge to improve health. Fortunately, the work immediately suggests future steps that can be taken to expand on these results to further our understanding of the mechanisms at play.

Appendices

A

An efficient counting method for the colored triad census

Contents

A.1 Abstract

The triad census is an important approach to understand local structure in social network analysis, providing comprehensive assessments of the observed relational configurations between triplets of actors in a network. However, researchers are often interested in combinations of relational and categorical nodal attributes. In this case, it is desirable to account for the label, or color, of the nodes in the triad census. In this paper, I describe an efficient algorithm for constructing the colored triad census, based, in part, on existing methods for the classic triad census. I evaluate the performance of the algorithm using empirical and simulated data for both undirected and directed graphs. The results of the simulation demonstrate that the proposed algorithm reduces computational time many-fold over the naïve approach. I also apply the colored triad census to the Zachary karate club network dataset. I simultaneously show the efficiency of the algorithm, and a way to conduct a statistical test on the census by forming a null distribution from $1,000$ realizations of a mixing-matrix conditioned graph and comparing the observed colored triad counts to the expected. From this, I demonstrate the method's utility in the discussion of results about homophily, heterophily, and bridging, simultaneously gained via the colored triad census. In sum, the proposed algorithm for the colored triad census brings novel utility to social network analysis in an efficient package.

A.2 Introduction

The triad census is an important approach towards understanding local network structure. Holland and Leinhardt [230] first presented the 16 isomorphism classes of structurally unique triads possible in a directed network without loops. To conduct

a triad census, one simply counts each occurrence of these structures, without respect to the labeling of the nodes (here I use node label, color, characteristic, and attribute interchangeably). This is useful insofar as specific triads, or combinations thereof, may relate to underlying social processes giving rise to an observed network. For example, bridges (triads with one null dyad and two non-null dyads) may be important in navigating social networks[231], and certain triads may be more or less favorable based on structural balance theory (e.g. the 300 is balanced but the 201 is not, see Figure A.1)[232]. Moreover, a variant of the triad census, motif analysis, investigates the statistics of various triad configurations (motifs), and has found wide application in biology[233].

Also important to network structure are nodal characteristics and how they relate to tie formation or dissolution. This has been the subject of research on homophily (individuals having similar attributes with those to whom they are connected)[234]. However, homophily is an observed phenomenon, not a process. The processes giving rise to homophily are varied, often confound the relationship between networks and outcomes, and are difficult to tease apart[158]. Methodological advances, such as stochastic actor-oriented models can disentangle these effects to some extent[59]. Other analyses have attempted to disentangle the processes leading to homophily from structural processes, such as triadic closure[235]. Additionally, the coloring of nodes in a network has been an important question for many graph theorists and indeed represents a major topic in this field[236].

Although nodal characteristics and the triad census are important, they have rarely been examined fully in conjunction. Yet, there are a few cases where specific colored triads have been studied. For example, Gould and Fernandez[237] study brokerage based on triad structure and group membership simultaneously. This same approach has been used to study brokerage in dynamic networks[238]. As well, a study by Marcum and Koehly[239] examined specific colored triads based on generational membership within families; in this work the authors showed that

inter-generational ties were observed in different quantities than expected based on the underlying null model. None of the past research evaluated the full census of colored triads, rather, researchers have focused instead on specific colored triads that were *a priori* expected to be relevant to the processes at hand. As a result, these foundational works were not exhaustive with respect to all alternatives. In other words, previous research examining a subset of colored triads likely had some number of false negatives due to not examining every colored triad; this could be addressed by censusing the colored triads.

The examination of node characteristics together with local structure is important as it provides opportunity to simultaneously study the occurrence of triadic structure, nodal attributes, and the interactions between them. For instance, certain colored triads may be forbidden, such as three-cycles between strict heterosexuals in mixed-orientation sexual contact networks[240]. Impermissible triads would be categorized the same as those that were not observed due to chance in a triad census, potentially missing important social processes or constraints at play in this type of network. Only by incorporating node coloring into the triad census can this pattern be fully elucidated.

Based on this methodological gap in the literature, I develop a method to census the colored triads for any one-mode binary network with arbitrary number of colors. Due to the large numbers of isomorphism classes of size 3 as the number of colors increases, this method requires computational efficiency in addition to mathematical accuracy. As well, one is often interested in forming a null distribution with which to compare observed colored triad counts. If the null distribution cannot be analytically solved, one would likely census the colored triads of many simulated networks, further increasing the need for the algorithm to be computationally efficient.

Current most-efficient methods for the triad census exploit the sparseness of networks[241], and scale sub-quadratically (as the number of edges increases the time to run the algorithm is faster than the number of edges squared). However,

methods that exploit network sparseness by inferring the number of null triads do not work in the colored case because they do not explicitly interrogate every triad, and there are multiple isomorphism classes among the null triads due to the coloring. Therefore, I extend the methodology of Moody[108], which is based on matrix algebra and interrogates every triad.

This paper (1) presents the colored triad census and its computational complexity, (2) shows that this approach can be used on large networks (tested for up to $10,000$ nodes) with up to 10 colors in relatively-efficient time, and (3) uses the method many times to create null distributions of colored triad censuses to form the basis of conditional uniform graph tests. (4) I illustrate the benefits of an analysis incorporating the colored triad census using a well-known dataset, Zachary's Karate Club[109].

A.3 Algorithm

Since the original appearance of the triad census in 1976, a number of papers have explored how to compute the triad census of a network in an efficient manner. Although approximately-quadratic methods (in terms of number of nodes) exist for calculating the triad census for sparse networks, (e.g. Batagelj and Mrvar[241]), I use the algorithm presented by Moody[108] here. This is because the more efficient methods avoid interrogating null triads directly by taking advantage of the sparseness of graphs, the subsequent large number of null (003) triads, and the known number of total triads. Instead, they interrogate all triads with at least one edge, and then subtract that count from the total number of triads in the network to arrive at the number of null triads. This is insufficient in the colored triad census as there are differently-colored null triads, and the count of each cannot therefore be algebraically determined. For example, if there are two colors, four different null colored triads exist (0-3 nodes of color A). The exact breakdown of the null triad

into the four colored triads cannot be determined without interrogating each null triad, thereby losing the efficiency gained when not considering colors. Moody's algorithm does not employ this limiting shortcut, and I therefore use it as a basis for the colored triad census algorithm. Additionally, because many networks are sparse, I can leverage computational techniques for increasing the efficiency of sparse matrix operations[242], further reducing the computational complexity of the algorithm.

Moody[108] showed that the count of each of the 16 triad isomorphism classes could be derived by using matrix algebra on the adjacency matrix of the graph and its derivatives. To review, let $\mathbf{A}$ be the adjacency matrix of a network, and $A_{ij} = 1$ when a tie exists from node i to node j. Let E be the symmetrized matrix A, formed by making any edge in A reciprocal via $E_{ij} = \max(A_{ij}, A_{ji})$. The complement of E, $\bar{E}$, is formed by subtracting the complete network adjacency matrix from E, so that $\bar{E_{ij}} = 1$ if and only if there is neither a tie from i to j nor a tie from j to i. Next, I have M, the mutual matrix of A, and is made by removing any asymmetric edges from A, or $M_{ij} = \min(A_{ij}, A_{ji})$. Finally, C is the matrix of only asymmetric edges, and is calculated by $C = A - M$. Therefore, $C_{ij} = 1 \iff A_{ij} = 1 \ \& \ A_{ji} = 0$. Based on these matrices, Moody demonstrates how to calculate the number of each of the 16 isomorphism classes for the case of unlabeled graphs (or, equivalently, for a graph consisting of nodes of the same single color). Generally, this was done by multiplying (either through dot-product or element-wise multiplication) the three matrices corresponding to the relevant edges in the triad of interest. There were two triads ($111U$ and $111D$) that were not directly amenable to this process and were calculated via addition and subtraction of other triad types, respectively.

To extend this work to the case of multiple colors, I introduce the out-coloring and in-coloring matrices, K^r and $K^{r'}$, respectively, where r is the focal color of matrix K. Here, the in-coloring matrix is the transpose of the out-coloring matrix. The out-coloring matrix is calculated by evaluating the color of the nodes row-wise,

such that rows indexing nodes of the focal color are composed in the following way:

$$K^r_{i\bullet} = \begin{cases} 1 & \text{if } R(i) = r \\ 0 & \text{if } R(i) \neq r \end{cases} \tag{A.1}$$

Where $R(i)$ is a function returning the color of node i. As above, the in-coloring matrix is the transpose of the out-coloring matrix in Eq. A.1.

My algorithm works by using the in- and out-coloring matrices to evaluate and "switch on" edges that have nodes of the focal colors at the *ends* (or tails) of edges in the adjacency matrix A of the network. I adapt the triad census nomenclature of Holland and Leinhardt [230] by appending the colors after the name of the triad. The colors are ordered from the top node proceeding clockwise in Figure A.1. I have arbitrarily adapted the orientation of the triads from the triad census figure in Holland and Leinhardt [230] for computational reasons. The orientation is important here because triads with the same orientation may no longer be isomorphic when color is introduced. Figure A.1 makes it possible to count unambiguously and name only unique colored triads. Therefore, $T_{102-123}$ is the triad consisting of 1 symmetric dyad and 2 null dyads, where the top node is of color 1, the bottom-right node is of color 2 and the bottom-left node is of color 3. This is distinct from the $T_{102-312}$ triad because the coloring of the nodes is not identical from the previous triad.

Following this, the general formula for an arbitrary triad "T" with an arbitrary coloring triplet is:

$$T = \mathrm{Tr}\left((K^1 \times H(T,1,2) \times K^{2'})(K^2 \times H(T,2,3) \times K^{3'})(K^3 \times H(T,3,1) \times K^{1'})\right) \tag{A.2}$$

In the above, "$\times$" refers to element-wise multiplication, and "Tr" is the trace function. For an arbitrary triad, "T" has a color triplet r_1, r_2, r_3. $H(T,i,j)$ is a

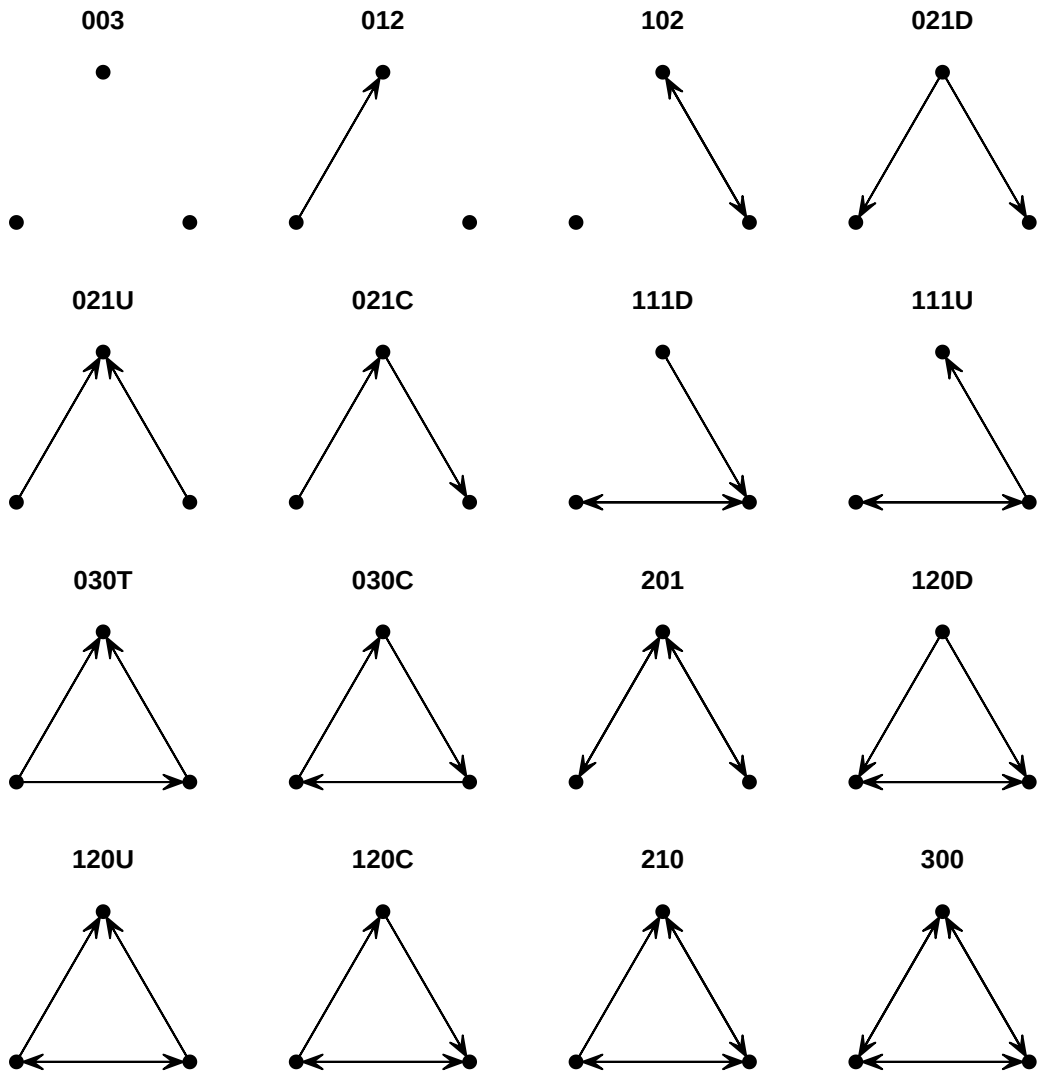

Figure A.1: The 16 isomorphism classes of triads and their orientation used here with respect to the color numbering. When colors are added to these triads, they are labeled starting from the top node and proceeding clockwise.

