Modeling Transmission Heterogeneity for Infectious Disease Outbreaks

by

Maimuna Shahnaz Majumder
S.M. Engineering Systems, Massachusetts Institute of Technology (2015)
M.P.H. Epidemiology & Biostatistics, Tufts University (2013)
B.S. Engineering Science, Tufts University (2013)

Submitted to the Institute for Data, Systems, and Society
in Partial Fulfillment of the Requirements for the Degree
of
Doctor of Philosophy in Engineering Systems
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
September 2018

© 2018 Massachusetts Institute of Technology. All rights reserved.

Signature of Author

Signature redacted

Institute for Data, Systems, and Society
August 31, 2018

Certified by

Signature redacted

Richard Larson
Professor Post-Tenure of Data, Systems, and Society
Thesis Chair

Certified by

Signature redacted

John Brownstein
Professor, Harvard Medical School
Doctoral Committee Member

Certified by

Signature redacted

Lydia Bourouiba
Assistant Professor of Civil and Environmental Engineering
Doctoral Committee Member

Certified by

Signature redacted

Stan N. Finkelstein
Senior Research Scientist of Data, Systems, and Society
Doctoral Committee Member

Accepted by

Signature redacted

Stephen Graves
Graduate Officer, Institute for Data, Systems, and Society
DISCLAIMER NOTICE

Due to the condition of the original material, there are unavoidable flaws in this reproduction. We have made every effort possible to provide you with the best copy available.

Thank you.

The images contained in this document are of the best quality available.
Modeling Transmission Heterogeneity During Infectious Disease Outbreaks
by
Maimuna Shahnaz Majumder

Submitted to the Institute for Data, Systems, and Society on August 31, 2018 in partial fulfillment of the Requirements for the Degree of Doctor of Philosophy

Abstract

The transmissibility of a given infectious disease is often described by its basic reproduction number ($R_0$) – namely, the average number of secondary infections caused by an index case in a fully susceptible population. Typical approaches to modeling transmission dynamics associated with infectious disease outbreaks frequently use $R_0$ to produce deterministic case count projections, in effect treating the affected population as homogeneous (i.e. as if every individual in the population interest has an equal likelihood of passing on the infection of interest). As a result, such approaches often fail to effectively capture transmission dynamics during real-world outbreaks in heterogeneous populations.

Here, we use analytical and simulation methods to show that the treatment of $R_0$ as the mean of a random variable (thus permitting the estimation of non-deterministic case count projections) allows us to better assess outbreak trajectory and likelihood of disease propagation in non-homogeneous populations (Chapter 2). We then empirically investigate predictors of in-population transmission heterogeneity (i.e. the fact that some individuals in a given population are more likely than others to pass on the infection of interest) within the context of Middle East Respiratory Syndrome in South Korea using a combination of statistical- and review-driven approaches (Chapter 3).

Then, in Chapter 4, we explore how in-population transmission heterogeneity can be used to our advantage through the deployment of risk-informed interventions (i.e. in which individuals who are more likely to pass on the infection of interest are exclusively targeted to receive the intervention) during infectious disease outbreaks. More specifically, we use the analytical and simulation methods first introduced in Chapter 2 – paired with in-population transmission heterogeneity data from Chapter 3 – to compare the utility of a variance-informed deployment scheme against a traditional, uniform deployment scheme (i.e. in which every individual has an equal likelihood of receiving the intervention).

Finally, building off of our findings in Chapters 2, 3, and 4, we recommend four interrelated policies in Chapter 5 that aim to (1) normalize the treatment and reporting of $R_0$ as the mean of a random variable and (2) improve access to the data required to sufficiently capture population heterogeneity when modeling disease propagation.

Thesis Chair: Richard Larson
Title: Professor of Data, Systems, and Society
Acknowledgements

Foremost, I would like to extend my deepest gratitude to my doctoral committee: Dick, for always encouraging me to challenge the status quo and teaching me (by example) that a life in academia is best-suited for those who thrive on curiosity; Stan, for being the first to introduce me to all that MIT had to offer to those interested in the rich intersection that lies between public health and engineering; Lydia, for showing me that the scale at which we ask our questions matter (and that multiple scales are better than just one); and John, for giving me the opportunity (and the financial means) to pursue the line of research I’d always dreamed of pursuing (and for taking a chance on me when no one else would). None of this – truly none – would have been possible without them.