function returning the matrix specific to the type of edge between nodes i and j in triad "T". For example, in a 102 triad, the first edge from the top node going clockwise is a symmetric edge from node one to node two (Figure A.1). $H(T_{102}, 1, 2)$ in this case would be the matrix E for the symmetric matrix, and the sandwiching color matrices would turn the proper edges on and off if nodes one and two were of the specified colors. If the edge is an asymmetric one, and the direction of the edge in Figure A.1 is counter-clockwise, then C' is used instead of C to force

the edge to go in the proper direction.

At this point, there are redundant triads due to certain colored triads being isomorphic. For instance, the $T_{003-122}$ is isomorphic with $T_{003-221}$ and $T_{003-212}$, and would be triple-counted. These are removed by checking for isomorphisms based on matrix row and column permutations of the triad. If two colored matrices are identical after such row and/or column permutations, then they are isomorphic, and all but one are. I arbitrarily decide to keep the triad whose coloring triplet name comes first alphanumerically. It should be noted that removing in this way is computationally expensive, particularly as the number of colors and nodes grows large. I therefore shorten this process by performing it once for 1 to 10 colors and storing the unique isomorphism classes. This leaves only unique isomorphism classes of colored triads, which can then be accessed in linear time.

The number of unique isomophism classes for a given number of colors can be shown for each of the 16 ismorphism classes in the triad census. The 16 classes separate into four types of colored triads, depending on how many structurally-distinct positions there are in the triad (e.g. the two ends of the edge in a 102 triad are not structurally-distinct from one another, but are distinct from the node with no edges). The calculation for the number of each isomorphism class for arbitrary number of colors (k) is shown in Table A.1. Each combinatoric term in each row (together with their respective leading permutation coefficients) counts the number of colored triads when there are three, two, or one unique color(s), respectively. For example, in a network with three colors,the '300' and '003' classes have only one accessible permutation when there are three colors present in the triad (i.e. $\binom{3}{3}$), six ways when there are two colors (i.e. $2\binom{3}{2}$), and three ways when there is one color in the triad (i.e. $\binom{3}{1}$).

If these numbers are summed over the 16 isomorphism classes, the total number of colored isomorphism classes of triads for k colors is returned. Similarly, the same can be done for undirected triads, solely summing over the 4 triads observed in the

Isomorphism classes	Number of colored triads
300 and 003	$\binom{k}{3} + 2\binom{k}{2} + \binom{k}{1}$
030C	$2\binom{k}{3} + 2\binom{k}{2} + \binom{k}{1}$
102 021D 021U 201 120D and 120U	$3\binom{k}{3} + 4\binom{k}{2} + \binom{k}{1}$
012 021C 111D 111U 030T 120C and 210	$6\binom{k}{3} + 6\binom{k}{2} + \binom{k}{1}$

Table A.1: Expression for the number of isomorphism classes within a triad class. k is the number of colors

undirected case. Table A.2 reports the total number of colored triads for undirected and directed networks over a range of k. Clearly, the number of isomorphism classes grows quite quickly as k increases.

Number of colors	Number of directed colored triads	Number of undirected colored triads
1	16	4
2	104	20
3	328	56
4	752	120
5	1440	220
6	2456	364
7	3864	560
8	5728	816
9	8112	1140
10	11080	1540

Table A.2: The number of colored triad isomorphism classes for directed and undirected networks for k ranging from 1 to 10.

The algorithm implemented as an R package is publicly available and is linked to this paper via github: `https://github.com/jlienert/ColoredTriadCensus`.

A.4 Algorithmic performance

If a naïve implementation of matrix multiplication is used, this algorithm runs with computational complexity $O(N^3 * 3^k)$. It scales with the number of nodes cubed (N^3) because of the matrix multiplication involved in the algorithm. However,

many software packages use algorithms that reduce the complexity of matrix multiplication to $O(N^{2.38})$[243]. Furthermore, by taking advantage of methods for matrix multiplication using sparse matrices (as appropriate due to the sparse nature of most social networks), this complexity is reduced closer to $O(N^2)$[244]. The exact benefit gained by using sparse matrix multiplication varies based on how sparse the matrix is. This ranges from the nearly-optimal $O(N^2)$ when very few edges exist, to worse than the optimized algorithm when many edges exist. The scaling with 3^k comes from the number of distinct colored triads the algorithm needs to evaluate, and the number of isomorphism classes scales in such a manner.

To test the efficiency of the algorithm, I apply it to networks ranging in size from $n = 10$ to $n = 10,000$ with the number of colors ranging from $k = 3$ to $k = 10$, all holding the average degree constant at 6 by creating Erdös-Rényi graphs with an edge probability of $\frac{6}{N-1}$. This reflects the average number of ties participants often enumerate in social networks surveys[245]. The runtime of the algorithm with these parameters can be seen in Figure A.2. In general, increasing K results in constant increases in $\log(runtime)$, which is what I expect based on the theoretical computational complexity. As expected, I also observe a super-quadratic increase in $\log(runtime)$ as N increases. Although it is super-linear, it is still below the curve that would exist if I used matrix multiplication not optimized for sparse matrices (dotted line in Figure A.2). This difference shows the expected time saved by using sparse matrix methods. Finally, I observe changes in the rank-order and decreases in runtime going from 10 to 100 nodes. This is also due to the computational time involved in initializing the sparse matrices and storing and operating on sparse matrices, and as such is not unexpected. Additionally, because the average degree was held constant, the smaller networks are much more dense, and therefore are actually less efficient than if they used standard matrix multiplication methods. To be perfectly optimized, therefore, the algorithm would use standard matrix multiplication for small networks, and switch to sparse methods for larger networks.

However, the gains would be minimal, generally under 10 seconds, and would require additional logical steps to check for network size, minimizing the gain. I therefore use sparse matrix methods for all network sizes.

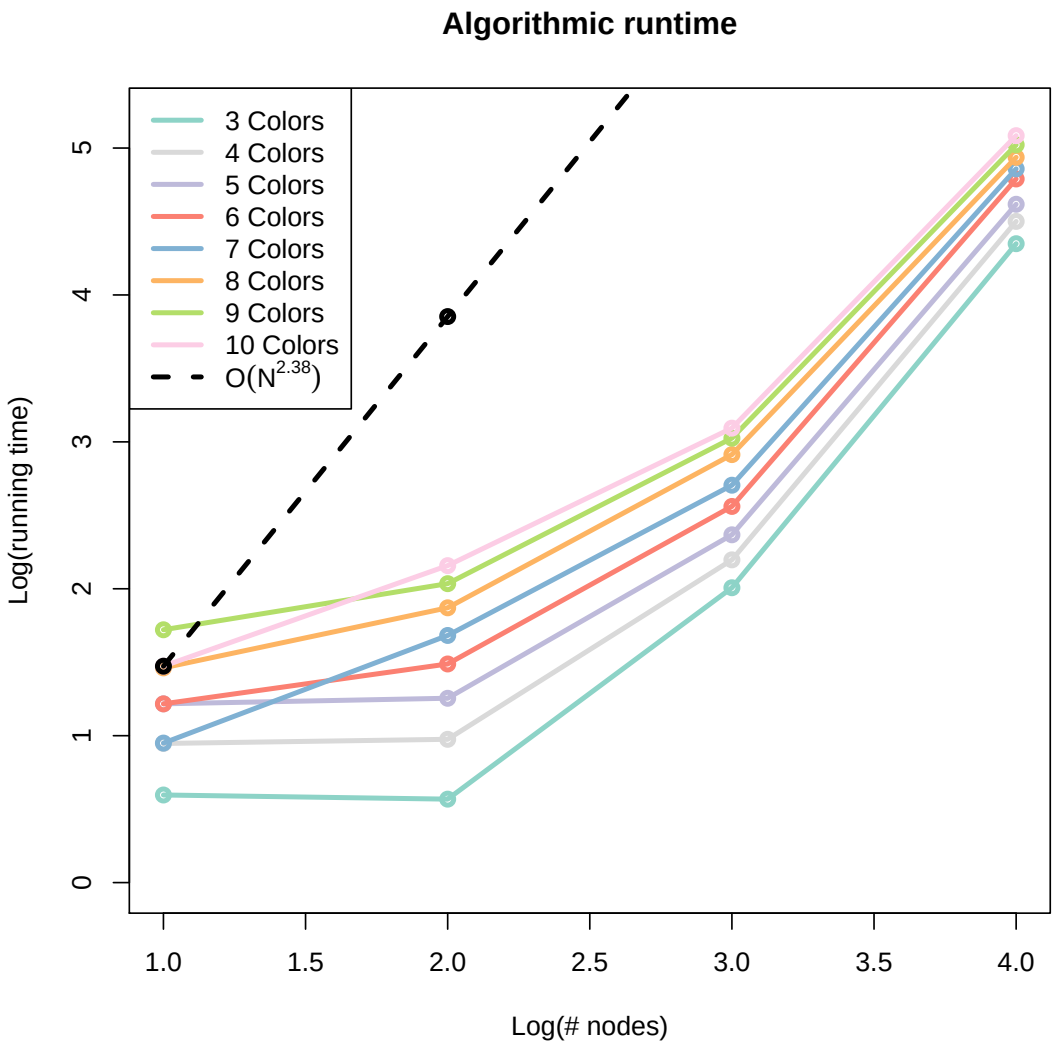

Figure A.2: Runtime of the algorithm on networks ranging from size 10 to 10,000 nodes in orders of magnitude, and from one to ten colors. Additionally, the dashed line represents the computational time that would be expected using standard matrix multiplication methods for $k = 10$. These runtimes were generated using virtual PCs including 1 dual core CPU and 10GBs of RAM.

A.5 Empirical use and example

To show the empirical value of this algorithm, I use the Zachary karate club social network[109]. This is a well-known historical network that describes the social relationships between 34 members of a university karate club. Ties exist between members if they overlapped in at least one of eight contexts representing undirected relations. These relations varied in terms of likely strength of the association. Likely at the weak end of the spectrum is being enrolled in the same class at the university, while likely at the strong end is being a student-teacher at the studio. Additionally, three ties are specific to activities with a part-time instructor.

Member “factions" were identified as a node attribute, taking one of five mutually exclusive values: strongly associated with the president, weakly associated with the president, neutral, weakly associated with the part-time instructor, or strongly associated with the part-time instructor. These are labeled "Zs", "Zw", "N", "Hw", and "Hs", respectively. These labels can be placed on an ordinal scale from -2 (Zs) to 2 (Hs) to quantify members' direction and strength of alignment. This undirected network with five colors represents a case that is rich in the number of colored triads (220) for detailed conclusions to be drawn using the proposed algorithm (which is general to both undirected and directed networks).

I initially ran the colored triad census on the social network using the faction as the nodal attribute. This provided the basis for the empirical observed colored triad census. To determine whether these triads were observed more or less often than expected by chance, I constructed a null model. As the choice of null model can have important ramifications for the null distribution of triads, I chose a model where edge formation is a function of the probability of ties between nodes of specific attributes[246]. The null model is a mixing-matrix conditioned uniform random graph distribution based on probabilities of edges between nodes of particular color combinations[247]. This matrix comprises empirical probabilities of ties between

groups, with the diagonal representing within-group tie probabilities. Observations of significantly over- or under-represented colored triads are the result of network effects beyond homophily and heterophily. Networks are then generated from this matrix via a Bernoulli random graph process á la Erdös and Rényi [248]. This null model therefore conditions on graph size, the distribution of node factions, and the probability of ties within and between factions. By generating networks from the null model, I can observe whether colored triad counts deviate from that expected based on the marginal distribution of faction mixing. Because I condition on the above parameters, if I observe statistical deviations in the colored triad census, it indicates that the structure of the network is dependent on parameters other than those on which I conditioned.

Moreover, for any triad, the expected number and variance can be calculated assuming each tie follows a Binomial distribution (which is a reasonable assumption for most binary social network data). The observed number can then be compared to these numerical results and a p-value extracted from an exact Binomial test. This equates to the following probability, expectation, and variance for an example colored triad:

$$\begin{aligned} P(T) =& P(A_{ij} = 1 | R(i) = r_1, R(j) = r_2) \times P(A_{ij} = 1 | R(i) = r_2, R(j) = r_3) \\ & \times P(A_{ij} = 1 | R(i) = r_3, R(j) = r_1) \end{aligned} \tag{A.3}$$

$$E(T) = P(T) \times \prod_{r=1}^{L(T)} \binom{\sum K^r_{\bullet 1}}{S(T, r)} \tag{A.4}$$

$$V(T) = E(T) \times (1 - E(T)) \tag{A.5}$$

The probability of T, $P(T)$ in Equation A.3 is based on the mixing-matrix of the three colors (r) involved in the triad T. As is standard for the mixing-matrix approach, this continues to assume that all edges in the graph are independent.

For the expected value of a specific triad, I multiply the probability of a single one of those triads by the total number of colored triplets that exist in the graph. In Equation A.4, the expectation of the triad, $L(T)$ returns the number of unique colors in T and $\sum K^r_{\bullet 1}$ is the number of nodes of color r in the graph. Also, I take the nodes one, two, or three at a time depending on how many times that color repeats in T, represented by $S(T, r)$. This expectation therefore follows a binomial distribution, and it's variance follows accordingly in Equation A.5.

However, to show that this method also works for null distributions that are not analytically solvable, I construct a null distribution based on simulated draws from the null model. As the number of trials increases, the simulated null distribution of the colored triad census should asymptotically approach the analytical solution shown above. For each of $1,000$ trials, I draw random networks from the null distribution, and run the triad census on all these networks. Comparing the observed count to the null distribution then allows us to get an approximate p-value for a conditional uniform graph test, and test the over- or under-representation of each colored triad. I now turn to these results.

A.6 Results

Figure A.3 is a heatmap of the approximate p-values associated with each binomial exact test against the null for each triad, clustered by the triad and the colored triplet as returned by the proposed algorithm. I use a clustering algorithm to group color triplets with similar profiles across the types of triads. This assists with identifying trends across different colored triads, leading to conclusions that would likely be missed if all the colored triads were individually examined. I find particular importance in three branch cutpoints in the clustering algorithm on the color triplets. The first branch in the clustering algorithm (A in Figure A.3) separates four color triplets, comprising 16 colored triads, with a pattern of over-observed 003 and 102

triads, and under-observed 201 and 300 triads. These results show that these color triplets are those that are less clustered than expected by chance. The color triplets all contain nodes of two factions with the first two nodes being Hs, that is, those strongly aligned with the part-time instructor. This indicates that those who are so aligned are likely to form ties to one another, but not to members of other factions. The only exception in this group is that two Hs nodes are more likely to form a tie from one of the Hs members to a Hw member, but even in this case the complete triad (003) is still observed less than expected by chance. This particular result is, perhaps, unsurprising, since Hs and Hw members are close in alignment, more so than with those aligning with the president. Therefore, given the tendency towards homophily they are likely to overlap, though less strongly than members of the same faction; hence, the under-observed $T_{003-HsHsHw}$.

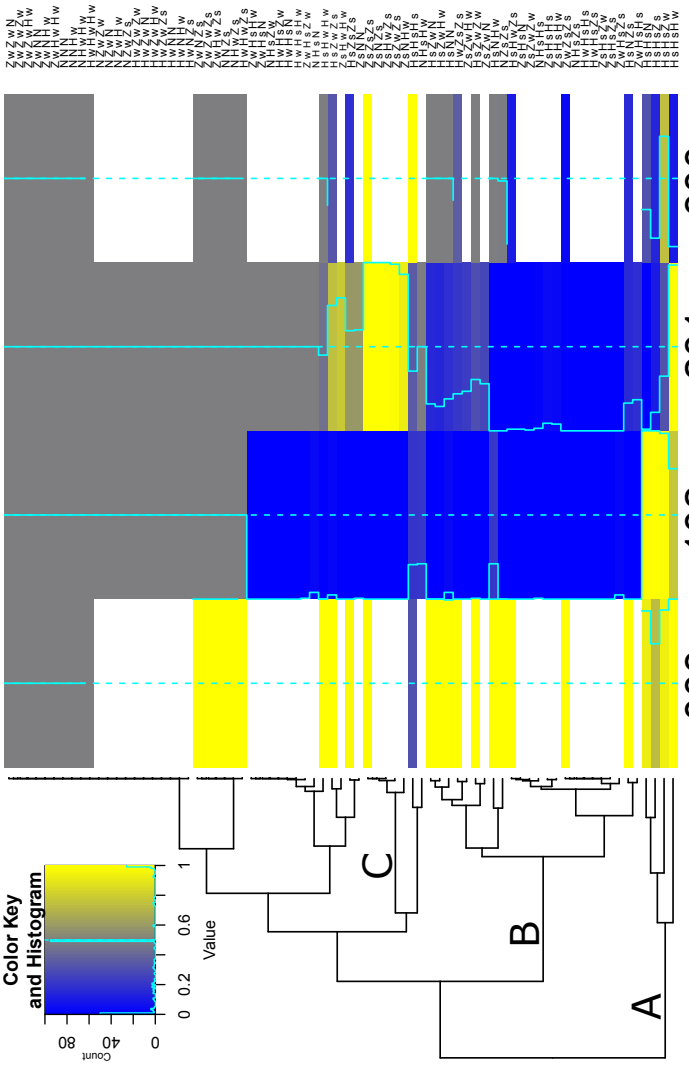

Figure A.3: Heatmap of colored triads and their corresponding p-value of how often they were observed in the empirical networks relative to the null distribution. The columns separate triads based on the MAN configuration, and the rows separate triads based on the triplet of colors. Standard clustering algorithms were used to create the dendrograms. White space indicates redundant isomorphism classes. Gray boxes are either those with 0 triads observed in the network or in any of the networks of the null distribution, and therefore have an undefined pseudo p-value, or those with a pseudo p-value of 0.5. The three labels correspond to three breakpoints in the clustering that separate meaningful groups. (A) is a group of four color triplets exhibiting homophily between Hs nodes. (B) is a group of 21 colored triplets exhibiting low clustering between heterogeneous nodes. (C) is a group of 6 colored triplets that show potential significant amounts of bridging.

The second branching point in the clustering (B in Figure A.3) separates the group of color triplets that are over-observed for the 003 triad, under-observed for the 102 and 201 triads, and observed about as much as expected for the 300 triads. All the triplets in question have nodes of different factions in the first and second position. Because the edge in the 102 triad is between the first and second node in the triplet (Figure A.1), this means that these are all triplets where the first edge is less likely than expected by chance, and the lack of formation of the first edge subsequently hampers the formation of the edge between the second and third nodes in the triplet (201 triad). The first two nodes of these triplets are often (e.g., 16 out of 21) from two factions at least a distance of two away (e.g. N and Hs), indicating members of a faction are not likely to overlap with members who are too disparate from their faction. Put another way, this pattern of triads shows a lack of faction heterophily.

The third branch point (unlabeled) is primarily singling out the group of color triplets that were not observed in the network, and I cannot draw conclusions about their prevalence. The fourth branch point (C in Figure A.3), however, distinguishes a group of five triplets that are under-observed for the 102 triad and over-observed for the 201 triad. This means that the edge between the first two nodes is less likely than expected by chance, but once that edge does occur, the second edge occurs more often than expected by chance. All these triplets begin with a Zs member, and the 201 triad in this case is effectively a bridging tie between it and another. Interestingly, the bridging node is anything other than an Hs (whom are primarily consigned to this role in branch A, as discussed above). The third node was another Zs member in four of five triplets. This indicates that Zs members of the karate club did not often overlap members of other factions, but when they did, provided it was not with an Hs, that second person also often overlapped with another Zs.

Although the above examples show homophily and bridging, analyzing the full colored triad census allows us to draw further conclusions by looking at other

colored triads. In particular, the homophily has mostly been a story of the Hs nodes, and the bridging primarily about the Zs nodes. The 300 triad of both of these factions, when comprising three nodes of the same faction, are observed more often than expected by chance in both cases, which has different implications on the previously-noted results. For the Hs nodes, homophily is strengthened, as not only do Hs nodes not often overlap with members of other factions, they also very strongly overlap with one another. This may partially be an artifact of the types of overlap, as stated before, three of the overlap activities involve direct participation in the part-time instructor's studio, but there are no corresponding groups for the president. This means that those who are Hs or Hw may have more opportunity to overlap with one another due solely to the structure of the data. On the other hand, the triplet of all Zs members also has an over-observed 300 triad. Although there are other triads that seem to indicate bridging between Zs members (C in Figure A.3), given that Zs members are also densely connected to one-another, the practical effect of these potential bridging ties is reduced. Observing this joint effect of homophily and bridging ties was possible only through the complete colored triad census. Neither a standard triad census nor a brokerage analysis would have revealed the intricacies of these results.

In sum, it is clear from these results that the colored triad census allows one to examine multiple trends simultaneously that are often done in isolated analyses, including homophily, heterophily, and brokerage. Importantly, it also allows for generalizations based on the clustering of various triads or color triplets, as well as specific results based on individual triads. In this manner, the colored triad census can yield results on multiple structural levels simultaneously, all while examining local structure, nodal attributes, and their interaction—that is, net of all alternatives involving mixtures of node coloring and triadic configurations.

A.7 Limitations

There are some limitations to this method. First, it is only computationally efficient relative to existing methods (including brute force counting). Networks of $10,000$ nodes or more will take over a day to run using the proposed algorithm for the colored triad census. However, this is an easily parallelizable process (by partitioning the separate algebraic steps, for example), and so the real time necessary to run the analysis can be greatly reduced by taking advantage of this feature. The time needed for the parallelized colored triad census is approximately inversely proportional to the number of computational cores used in the calculation (plus some overhead). Second, the interpretation and visualization of these results is complicated, particularly as the number of colors increases. Examining all of the triads simultaneously reduces the likelihood of missing interesting results because a specific colored triad was excluded. However, the sheer number of colored triads means that making complete sense of results can be difficult. Even if the results are carefully examined for all colored triads, it is conceivable that one might miss an important result out of the $11,080$ colored triads in a directed, 10-color network, no matter how meticulous the examiner's eye. However, use of standard clustering algorithms and heatmaps (as in Figure A.3) may help to ease interpretation of the results at both a coarse- (general groups of triads) or fine-grained (individual colored triads) perspective. That said, I recognize that the interpretation is not straightforward and that this is a first effort at understanding these results, but I believe that having an algorithm to efficiently calculate the colored triad census will spur additional work towards interpreting and using the results. As a result better approaches therein will emerge with time and use.