Next, I would like to thank the two communities that have made me feel so at home (and at peace) over the last 4-5 years, throughout my SM and my PhD: the Engineering Systems Division (now, the Institute for Data, Systems, and Society) at MIT and the HealthMap Computational Epidemiology Group at Children’s Hospital and Harvard Med. The friends and mentors I have made thanks to these two groups of inspirational individuals – and the wisdom they’ve imparted upon me – will always stay with me: Abby, Ahmed, Alec, Angel, Ash, Colleen, Denise, Derek, Elaine, Fatima, Henry, Jane, Jared, Joanna, Jon, Kara, Kristin, Larry, Leo, Mauricio, Rachel, Sarah, Sheryl, Shlomiya, Sumi, Yiyao... The list is endless, really.

Beyond these “real-world” communities, I would also like to take a moment to appreciate the ever-illustrious infectious disease Twitterati, without which I would have never met some of my most trusted advisors and colleagues: Adia B., Alexandra P., Caitlin R., Christian A., Ciro C., Crawford K., David F., Eric L., Gregg G., Helen B., Helen P., Ian M., Kelly H., Kim D., Maryn M., Matt W., Michael E., Michael G., Michael O., Michaelleen D., Nick E., Madhu M., Marc L., Nikhil P., Stephen G., Tara S., and so many others. The online community I discovered when I first began to explore the 2014-2016 West African Ebola outbreak – during my first year at MIT, before starting at HealthMap – completely changed the course of my career trajectory, and for that, I am forever indebted.

Penultimately, I would like to acknowledge the folks who have made the last few years as seamless as possible (administratively, technically, and otherwise): Beth, the greatest department administrator ever; Ron, my technical writing tutor; and Leila, Emily and Armine – the spectacular trio that keeps HealthMap up and running.

And finally, to my incredible immediate family – my siblings, Myisha, Adnan, and Jibran; my sister-from-another-mister, Chandni; my parents, Naz and Shafik; and especially to my wonderful husband, Imran... This achievement is as much yours as it is mine. I love you all, more than anything.
Proposal 4: Acknowledging Context-Dependence When Reporting $R_0$ Across Populations

IV. Summary of Recommendations

Appendix I. Other Sources of Non-Traditional Data and Their Uses

References
Chapter 1: Introduction & Framing

In 1766, Daniel Bernoulli attempted – for the first time in recorded history – to use mathematics to glean insight into the transmission dynamics of infectious disease (Dietz and Heesterbeek 2000). A true proponent of inoculation (a position as political then as it is now), Bernoulli used a combination of empirical data and analytical modeling to demonstrate the increases in life expectancy that could be expected given universal variolation against small pox (Blower and Bernoulli 2003).

250 years later, the message of Bernoulli’s work is more pertinent than ever. Recent emerging infectious disease outbreaks – including Ebola, Middle East respiratory syndrome (MERS), and pandemic influenza – highlight the increasingly intimate relationship between infectious disease, mathematics, and policy (Larson and Teytelman 2012; Teytelman and Larson 2012; Althaus 2014; Fisman et al. 2014; Majumder et al. 2015a; Majumder et al. 2016; Majumder et al. 2017).

I. $R_0$, A History

Today, mathematical modeling of infectious disease typically involves the estimation of the basic reproduction number ($R_0$), which is often defined as the average number of secondary infections generated by a primary (i.e. index) case of a given disease in a fully susceptible population (Diekmann et al. 1990; Larson 2007). However, $R_0$ finds its beginnings in population demography and many of the assumptions made during its
conception – and limitations therein – remain intact today in its applications within outbreak research.

The initial concept of “the total reproduction of the population” was first described by German demographer Richard Böckh in 1886 (Heesterbeek 2002). Four decades later (though not in response to Böckh’s original work), Dublin and Lotka formally formed the term “net fertility”, which is equivalent to our modern-day interpretation of $R_0$ within the population growth context (Dublin and Lotka 1925; Heesterbeek 2002).