A.8 Conclusions

In this paper, I have extended the matrix algebra methods of Moody[108] to calculate the colored triad census for any network, directed or undirected, with an arbitrary number of colors in a relatively computationally efficient manner. I have shown a number of mathematical results regarding the colored triad census, including a generalized equation for an arbitrary colored triad, the number of isomorphism classes for arbitrary numbers of colors, and the expectation and variances for colored triads. I analyzed an empirical social network using the algorithm, and calculated approximate p-values for each colored triad, based on an analytic exact binomial test for less complex null distributions, or approximately through simulation for more complex null distributions. I have also shown the type of conclusions that can be drawn from these results, observing results that would not be feasible with many other currently available methods.

One additional benefit of this method is that it can be directly used as a counting tool for sufficient statistics in network inference models, such as exponential random graph models (ERGM). The colored triad census essentially allows one to simultaneously evaluate the effect of local structure and node attribute on network structure in an ERGM, building off previous work where researchers explicated the ERGMs capacity for including the triad census[249]. I believe that the colored triad census is a useful technique with an efficient implementation that can be widely-applicable in social networks research, showing the continued importance of the triad census even in this era of stochastic models for complex networks.

Bibliography

[1] US Department of Health, Human Services, et al. The health consequences of smoking—50 years of progress: a report of the surgeon general. *Atlanta, GA: US Department of Health and Human Services, Centers For Disease Control and Prevention, National Center For Chronic Disease Prevention and Health Promotion, office On Smoking and Health*, 17, 2014.

[2] Sheldon Cohen and Thomas A Wills. Stress, social support, and the buffering hypothesis. *Psychological Bulletin*, 98(2):310–357, 1985. ISSN 0033-2909. doi: 10.1037//0033-2909.98.2.310.

[3] Andrew J Tatem, David J Rogers, and Simon I Hay. Global transport networks and infectious disease spread. *Advances in Parasitology*, 62:293–343, 2006.

[4] Irene H Yen and S Leonard Syme. The social environment and health: a discussion of the epidemiologic literature. *Annual Review of Public Health*, 20 (1):287–308, 1999.

[5] Richard G Wilkinson and Michael Marmot. *Social determinants of health: the solid facts*. World Health Organization, 2003.

[6] Laura L Carstensen. Social and emotional patterns in adulthood: support for socioemotional selectivity theory. *Psychology and Aging*, 7(3):331, 1992.

[7] Paul Drew and John Heritage. *Talk at work: Interaction in institutional settings*. Cambridge Univ Pr, 1992.

[8] Sheldon Cohen. Social relationships and health. *American Psychologist*, 59 (8):676, 2004.

[9] Ichiro Kawachi and Lisa F Berkman. Social ties and mental health. *Journal of Urban Health*, 78(3):458–467, 2001.

[10] Ichiro Kawachi and Lisa F Berkman. *Neighborhoods and health*. Oxford University Press, 2003.

[11] Kristina Orth-Gomer and Jeffrey V Johnson. Social network interaction and mortality: A six year follow-up study of a random sample of the swedish population. *Journal of Chronic Diseases*, 40(10):949–957, 1987.

[12] Maarit Kauppi, Ichiro Kawachi, George David Batty, Tuula Oksanen, Marko Elovainio, Jaana Pentti, Ville Aalto, Marianna Virtanen, Markku Koskenvuo, Jussi Vahtera, et al. Characteristics of social networks and mortality risk: Evidence from two prospective cohort studies. *American Journal of Epidemiology*, 2017.

[13] Mark S Granovetter. The strength of weak ties. In *Social networks*, pages 347–367. Elsevier, 1977.

[14] Leonard I Pearlin, Scott Schieman, Elena M Fazio, and Stephen C Meersman. Stress, health, and the life course: Some conceptual perspectives. *Journal of Health and Social Behavior*, 46(2):205–219, 2005.

[15] Sheldon Cohen, Ronald C Kessler, Lynn U Gordon, et al. Strategies for measuring stress in studies of psychiatric and physical disorders. *Measuring Stress: a Guide For Health and Social Scientists*, pages 3–26, 1995.

[16] Berton H Kaplan, John C Cassel, and Susan Gore. Social support and health. *Medical Care*, 15(5):47–58, 1977.

[17] Catherine Schaefer, James C Coyne, and Richard S Lazarus. The health-related functions of social support. *Journal of Behavioral Medicine*, 4(4): 381–406, 1981.

[18] Sheldon Cohen, William J Doyle, David P Skoner, Bruce S Rabin, and Jack M Gwaltney. Social ties and susceptibility to the common cold. *JAMA*, 277(24): 1940–1944, 1997.

[19] Bert N Uchino. Social support and health: a review of physiological processes potentially underlying links to disease outcomes. *Journal of Behavioral Medicine*, 29(4):377–387, 2006.

[20] Robert B Zajonc. Social Facilitation. *Science*, 149(3681):269–274, 1965.

[21] Jo Corbett, Martin J Barwood, Alex Ouzounoglou, Richard Thelwell, and Matthew Dicks. Influence of competition on performance and pacing during cycling exercise. *Medicine & Science in Sports & Exercise*, 44(3):509–515, 2012.

[22] Hazel Markus. The effect of mere presence on social facilitation: An unobtrusive test. *Journal of Experimental Social Psychology*, 14(4):389–397, 1978.

[23] Tegan Cruwys, Kirsten E Bevelander, and Roel CJ Hermans. Social modeling of eating: A review of when and why social influence affects food intake and choice. *Appetite*, 86:3–18, 2015.

[24] C Peter Herman. The social facilitation of eating. a review. *Appetite*, 86: 61–73, 2015.

[25] Daniel M Landers and Stephen H Boutcher. Arousal-performance relationships. *Applied Sport Psychology: Personal Growth to Peak Performance*, 4:206–228, 1986.

[26] Charles F. Bond and Linda J. Titus. Social facilitation: A meta-analysis of 241 studies. *Psychological Bulletin*, 94(2):265–292, 1983. ISSN 0033-2909. doi: 10.1037/0033-2909.94.2.265.

[27] Stephen W Epley. Reduction of the behavioral effects of aversive stimulation by the presence of companions. *Psychological Bulletin*, 81(5):271, 1974.

[28] Ryan K Masters, Eric N Reither, Daniel A Powers, Y Claire Yang, Andrew E Burger, and Bruce G Link. The impact of obesity on us mortality levels: the importance of age and cohort factors in population estimates. *American Journal of Public Health*, 103(10):1895–1901, 2013.

[29] Cora Lynn Craig, Christine Cameron, Storm J Russell, and Angèle Beaulieu. Increasing physical activity. *Canadian Fitness and Lifestyle Research institute, Ottawa, Ontario*, 2001.

[30] Diane L Gill. Competitiveness among females and males in physical activity classes. *Sex Roles*, 15(5-6):233–247, 1986.

[31] James Paget Henry and Patricia M Stephens. *Stress, health, and the social environment: A sociobiologic approach to medicine*. Springer Science & Business Media, 2013.

[32] Ralf Schwarzer. Modeling health behavior change: How to predict and modify the adoption and maintenance of health behaviors. *Applied Psychology*, 57(1): 1–29, 2008.

[33] Alan E Kazdin. Effects of covert modeling, multiple models, and model reinforcement on assertive behavior. *Behavior therapy*, 7(2):211–222, 1976.

[34] Urte Scholz, Falko F Sniehotta, and Ralf Schwarzer. Predicting physical exercise in cardiac rehabilitation: The role of phase-specific self-efficacy beliefs. *Journal of Sport and Exercise Psychology*, 27(2):135–151, 2005.

[35] Britta Renner, Sunkyo Kwon, Byung-Hwan Yang, Ki-Chung Paik, Seok Hyeon Kim, Sungwon Roh, Jaechul Song, and Ralf Schwarzer. Social-cognitive predictors of dietary behaviors in south korean men and women. *International Journal of Behavioral Medicine*, 15(1):4–13, 2008.

[36] Ralf Schwarzer, Benjamin Schüz, Jochen P Ziegelmann, Sonia Lippke, Aleksandra Luszczynska, and Urte Scholz. Adoption and maintenance of four health behaviors: Theory-guided longitudinal studies on dental flossing, seat belt use, dietary behavior, and physical activity. *Annals of Behavioral Medicine*, 33(2):156–166, 2007.

[37] Anne W Garcia and Abby C King. Predicting long-term adherence to aerobic exercise: A comparison of two models. *Journal of Sport and Exercise Psychology*, 13(4):394–410, 1991.

[38] Nicholas A Christakis and James H Fowler. The spread of obesity in a large social network over 32 years. *The New England Journal of Medicine*, 357(4): 370–379, 2007.

[39] Nathan Eagle, Alex Sandy Pentland, and David Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, 106(36):15274–15278, 2009.

[40] Carlyne L Kujath. Facebook and myspace: Complement or substitute for face-to-face interaction? *Cyberpsychology, Behavior, and Social Networking*, 14(1-2):75–78, 2011.

[41] Allison Davis, Burleigh Bradford Gardner, and Mary R Gardner. *Deep South: A social anthropological study of caste and class*. Univ of South Carolina Press, 2009.

[42] Moira Burke, Cameron Marlow, and Thomas Lento. Social network activity and social well-being. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1909–1912. ACM, 2010.

[43] Susan C Duncan, Terry E Duncan, and Lisa A Strycker. Sources and types of social support in youth physical activity. *Health Psychology*, 24(1):3, 2005.

[44] Y Lunsky and BA Benson. Association between perceived social support and strain, and positive and negative outcome for adults with mild intellectual disability. *Journal of intellectual Disability Research*, 45(2):106–114, 2001.

[45] Marcel Salathé, Maria Kazandjieva, Jung Woo Lee, Philip Levis, Marcus W Feldman, and James H Jones. A high-resolution human contact network for infectious disease transmission. *Proceedings of the National Academy of Sciences*, page 201009094, 2010.

[46] Francesco Calabrese, Francisco C Pereira, Giusy Di Lorenzo, Liang Liu, and Carlo Ratti. The geography of taste: analyzing cell-phone mobility and social events. In *International conference on pervasive computing*, pages 22–37. Springer, 2010.

[47] DB Roszak and RR Colwell. Survival strategies of bacteria in the natural environment. *Microbiological Reviews*, 51(3):365, 1987.

[48] World Health Organization et al. *Contact tracing during an outbreak of Ebola virus disease*. World Health Organization, 2014.

[49] Mirjam Kretzschmar, Susan Van den Hof, Jacco Wallinga, and Jan Van Wijngaarden. Ring vaccination and smallpox control. *Emerging infectious Diseases*, 10(5):832, 2004.

[50] Denis Gerlier. Anti-ebola vaccination of humans using a chimeric virus: rational of a hope. *Virologie*, 19(5):197–203, 2015.

[51] Jacob Levy Moreno. *Who shall survive?: A new approach to the problem of human interrelations.* Nervous and Mental Disease Publishing Co, 1934.

[52] Alessandro Vespignani. Twenty years of network science. *Nature*, 2018.

[53] Stanley Wasserman and Katherine Faust. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.

[54] Kristina Orth-Gomer and Jeffrey V Johnson. Social network interaction and mortality: a six year follow-up study of a random sample of the swedish population. *Journal of Chronic Diseases*, 40(10):949–957, 1987.

[55] Dean Lusher, Johan Koskinen, and Garry Robins. *Exponential random graph models for social networks: Theory, methods, and applications*. Cambridge University Press, 2013.

[56] Kayla De La Haye, Garry Robins, Philip Mohr, and Carlene Wilson. Homophily and contagion as explanations for weight similarities among adolescent friends. *Journal of Adolescent Health*, 49(4):421–427, 2011. ISSN 1054139X. doi: 10.1016/j.jadohealth.2011.02.008.

[57] Yun Huang, Cuihua Shen, Dmitri Williams, and Noshir Contractor. Virtually there: Exploring proximity and homophily in a virtual world. In *Computational Science and Engineering, 2009. CSE'09. International Conference on*, volume 4, pages 354–359. IEEE, 2009.

[58] Pavel N Krivitsky, Mark S Handcock, et al. tergm: Fit, simulate and diagnose models for network evolution based on exponential-family random graph models. *The Statnet Project (Http://www. Statnet. Org). R Package Version*, 3(1), 2017.

[59] Tom AB Snijders. Stochastic actor-oriented models for network change. *Journal of Mathematical Sociology*, 21(1-2):149–172, 1996.

[60] Petter Holme and Jari Saramäki. Temporal networks. *Physics Reports*, 519 (3):97–125, 2012.

[61] Petter Holme. Information content of contact-pattern representations and predictability of epidemic outbreaks. *Scientific Reports*, 5:14462, 2015.

[62] Michael J Barber. Modularity and community detection in bipartite networks. *Physical Review E*, 76(6):066102, 2007.

[63] Mark EJ Newman, Steven H Strogatz, and Duncan J Watts. Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, 64(2):026118, 2001.

[64] Andrew JK Conlan, Ken TD Eames, Jenny A Gage, Johann C von Kirchbach, Joshua V Ross, Roberto A Saenz, and Julia R Gog. Measuring social networks in british primary schools through scientific engagement. *Proceedings of the Royal Society of London B: Biological Sciences*, page rspb20101807, 2010.

[65] Stephen Eubank, Hasan Guclu, VS Anil Kumar, Madhav V Marathe, Aravind Srinivasan, Zoltan Toroczkai, and Nan Wang. Modelling disease outbreaks in realistic urban social networks. *Nature*, 429(6988):180, 2004.

[66] Aamena Alshamsi, Fabio Pianesi, Bruno Lepri, Alex Pentland, and Iyad Rahwan. Beyond contagion: Reality mining reveals complex patterns of social influence. *PloS One*, 10(8):e0135740, 2015.

[67] Alain Barrat and Ciro Cattuto. Face-to-face interactions. In *Social Phenomena*, pages 37–57. Springer, 2015.

[68] JM Read, WJ Edmunds, S Riley, J Lessler, and DAT Cummings. Close encounters of the infectious kind: methods to measure social mixing behaviour. *Epidemiology & infection*, 140(12):2117–2130, 2012.

[69] W John Edmunds, CJ O'Callaghan, and DJ Nokes. Who mixes with whom? A method to determine the contact patterns of adults that may lead to the spread of airborne infections. *Proceedings of the Royal Society of London B: Biological Sciences*, 264(1384):949–957, 1997.

[70] Timo Smieszek, Elena U Burri, Robert Scherzinger, and Roland W Scholz. Collecting close-contact social mixing data with contact diaries: reporting errors and biases. *Epidemiology & infection*, 140(4):744–752, 2012.

[71] Timo Smieszek, Victoria C Barclay, Indulaxmi Seeni, Jeanette J Rainey, Hongjiang Gao, Amra Uzicanin, and Marcel Salathé. How should social mixing be measured: comparing web-based survey and sensor-based methods. *BMC infectious Diseases*, 14(1):136, 2014.

[72] Laura Fumanelli, Marco Ajelli, Piero Manfredi, Alessandro Vespignani, and Stefano Merler. Inferring the structure of social contacts from demographic data in the analysis of infectious diseases spread. *PLoS Computational Biology*, 8(9):e1002673, 2012.

[73] Fabrizio Iozzi, Francesco Trusiano, Matteo Chinazzi, Francesco C Billari, Emilio Zagheni, Stefano Merler, Marco Ajelli, Emanuele Del Fava, and Piero Manfredi. Little Italy: an agent-based approach to the estimation of contact patterns-fitting predicted matrices to serological data. *PLoS Computational Biology*, 6(12):e1001021, 2010.

[74] Pan Hui, Augustin Chaintreau, James Scott, Richard Gass, Jon Crowcroft, and Christophe Diot. Pocket switched networks and human mobility in conference environments. In *Proceedings of the 2005 ACM SIGCOMM workshop on Delay-tolerant networking*, pages 244–251. ACM, 2005.

[75] Ciro Cattuto, Wouter Van den Broeck, Alain Barrat, Vittoria Colizza, Jean-François Pinton, and Alessandro Vespignani. Dynamics of person-to-person interactions from distributed rfid sensor networks. *PloS One*, 5(7):e11596, 2010.

[76] Mathieu Génois, Christian L Vestergaard, Julie Fournet, André Panisson, Isabelle Bonmarin, and Alain Barrat. Data on face-to-face contacts in an office building suggest a low-cost vaccination strategy based on community linkers. *Network Science*, 3(3):326–347, 2015.

[77] Julie Fournet and Alain Barrat. Contact patterns among high school students. *PloS One*, 9(9):e107878, 2014.

[78] Nicolas Voirin, Cécile Payet, Alain Barrat, Ciro Cattuto, Nagham Khanafer, Corinne Régis, Byeul-a Kim, Brigitte Comte, Jean-Sébastien Casalegno, Bruno Lina, et al. Combining high-resolution contact data with virological data to investigate influenza transmission in a tertiary care hospital. *infection Control & Hospital Epidemiology*, 36(3):254–260, 2015.

[79] Leon Danon, Jonathan M Read, Thomas A House, Matthew C Vernon, and Matt J Keeling. Social encounter networks: characterizing great britain. *Proc. R. Soc. B*, 280(1765):20131037, 2013.

[80] Anna Machens, Francesco Gesualdo, Caterina Rizzo, Alberto E Tozzi, Alain Barrat, and Ciro Cattuto. An infectious disease model on empirical networks of human contact: bridging the gap between dynamic network data and contact matrices. *BMC infectious Diseases*, 13(1):185, 2013.

[81] Ronald S Burt. Network items and the general social survey. *Social Networks*, 6(4):293–339, 1984.

[82] Claude Lévi-Strauss. *The elementary structures of kinship*. Number 340. Beacon Press, 1971.

[83] Matthew J Salganik. *Bit by bit: social research in the digital age*. Princeton University Press, 2017.

[84] Samuel Fosso Wamba, Shahriar Akter, Andrew Edwards, Geoffrey Chopin, and Denis Gnanzou. How 'Big Data' can make big impact: Findings from a systematic review and a longitudinal case study. *International Journal of Production Economics*, 165:234–246, 2015.

[85] R Webster Crowley, Hian K Yeoh, George J Stukenborg, Ricky Medel, Neal F Kassell, and Aaron S Dumont. Influence of weekend hospital admission on short-term mortality after intracerebral hemorrhage. *Stroke*, 40(7):2387–2392, 2009.

[86] Tim Harford. Big data: A big mistake? *Significance*, 11(5):14–19, 2014.

[87] Bingshan Li and Suzanne M Leal. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *The American Journal of Human Genetics*, 83(3):311–321, 2008.

[88] Frank Kalter, Anthony F Heath, Miles Hewstone, J Jonsson, Matthijs Kalmijn, Irena Kogan, Frank Van Tubergen, C Kroneberg, L Andersson Rydell, S Brolin Låftman, et al. Children of immigrants longitudinal survey in four european countries (cils4eu)[dataset]. 2013.

[89] Raudenbush Stephen and Bryk Anthony. *Hierarchical Linear Models*. Sage Publications, Thousand Oaks, CA, 2002.

[90] Karthik Kambatla, Giorgos Kollias, Vipin Kumar, and Ananth Grama. Trends in big data analytics. *Journal of Parallel and Distributed Computing*, 74(7): 2561–2573, 2014.

[91] Jianqing Fan, Fang Han, and Han Liu. Challenges of big data analysis. *National Science Review*, 1(2):293–314, 2014.

[92] Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J Niels Rosenquist. Understanding the demographics of twitter users. *ICWSM*, 11(5th):25, 2011.

[93] Hugh Louch. Personal network integration: transitivity and homophily in strong-tie relations. *Social Networks*, 22(1):45–64, 2000.

[94] Giuliana Carullo, Aniello Castiglione, Alfredo De Santis, and Francesco Palmieri. A triadic closure and homophily-based recommendation system for online social networks. *World Wide Web*, 18(6):1579–1601, 2015.

[95] Carole Cadwalladr and Emma Graham-Harrison. Revealed: 50 million facebook profiles harvested for cambridge analytica in major data breach. *The Guardian*, 17, 2018.

[96] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *Security and Privacy, 2008. SP 2008. IEEE Symposium on*, pages 111–125. IEEE, 2008.

[97] Jan Philipp Albrecht. How the GDPR will change the world. *Eur. Data Prot. L. Rev.*, 2:287, 2016.

[98] Stuart Barton. Which clinical studies provide the best evidence?: The best RCT still trumps the best observational study. *BMJ: British Medical Journal*, 321(7256):255, 2000.

[99] Ashish K Jha, Catherine M DesRoches, Eric G Campbell, Karen Donelan, Sowmya R Rao, Timothy G Ferris, Alexandra Shields, Sara Rosenbaum, and David Blumenthal. Use of electronic health records in US hospitals. *New England Journal of Medicine*, 360(16):1628–1638, 2009.

[100] Lisa I Iezzoni. Assessing quality using administrative data. *Annals of internal Medicine*, 127(8):666–674, 1997.

[101] World Health Organization. *International Statistical Classification of Diseases and Related Health Problems, 10th Revision (ICD-10)*. WHO, Geneva, 1992.

[102] Sandra K Thygesen, Christian F Christiansen, Steffen Christensen, Timothy L Lash, and Henrik T Sørensen. The predictive value of icd-10 diagnostic coding used to assess charlson comorbidity index conditions in the population-based danish national registry of patients. *BMC Medical Research Methodology*, 11 (1):83, 2011.

[103] John W Peabody, Jeff Luck, Sharad Jain, Dan Bertenthal, and Peter Glassman. Assessing the accuracy of administrative data in health information systems. *Medical Care*, 42(11):1066–1072, 2004.

[104] Riitta A Marjamaa, Paulus M Torkki, Markus I Torkki, and Olli A Kirvelä. Time accuracy of a radio frequency identification patient tracking system for recording operating room timestamps. *Anesthesia & Analgesia*, 102(4): 1183–1186, 2006.

[105] Graham P Martin. 'Ordinary people only': knowledge, representativeness, and the publics of public participation in healthcare. *Sociology of Health & Illness*, 30(1):35–54, 2008.

[106] Oliver J Old, Richard J Egan, Sally A Norton, and Justin DT Morgan. Ethnic minorities have equal access to bariatric surgery in the uk and ireland. *Obesity Surgery*, 23(5):727–729, 2013.

[107] Marci Meingast, Tanya Roosta, and Shankar Sastry. Security and privacy issues with health care information technology. In *Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE*, pages 5453–5458. IEEE, 2006.

[108] James Moody. Matrix methods for calculating the triad census. *Social Networks*, 20(4):291–299, 1998. ISSN 03788733. doi: 10.1016/S0378-8733(98) 00006-9.

[109] Wayne W Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33(4):452–473, 1977.

[110] A Sarah Walker, David W Eyre, David H Wyllie, Kate E Dingle, David Griffiths, Brian Shine, Sarah Oakley, Lily O'Connor, John Finney, Alison Vaughan, et al. Relationship between bacterial strain type, host biomarkers, and mortality in clostridium difficile infection. *Clinical infectious Diseases*, 56 (11):1589–1600, 2013.

[111] John M Finney, A Sarah Walker, Tim EA Peto, and David H Wyllie. An efficient record linkage scheme using graphical analysis for identifier error detection. *BMC Medical informatics and Decision Making*, 11(1):7, 2011.

[112] Hude Quan, Bing Li, L Duncan Saunders, Gerry A Parsons, Carolyn I Nilsson, Arif Alibhai, and William A Ghali. Assessing validity of icd-9-cm and icd-10 administrative data in recording clinical conditions in a unique dually coded database. *Health Services Research*, 43(4):1424–1441, 2008.

[113] Oxfordshire's Health and Wellbeing Board. Joint strategic needs assessment report 2017. Technical report, University of Oxfordshire Hospitals, Mar 2017.

[114] World Health Organization. *International statistical classification of diseases and related health problems. - 10th revision, Fifth edition.* 2016. ISBN 978 92 4 154916 5.

[115] John N Newton, Adam DM Briggs, Christopher JL Murray, Daniel Dicker, Kyle J Foreman, Haidong Wang, Mohsen Naghavi, Mohammad H Forouzanfar, Summer Lockett Ohno, Ryan M Barber, et al. Changes in health in england, with analysis by english regions and areas of deprivation, 1990–2013: a systematic analysis for the global burden of disease study 2013. *The Lancet*, 386(10010):2257–2274, 2015.

[116] Cancer Research UK. http://www.cancerresearchuk.org/content/cancer-mortality-statistics, 2012.

[117] Jacob E. Simmering, Linnea A. Polgreen, David R. Campbell, Joseph E. Cavanaugh, and Philip M. Polgreen. Hospital transfer network structure as a risk factor for clostridium difficile infection. *Infection Control; Hospital Epidemiology*, 36(9):1031â€“1037, 2015. doi: 10.1017/ice.2015.130.

[118] Mandy Odell. The patient's thoughts and feelings about their transfer from intensive care to the general ward. *Journal of Advanced Nursing*, 31(2): 322–329, 2000.

[119] Chiung-Jung Jo Wu and Fiona Coyer. Reconsidering the transfer of patients from the intensive care unit to the ward: a case study approach. *Nursing & Health Sciences*, 9(1):48–53, 2007.

[120] Tao Zhou, Jie Ren, Matúš Medo, and Yi-Cheng Zhang. Bipartite network projection and personal recommendation. *Physical Review E*, 76(4):046115, 2007.

[121] Evan S Snitkin, Adrian M Zelazny, Pamela J Thomas, Frida Stock, David K Henderson, Tara N Palmore, Julia A Segre, NISC Comparative Sequencing Program, et al. Tracking a hospital outbreak of carbapenem-resistant klebsiella pneumoniae with whole-genome sequencing. *Science Translational Medicine*, 4(148):148ra116–148ra116, 2012.

[122] Ursula Beckmann, Donna M Gillies, Sean M Berenholtz, Albert W Wu, and Peter Pronovost. Incidents relating to the intra-hospital transfer of critically ill patients. *Intensive Care Medicine*, 30(8):1579–1585, 2004.

[123] Jacob E Simmering, Linnea A Polgreen, David R Campbell, Joseph E Cavanaugh, and Philip M Polgreen. Hospital transfer network structure as a risk factor for clostridium difficile infection. *infection Control & Hospital Epidemiology*, 36(9):1031–1037, 2015.

[124] Theodore J Iwashyna, Jason D Christie, Jeremy M Kahn, and David A Asch. Uncharted paths: hospital networks in critical care. *Chest*, 135(3):827–833, 2009.

[125] Kwang-Il Goh, Michael E Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-László Barabási. The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690, 2007.

[126] Amy N Duckro, Donald W Blom, Elizabeth A Lyle, Robert A Weinstein, and Mary K Hayden. Transfer of vancomycin-resistant enterococci via health care worker hands. *Archives of internal Medicine*, 165(3):302–307, 2005.

[127] N Bossard, M Velten, L Remontet, a Belot, N Maarouf, a M Bouvier, a V Guizard, B Tretarre, G Launoy, M Colonna, a Danzon, F Molinie, X Troussard, N Bourdon-Raverdy, P M Carli, a Jaffré, C Bessaguet, E Sauleau, C Schvartz, P Arveux, M Maynadié, P Grosclaude, J Estève, and J Faivre. Survival of cancer patients in France: a population-based study from The Association of the French Cancer Registries (FRANCIM). *European Journal of Cancer (Oxford, England : 1990)*, 43(1):149–60, jan 2007. ISSN 0959-8049. doi: 10.1016/j.ejca.2006.07.021.

[128] M P Coleman, D Forman, H Bryant, J Butler, B Rachet, C Maringe, U Nur, E Tracey, M Coory, J Hatcher, C E McGahan, D Turner, L Marrett, M L Gjerstorff, T B Johannesen, J Adolfsson, M Lambe, G Lawrence, D Meechan, E J Morris, R Middleton, J Steward, and M a Richards. Cancer survival in Australia, Canada, Denmark, Norway, Sweden, and the UK, 1995-2007 (the International Cancer Benchmarking Partnership): an analysis of population-based cancer registry data. *Lancet*, 377(9760):127–38, jan 2011. ISSN 1474-547X. doi: 10.1016/S0140-6736(10)62231-3.

[129] Rebecca L Siegel, Kimberly D Miller, and Ahmedin Jemal. Cancer statistics, 2015. *CA: a Cancer Journal For Clinicians*, 65(1):5–29, jan 2015. ISSN 1542-4863. doi: 10.3322/caac.21254.

[130] Denise Ernst, Clyde R Pope, and Jack F Hollis. Social networks as predictors of ischemic heart disease, cancer, stroke and hypertension: incidence, survival and mortality. *J Clin Epidemiol*, 45(6):659–66, 1992.

[131] Julianne Holt-Lunstad, Timothy B Smith, and J Bradley Layton. Social relationships and mortality risk: a meta-analytic review. *PLoS Medicine*, 7 (7):e1000316, jul 2010. ISSN 1549-1676. doi: 10.1371/Journal.pmed.1000316.

[132] Eran Shor and David J Roelfs. Social contact frequency and all-cause mortality: a meta-analysis and meta-regression. *Social Science & Medicine (1982)*, 128 (1):76–86, mar 2015. ISSN 1873-5347. doi: 10.1016/j.socscimed.2015.01.010.

[133] Edna Maria, Vissoci Reiche, Sandra Odebrecht, Vargas Nunes, and Helena Kaminami Morimoto. Stress, depression, the immune system, and cancer. *The Lancet Oncology*, 5(10):617–625, 2004.

[134] E Tiligada. Chemotherapy: induction of stress responses. *Endocrine-Related Cancer*, 13(1):S115–24, dec 2006. ISSN 1351-0088. doi: 10.1677/erc.1.01272.

[135] S Cohen. Psychosocial models of the role of social support in the etiology of physical disease. *Health Psychology : official Journal of the Division of Health Psychology, American Psychological Association*, 7(3):269–97, 1988.

[136] Elliott a Beaton, Louis a Schmidt, Jay Schulkin, Martin M Antony, Richard P Swinson, and Geoffrey B Hall. Different neural responses to stranger and personally familiar faces in shy and bold adults. *Behavioral Neuroscience*, 122 (3):704–9, jun 2008. ISSN 0735-7044. doi: 10.1037/0735-7044.122.3.704.

[137] Sidney Cobb. Social Support as a Moderator of Life Stress. *Psychometric Medicine*, 38(5):300–314, 1976.

[138] James S House. Social support and social structure. In *Sociological forum*, volume 2, pages 135–146. Springer, 1987.

[139] Yang Claire Yang, Courtney Boen, Karen Gerken, Ting Li, Kristen Schorpp, and Kathleen Mullan Harris. Social relationships and physiological determinants of longevity across the human life span. *Proceedings of the National Academy of Sciences*, 113(3):201511085, jan 2016. ISSN 0027-8424. doi: 10.1073/pnas.1511085112.

[140] Aleksandra Luszczynska, Yagnaseni Sarkar, and Nina Knoll. Received social support, self-efficacy, and finding benefits in disease as predictors of physical functioning and adherence to antiretroviral therapy. *Patient Education and Counseling*, 66(1):37–42, apr 2007. ISSN 0738-3991. doi: 10.1016/j.pec.2006.10.002.

[141] Cynthia A Berg, Deborah J Wiebe, Jonathan Butner, Lindsey Bloor, Chester Bradstreet, Renn Upchurch, John Hayes, Robert Stephenson, Lillian Nail, and Gregory Patton. Collaborative coping and daily mood in couples dealing with prostate cancer. *Psychology and Aging*, 23(3):505–16, sep 2008. ISSN 0882-7974. doi: 10.1037/a0012687.

[142] Robert T. Croyle and Julie R. Hunt. Coping with health threat: Social influence processes in reactions to medical test results. *Journal of Personality and Social Psychology*, 60(3):382–389, 1991. ISSN 0022-3514. doi: 10.1037//0022-3514.60.3.382.

[143] Jeana H Frost and Michael P Massagli. Social uses of personal health information within PatientsLikeMe, an online patient community: what can happen when patients have access to one another's data. *Journal of Medical internet Research*, 10(3):e15, jan 2008. ISSN 1438-8871. doi: 10.2196/jmir.1053.

[144] JB McKinlay. Social networks, lay consultation and help-seeking behavior. *Social Forces*, 51(3):275–292, 1973.

[145] Karen S. Rook. The Negative Side of Social Interaction: Impact on Psychological Well-Being. *Journal of Personality and Social Psychology*, 46(5):1097–1108, 1984. ISSN 0022-3514. doi: 10.1037/0022-3514.46.5.1097.

[146] A Armario, R Ortiz, and J Balasch. Corticoadrenal and behavioral response to open field in paris of male rats either familiar or non-familiar to each other. *Experientia*, 39(11):1316–1317, 1983.

[147] Eileen N. Cain, Ernest I. Kohorn, Donald M. Quinlan, Kate Latimer, and Peter E. Schwartz. Psychosocial benefits of a cancer support group. *Cancer*, 57 (1):183–189, jan 1986. ISSN 0008-543X. doi: 10.1002/1097-0142(19860101)57:1<183::AID-CNCR2820570135>3.0.CO;2-3.

[148] Peggy A Thoits. Mechanisms linking social ties and support to physical and mental health. *Journal of Health and Social Behavior*, 52(2):145–61, jun 2011. ISSN 2150-6000. doi: 10.1177/0022146510395592.

[149] Felix Elwert and Nicholas A Christakis. The Effect of Widowhood on Mortality by the Causes of Death of Both Spouses. *American Journal of Public Health*, 98 (11):2092–2098, nov 2008. ISSN 0090-0036. doi: 10.2105/AJPH.2007.114348.

[150] Geraldine P Mineau, Ken R Smith, and Lee L Bean. Historical trends of survival among widows and widowers. *Social Science & Medicine*, 54(2): 245–254, jan 2002. ISSN 02779536. doi: 10.1016/S0277-9536(01)00024-7.

[151] Mari Oyama, Kazutoshi Nakamura, Yuko Suda, and Toshiyuki Someya. Social network disruption as a major factor associated with psychological distress 3 years after the 2004 Niigata-Chuetsu earthquake in Japan. *Environmental Health and Preventive Medicine*, 17(2):118–23, mar 2012. ISSN 1347-4715. doi: 10.1007/s12199-011-0225-y.

[152] Brea L Perry. Understanding Social Network Disruption : The Case of Youth in Foster Care. *Social Problems*, 53(3):371–391, 2016.

[153] Marilyn T Oberst, Suzanne E Thomas, Kathleen A Gass, and Sandra E Ward. Caregiving demands and apprasial of stress among family caregivers. *Cancer Nursing*, 12(4):209–215, 1989.

[154] Laura M Koehly, June A Peters, Natalia Kuhn, Lindsey Hoskins, Anne Letocha, Regina Kenen, Jennifer Loud, and Mark H Greene. Sisters in hereditary breast and ovarian cancer families: Communal coping, social integration, and psychological well-being. *Psycho-Oncology*, 17(8):812–821, 2011. doi: 10.1002/pon.1373.Sisters.

[155] Liesbeth Mercken, Christian Steglich, Philip Sinclair, Jo Holliday, and Laurence Moore. A longitudinal social network analysis of peer influence, peer selection, and smoking behavior among adolescents in British schools. *Health Psychology: Official Journal of the Division of Health Psychology, American Psychological Association*, 31(4):450–9, jul 2012. ISSN 1930-7810. doi: 10.1037/a0026876.

[156] Jessica M Perkins, S V Subramanian, and Nicholas A Christakis. Social networks and health: A systematic review of sociocentric network studies in low- and middle-income countries. *Social Science & Medicine*, 125(1):60–78, aug 2014. ISSN 1873-5347. doi: 10.1016/j.socscimed.2014.08.019.

[157] Thomas W. Valente, Beth R. Hoffman, Annamara Ritt-Olson, Kara Lichtman, and C. Anderson Johnson. Effects of a Social-Network Method for Group Assignment Strategies on Peer-Led Tobacco Prevention Programs in Schools. *American Journal of Public Health*, 93(11):1837–1843, nov 2003. ISSN 0090-0036. doi: 10.2105/AJPH.93.11.1837.

[158] Cosma Rohila Shalizi and Andrew C Thomas. Homophily and contagion are generically confounded in observational social network studies. *SMR*, 40(2): 211–39, 2012. doi: 10.1177/0049124111404820.Homophily.

[159] Sinan Aral and Dylan Walker. Tie Strength , Embeddedness , and Social Influence : A Large-Scale Networked Experiment. *Management Science*, 6: 1352–70, 2014.

[160] Candyce H Kroenke, Laura D Kubzansky, Eva S Schernhammer, Michelle D Holmes, and Ichiro Kawachi. Social networks, social support, and survival after breast cancer diagnosis. *Journal of Clinical Oncology : official Journal of the American Society of Clinical Oncology*, 24(7):1105–11, mar 2006. ISSN 1527-7755. doi: 10.1200/JCO.2005.04.2846.

[161] Kenji Shibuya, Colin D Mathers, Cynthia Boschi-pinto, Alan D Lopez, and Christopher J L Murray. Global and regional estimates of cancer mortality and incidence by site : II . results for the global burden of disease 2000. *BMC Cancer*, 2(37):1–26, 2002.

[162] MB Amin, S Edge, F Greene, DR Byrd, RK Brookland, MK Washington, JE Gershenwald, CC Compton, KR Hess, DC Sullivan, JM Jessup, JD Brierley, LE Gaspar, RL Schilsky, CM Balch, DP Winchester, EA Asare, M Madera, DM Gress, and LR Mayer. *AJCC Cancer Staging Manual*. American Joint Committee on Cancer, 2017.

[163] Melinda Mills. *Introducing Survival and Event History Analaysis*. SAGE Publications Ltd, London, 2011.

[164] Tyler H McCormick, Matthew J Salganik, and Tian Zheng. How many people do you know?: Efficiently estimating personal network size. *Journal of the American Statistical Association*, 105(489):59–70, 2010. ISSN 0162-1459. doi: 10.1198/jasa.2009.ap08518.

[165] G G Grabenbauer, I H Schneider, F P Gall, and R Sauer. Epidermoid carcinoma of the anal canal: treatment by combined radiation and chemotherapy. *Radiotherapy and Oncology : Journal of the European Society For therapeutic Radiology and Oncology*, 27(1):59–62, 1993. ISSN 0167-8140 (Print).

[166] E Rapp, JL Pater, A Willan, Y Cormier, N Murray, WK Evans, DI Hodson, DA Clark, R Feld, and AM Arnold. Chemotherapy can prolong survival in patients with advanced non-small-cell lung cancer–report of a Canadian multicenter randomized trial. *J. Clin. Oncol.*, 6(4):633–641, 1988.

[167] Department of Health. The NHS Cancer Plan. Technical Report September, National Health Service, 2000.

[168] Tyler J VanderWeele. Sensitivity analysis for contagion effects in social networks. *Sociological Methods & Research*, 40(2):240–255, 2011.

[169] NHS. Cancer Patient Experience Survey 2013 National Report. Technical report, NHS, 2013.

[170] L Gabriel and A Beriot-Mathiot. Hospitalization stay and costs attributable to clostridium difficile infection: a critical review. *Journal of Hospital infection*, 88(1):12–21, 2014.

[171] William R Jarvis. Selected aspects of the socioeconomic impact of nosocomial infections: Morbidity , mortality , cost , and prevention. *Infection Control and Hospital Epidemiology*, 17(8):552–557, 1996.

[172] R Plowman, N Graves, MAS Griffin, JA Roberts, AV Swan, B Cookson, and L Taylor. The rate and cost of hospital-acquired infections occurring in patients admitted to selected specialties of a district general hospital in england and the national burden imposed. *Journal of Hospital infection*, 47 (3):198–209, 2001.

[173] Binila Chacko, Kurien Thomas, Thambu David, Hema Paul, Lakshmanan Jeyaseelan, and John Victor Peter. Attributable cost of a nosocomial infection in the intensive care unit: a prospective cohort study. *World Journal of Critical Care Medicine*, 6(1):79, 2017.

[174] Christophe Adrie, Maité Garrouste-Orgeas, Wafa Ibn Essaied, Carole Schwebel, Michael Darmon, Bruno Mourvillier, Stéphane Ruckly, Anne-Sylvie Dumenil, Hatem Kallel, Laurent Argaud, et al. Attributable mortality of icu-acquired bloodstream infections: impact of the source, causative micro-organism, resistance profile and antimicrobial therapy. *Journal of infection*, 74(2):131–141, 2017.

[175] Didier Pittet, Stephane Hugonnet, Stephan Harbarth, Philippe Mourouga, Valerie Sauvan, Sylvie Touveneau, and Thomas V Perneger. Effectiveness of a hospital-wide programme to improve compliance with hand hygiene. *The Lancet*, 356(9238):1307–1312, 2000. doi: 10.1016/S0140-6736(00)02814-2.

[176] Cheng-Chuan Hsu, Yusen E Lin, Yao-Shen Chen, Yung-Ching Liu, and Robert R Muder. Validation study of artificial neural network models for prediction of methicillin-resistant staphylococcus aureus carriage. *Infection Control and Hospital Epidemiology : the official Journal of the Society of Hospital Epidemiologists of America*, 29(7):607–14, 2008.

[177] Eli N Perencevich, David N Fisman, Marc Lipsitch, Anthony D Harris, J Glenn Morris, and David L Smith. Projected benefits of active surveillance for vancomycin-resistant enterococci in intensive care units. *Clinical infectious Diseases : An official Publication of the infectious Diseases Society of America*, 38(8):1108–1115, 2004. ISSN 1537-6591. doi: 10.1086/382886.

[178] Zia Sadique, Ben Lopman, Ben S Cooper, and W John Edmunds. Cost-effectiveness of ward closure to control outbreaks of norovirus infection in united kingdom national health service hospitals. *The Journal of infectious Diseases*, 213(suppl_1):S19–S26, 2015.

[179] Morten OA Sommer, Christian Munck, Rasmus Vendler Toft-Kehler, and Dan I Andersson. Prediction of antibiotic resistance: time for a new preclinical paradigm? *Nature Reviews Microbiology*, 15(11):689, 2017.

[180] Gemma Johnson, Michael R Millar, Stuart Matthews, Margaret Skyrme, Peter Marsh, Emma Barringer, Stephen O'Hara, and Mark Wilks. Evaluation of baclite rapid mrsa, a rapid culture based screening test for the detection of ciprofloxacin and methicillin resistant s. aureus (mrsa) from screening swabs. *BMC Microbiology*, 6(1):83, 2006.

[181] Longxiang Su, Bingchao Han, Changting Liu, Liling Liang, Zhaoxu Jiang, Jie Deng, Peng Yan, Yanhong Jia, Dan Feng, and Lixin Xie. Value of soluble TREM-1, procalcitonin, and C-reactive protein serum levels as biomarkers for detecting bacteremia among sepsis patients with new fever in intensive care units: a prospective cohort study. *BMC infectious Diseases*, 12(1):157, 2012. ISSN 1471-2334. doi: 10.1186/1471-2334-12-157.

[182] Liliana Simon, France Gauvin, Devendra K Amre, Patrick Saint-Louis, and Jacques Lacroix. Serum procalcitonin and c-reactive protein levels as markers of bacterial infection: A systematic review and meta-analysis. *Clinical infectious Diseases*, 39(2):206–217, 2004.

[183] N Bagdasarian, HC Chan, S Ang, MS Isa, SM Chan, and DA Fisher. A "stone in the pond" approach to contact tracing: Responding to a large-scale, nosocomial tuberculosis exposure in a moderate tb-burden setting. *Infection Control & Hospital Epidemiology*, pages 1–3, 2017.

[184] European Centre for Disease Prevention and Control. Risk assessment guidelines for infectious diseases transmitted on aircraft (RAGIDA) - influenza, 2014.

[185] Evan S Snitkin, Adrian M Zelazny, Pamela J Thomas, Frida Stock, David K Henderson, Tara N Palmore, and Julia a Segre. Tracking a hospital outbreak of carbapenem-resistant Klebsiella pneumoniae with whole-genome sequencing. *Science Translational Medicine*, 4(148):148ra116, aug 2012. ISSN 1946-6242. doi: 10.1126/scitranslmed.3004129.

[186] Centers for Disease Control, Prevention, et al. CDC methods for implementing and managing contact tracing for ebola virus disease in less-affected countries, 2017.

[187] Ruogu Fang, Samira Pouyanfar, Yimin Yang, Shu-Ching Chen, and SS Iyengar. Computational health informatics in the big data age: a survey. *ACM Computing Surveys (CSUR)*, 49(1):12, 2016.

[188] Marco Cusumano-Towner, Daniel Y Li, Shanshan Tuo, Gomathi Krishnan, and David M Maslove. A social network of hospital acquired infection built from electronic medical record data. *Journal of the American Medical informatics Association : JAMIA*, 20(3):427–34, may 2013. ISSN 1527-974X. doi: 10.1136/amiajnl-2012-001401.

[189] Sharon E Perlman, Katharine H McVeigh, Remle Newton-Dame, Lorna E Thorpe, Elisabeth F Snell, Claudia Chernov, Jesse Singer, and Carolyn M Greene. Innovations in population health surveillance using electronic health record data. *Online Journal of Public Health informatics; Vol 6, No 1 (2014)*, pages 1–5, 2014.

[190] Juan Fernández-Gracia, Jukka-Pekka Onnela, Michael L Barnett, Víctor M Eguíluz, and Nicholas A Christakis. Influence of a patient transfer network of US inpatient facilities on the incidence of nosocomial infections. *Scientific Reports*, 7(1):2930, 2017.

[191] Patrick M Bossuyt, Johannes B Reitsma, David E Bruns, Constantine A Gatsonis, Paul P Glasziou, Les Irwig, Jeroen G Lijmer, David Moher, Drummond Rennie, Henrica C W de Vet, Herbert Y Kressel, Nader Rifai, Robert M Golub, Douglas G Altman, Lotty Hooft, Daniël A Korevaar, and Jérémie F Cohen. Stard 2015: an updated list of essential items for reporting diagnostic accuracy studies. *BMJ*, 351, 2015. doi: 10.1136/bmj.h5527.

[192] Teresa C Horan, Mary Andrus, and Margaret A Dudeck. Cdc/nhsn surveillance definition of health care–associated infection and criteria for specific types of infections in the acute care setting. *American Journal of infection Control*, 36(5):309–332, 2008.

[193] Nita Pal, Rajni Sharma, Suman Rishi, and Leela Vyas. Optimum time to detection of bacteria and yeast species with bactec 9120 culture system from blood and sterile body fluids. *Journal of Laboratory Physicians*, 1(2):69–72, 2009.

[194] Roy M Anderson. *The population dynamics of infectious diseases: theory and applications*. Springer Science, 1982.

[195] Public Health Agency of Canada. Pseudomonas spp. pathogen safety data sheet, 2012.

[196] CB Dalton, ED Mintz, JG Wells, CA Bopp, and RV Tauxe. Outbreaks of enterotoxigenic escherichia coli infection in american adults : a clinical and epidemiologic profile. *Epidemiology and infection*, 123(May):9–16, 1999.

[197] Public Health Agency of Canada. Staphylococcus aureus pathogen safety data sheet, 2012.

[198] Stuart H Cohen, Dale N Gerding, Stuart Johnson, Ciaran P Kelly, Vivian G Loo, L Clifford McDonald, Jacques Pepin, and Mark H Wilcox. Clinical practice guidelines for clostridium difficile infection in adults: 2010 update by the society for healthcare epidemiology of america (shea) and the infectious diseases society of america (idsa). *Infection Control & Hospital Epidemiology*, 31(5):431–455, 2010.

[199] JC Rahamat-Langendoen, M Lokate, EH Schölvinck, AW Friedrich, and HGM Niesters. Rapid detection of a norovirus pseudo-outbreak by using real-time sequence based information. *Journal of Clinical Virology : the official Publication of the Pan American Society For Clinical Virology*, 58(1): 245–8, sep 2013.

[200] Benedetta Allegranzi, Sepideh Bagheri Nejad, Christophe Combescure, Wilco Graafmans, Homa Attar, Liam Donaldson, and Didier Pittet. Burden of endemic health-care-associated infection in developing countries: Systematic review and meta-analysis. *The Lancet*, 377(9761):228–241, 2011. ISSN 01406736. doi: 10.1016/S0140-6736(10)61458-4.

[201] P. N. Wiegand, D. Nathwani, M. H. Wilcox, J. Stephens, A. Shelbaya, and S. Haider. Clinical and economic burden of Clostridium difficile infection in Europe: A systematic review of healthcare-facility-acquired infection. *Journal of Hospital infection*, 81(1):1–14, 2012. ISSN 01956701. doi: 10.1016/j.jhin.2012.02.004.

[202] Madhukar Pai, Alice Zwerling, and Dick Menzies. Systematic Review: T-Cell – based Assays for the Diagnosis of Latent Tuberculosis Infection: An Update. *Annals of internal Medicine*, 149(3):177–84, 2008.

[203] L Simon, F Gauvin, Dk Amre, P Saint-Louis, and J Lacroix. Serum procalcitonin and C-reactive protein levels as markers of bacterial infection: a systematic review and meta-analysis. *Clinical infectious Diseases*, 39:206–17, 2004. ISSN 1537-6591. doi: 10.1086/421997.

[204] JM Pollock and SD Neill. Mycobacterium boviss infection and tuberculosis in cattle. *The Veterinary Journal*, 163(2):115–127, 2002.

[205] Anthony O'Hare, RJ Orton, Paul R Bessell, and Rowland R Kao. Estimating epidemiological parameters for bovine tuberculosis in british cattle using a bayesian partial-likelihood approach. *Proceedings of the Royal Society of London B: Biological Sciences*, 281(1783):20140248, 2014.

[206] O Clerc and G Greub. Routine use of point-of-care tests: usefulness and application in clinical microbiology. *Clinical Microbiology and infection*, 16 (8):1054–1061, 2010.

[207] MJ Espy, JR Uhl, LM Sloan, SP Buckwalter, MF Jones, EA Vetter, JDC Yao, NL Wengenack, JE Rosenblatt, FR 3 Cockerill, et al. Real-time PCR in clinical microbiology: applications for routine laboratory testing. *Clinical Microbiology Reviews*, 19(1):165–256, 2006.

[208] Cara M Cannon, Jackson S Musuuza, Anna K Barker, Megan Duster, Mark B Juckett, Aurora E Pop-Vicas, and Nasia Safdar. Risk of *Clostridium difficile* infection in hematology-oncology patients colonized with toxigenic *C. difficile*. *Infection Control & Hospital Epidemiology*, 38(6):718–720, 2017.

[209] Guillaume Gingras, Marie-Hélène Guertin, Jean-François Laprise, Mélanie Drolet, and Marc Brisson. Mathematical modeling of the transmission dynamics of clostridium difficile infection and colonization in healthcare settings: a systematic review. *PloS One*, 11(9):e0163880, 2016.

[210] Ioannis M Zacharioudakis, Fainareti N Zervou, Elina Eleftheria Pliakos, Panayiotis D Ziakas, and Eleftherios Mylonakis. Colonization with toxinogenic *C. difficile* upon hospital admission, and risk of infection: a systematic review and meta-analysis. *The American Journal of Gastroenterology*, 110(3):381, 2015.

[211] Andrew S Ross, Christopher Baliga, Punam Verma, Jeffrey Duchin, and Michael Gluck. A quarantine process for the resolution of duodenoscope-associated transmission of multidrug-resistant escherichia coli. *Gastrointestinal Endoscopy*, 82(3):477–483, 2015.

[212] John D Birkmeyer. Using administative data for clinical research. *Surgical Research*, pages 127–136, 2001.

[213] Robert H Miller and Ida Sim. Physicians' use of electronic medical records: barriers and solutions. *Health Affairs*, 23(2):116–126, 2004.

[214] Kevin J. Downes and Samir S. Shah. Biomarkers in infectious diseases. *Journal of the Pediatric infectious Diseases Society*, 1(4):343–346, 2012. ISSN 20487207. doi: 10.1093/jpids/pis099.

[215] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural Networks*, 32:323–332, 2012.

[216] Xiaojin Zhu. Semi-supervised learning literature survey. *Computer Science, University of Wisconsin-Madison*, 2005.

[217] Maarten van Smeden, Christiana A Naaktgeboren, Johannes B Reitsma, Karel G M Moons, and Joris a H de Groot. Latent Class Models in Diagnostic Studies When There is No Reference Standard–A Systematic Review. *American Journal of Epidemiology*, 179(4):423–431, 2014. ISSN 0002-9262. doi: 10.1093/aje/kwt286.

[218] a. W. Rutjes, J. B. Reitsma, a. Coomarasamy, K. S. Khan, and P. M. Bossuyt. Evaluation of diagnostic tests when there is no gold standard. A review of methods. *Health Technology Assessment*, 11(50):iii, ix–51–, 2007. ISSN 1366-5278. doi: 06/90/23[pii].

[219] Yinsheng Qu, Ming Tan, and Michael H Kutner. Random Effects Models in Latent Class Analysis for Evaluating Accuracy of Diagnostic Tests. *Biometrics*, 52(3):797–810, 1996.

[220] Didier Pittet, Debra Tarara, and Richard P Wenzel. Nosocomial bloodstream infection in critically ill patients: Excess length of stay, extra costs, and attributable mortality. *JAMA*, 271(20):1598–1601, 1994.

[221] Carley B Emerson, Lindsay M Eyzaguirre, Jennifer S Albrecht, Angela C Comer, Anthony D Harris, and Jon P Furuno. Healthcare-associated infection and hospital readmission. *Infection Control & Hospital Epidemiology*, 33(6): 539–544, 2012.

[222] James Moody. The importance of relationship timing for diffusion. *Social Forces*, 81(1):25–56, 2002.

[223] Ugo Fano. Ionization yield of radiations. ii. the fluctuations of the number of ions. *Physical Review*, 72(1):26, 1947.

[224] Siddhartha Chib and Edward Greenberg. Understanding the metropolis-hastings algorithm. *The American Statistician*, 49(4):327–335, 1995.

[225] Axel Kramer, Ingeborg Schwebke, and Günter Kampf. How long do nosocomial pathogens persist on inanimate surfaces? A systematic review. *BMC infectious Diseases*, 6(1):130, 2006.

[226] Thomas Haustein, John P Harris, Richard Pebody, and Ben A Lopman. Hospital admissions due to norovirus in adult and elderly patients. *Clinical infectious Diseases*, 49(12):1890–1892, 2009.

[227] S Harbarth, C Fankhauser, J Schrenzel, J Christenson, P Gervaz, C Bandiera-Clerc, G Renzi, N Vernaz, H Sax, and D Pittet. Universal screening for methicillin-resistant Staphylococcus aureus at hospital admission and nosocomial infection in surgical patients. *JAMA*, 299(1538-3598 (Electronic)): 1149–1157, 2008. ISSN 1538-3598. doi: 10.1001/jama.299.10.1149.

[228] Elisabeth Andritsch, Herbert Stöger, Thomas Bauernhofer, Hans Andritsch, Anne-Katrin Kasparek, Renate Schaberl-Moser, Ferdinand Ploner, and Hellmut Samonigg. The ethics of space, design and color in an oncology ward. *Palliative & Supportive Care*, 11(3):215–221, 2013.

[229] Mara Bloom, Sarah Markovitz, Susan Silverman, and Carl Yost. Ten trends transforming cancer care and their effects on space planning for academic medical centers. *HERD: Health Environments Research & Design Journal*, 8 (2):85–94, 2015.

[230] Paul W Holland and Samuel Leinhardt. Local Structure in Social Networks. *Sociological Methodology*, 7(1):1–45, 1976.

[231] Mark S Granovetter. The strength of weak ties. *American Journal of Sociology*, 78(6):1360–1380, 1973. ISSN 0002-9602. doi: 10.1086/225469.

[232] Dorwin Cartwright and Frank Harary. Structural balance: a generalization of Heider's theory. *Psychological Review*, 63(5):277–293, 1956. ISSN 0033-295X. doi: 10.1037/h0046049.

[233] Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

[234] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather : Homophily in social networks. *Annual Review of Sociology*, 27(2001):415–444, 2014.

[235] Steven M Goodreau, James A Kitts, and Martina Morris. Birds of a feather, or friend of a friend? using exponential random graph models to investigate adolescent social networks. *Demography*, 46(1):103–125, 2009.

[236] Tommy R Jensen and Toft Bjarne. *Graph coloring problems*. John Wiley and Sons, 2011.

[237] Roger V Gould and Roberto M Fernandez. Structures of mediation: A formal approach to brokerage in transaction networks. *Sociological Methodology*, pages 89–126, 1989.

[238] Emma S. Spiro, Ryan M. Acton, and Carter T. Butts. Extended structures of mediation: Re-examining brokerage in dynamic networks. *Social Networks*, 35 (1):130 – 143, 2013. ISSN 0378-8733. doi: http://dx.doi.org/10.1016/j.socnet.2013.02.001.

[239] Christopher Steven Marcum and Laura M. Koehly. Inter-generational contact from a network perspective. *Advances in Life Course Research*, 24:10–20, 2015. ISSN 18796974. doi: 10.1016/j.alcr.2015.04.001.

[240] Christopher Steven Marcum, Jielu Lin, and Laura Koehly. Growing-up and coming-out: Are 4-cycles present in adult hetero/gay hook-ups? *Network Science*, 2016.

[241] Vladimir Batagelj and Andrej Mrvar. A subquadratic triad census algorithm for large sparse networks with small maximum degree. *Social Networks*, 23 (3):237–243, 2001. ISSN 03788733. doi: 10.1016/S0378-8733(01)00035-1.

[242] Iain S Duff, Albert Maurice Erisman, and John Ker Reid. *Direct methods for sparse matrices*. Oxford University Press, 2017.

[243] A. M. Davie and A. J. Stothers. Improved bound for complexity of matrix multiplication. *Proceedings of the Royal Society of Edinburgh: Section a Mathematics*, 143(2):351â€“369, 2013. doi: 10.1017/S0308210511001648.

[244] Raphael Yuster and Uri Zwick. Fast sparse matrix multiplication. *ACM Transactions On Algorithms (TALG)*, 1(1):2–13, 2005.

[245] Peter V Marsden. Interviewer effects in measuring network size using a single name generator. *Social Networks*, 25(1):1–16, 2003.

[246] Katherine Faust. A puzzle concerning triads in social networks: Graph constraints and the triad census. *Social Networks*, 32(3):221–233, 2010.

[247] Mark EJ Newman. Mixing patterns in networks. *Physical Review E*, 67(2): 026126, 2003.

[248] Paul Erdös and Alfréd Rényi. On random graphs I. *Publicationes Mathematicae*, 6:290–297, 1959.

[249] Omer Nebil Yaveroglu, Sean M. Fitzhugh, Maciej Kurant, Athina Markopoulou, Carter T. Butts, and Natasa Przulj. ERGM graphlets: A package for erg modeling based on graphlet statistics. *Journal of Statistical Software*, 65(12):1–29, 2015.