Prior to this seminal publication however, Lotka had already begun to react to the work of Ross and Hudson, both of whom were working on malaria (Heesterbeek 2002). Ross is perhaps best known for his utilization of epidemic modeling to show that malaria could be controlled via reduction of vectors below a “critical density” or threshold (Ross 1911). Soon thereafter, Ross and Hudson published a series of three papers that serve as the first “developments in abstract epidemic theory”, though this work was developed specifically for malaria and did not themselves generalize further to other infectious diseases (Ross 1916; Ross and Hudson 1917a; Ross and Husdon 1917b; Heesterbeek 2002). This task was instead left to their students Kermack and McKendrick, who published a generalizable threshold theorem for epidemic disease transmission in 1927 – and introduced the now-famous SIR (susceptible-infected-recovered) model to the world (Kermack and McKendrick 1927).
Interestingly, it was not until 1952 that the SIR model – a deterministic\(^1\), ordinary differential equation model in which populations are assumed to be both randomly and homogeneously mixed with respect disease transmission risk – found its first (published) application, by MacDonald within the malaria context, where he believed that quantitative approaches had previously been constrained to "mathematical exercises" by the aforementioned Lotka, Ross, and others, instead of "practical use" (MacDonald 1952). In this paper, MacDonald introduces the phrase "basic reproduction rate" for the first time, though other similar terms had previously been introduced in population demography, as aforementioned (Heesterbeek 2002).

However, at this time, it is worth noting that key assumptions made by Lotka and his fellow demographers in their original realization of \( R_0 \) (or "net fertility") within the population growth context may not be readily applicable within the outbreak research context. Namely, as indicated by Larson (2007), population growth occurs over a span of decades, in a system that may be assumed to be approximately at equilibrium – an assumption that Lotka himself explores in his 1911 paper with Sharpe (Sharpe and Lotka 1911). Though these assumptions are somewhat applicable to endemic diseases (e.g. malaria), epidemic diseases occur over substantially shorter periods of time, in systems that are dynamic and a far cry from any semblance of equilibrium (Larson 2007).

\(^1\)Please note that in this thesis, any reference to SIR- or SIR-type models refers to the traditional, deterministic version of such models which are most frequently used today (unless otherwise specified).
II. Reproduction Numbers Today

Nevertheless, the basic reproduction number – and the SIR-type models that are typically used to estimate it – remains the most widely-accepted descriptor of infectious disease transmissibility today, both within the endemic and (as is the focus of this thesis) the epidemic (or outbreak) context. With this in mind, $R_0$ may be quantified as (Equation 1):

$$R_0 = Cp$$

Where $C$ is the average contact rate per person in the population of interest and $p$ is the average probability of infection given contact with a susceptible individual, the latter of which adequately captures both “susceptibility” and “infectiousness” (Teytelman and Larson 2012; Finkelstein et al. 2015). Deterministic SIR- and SIR-type models assume that both $C$ and $p$ are homogeneous throughout the population of interest. In general (but not always, as is shown in Chapters 2, 3, and 4), diseases with estimated $R_0$ values above 1.0 are expected to result in outbreaks with early exponential growth after introduction into a fully susceptible population, while diseases with lower estimated $R_0$ values are expected to cause linear or sub-linear growth in cumulative cases over time (van den Driessche and Watmough 2002; Larson 2007; Finkelstein et al. 2015) (Figure 1).
Figure 1. Expected growth in cumulative cases over time for various values of $R_0$ in an infinitely large, randomly mixed population, such that susceptibility may always be considered 100%.

Because $R_0$ is generally unobservable, it represents a theoretical quantity that must be mathematically derived (Johansson et al. 2011). $R_0$ can be estimated using a variety of techniques (e.g. deterministic compartmental models) in combination with known attributes of the disease itself and data available from previous or ongoing outbreaks (Dietz 1993; Larson 2007; Johansson et al. 2011; Larson and Teytelman 2012).

Though not the focus of this thesis, other reproduction numbers – which aim to address the dynamic (rather than steady-state) nature of outbreaks – have more recently been used to understand disease dynamics, though they are cited far less frequently in the literature.

The effective reproduction number ($R_{\text{eff}}$) represents the average number of secondary
infections caused per new case at some point in time after an outbreak begins, taking into account control measures that have been put in place (e.g. vaccination) or the diminishing pool of individuals susceptible to infection as a function of time (Farrington and Whitaker 2003; Riley et al. 2003; Kretzschmar et al. 2004; Biggerstaff et al. 2014). It may be calculated simply as the basic reproduction number multiplied by the proportion of the population that is susceptible to the disease, $S$ (Figures 2 and 3) (Equation 2):

$$ R_{Eff} = R_0 S $$

*Figure 2. $R_{Eff}$ and cumulative cases over the course of a hypothetical outbreak in a population of $N = 100$ with $R_0 = 1.5$.断*
Similarly, the observed reproduction number ($R_{obs}$) represents the observed rate of transmission and is typically determined using empirical data (Majumder et al. 2015b; Majumder et al. 2016). It may be estimated at any generation time as follows (Equation 3):

$$R_{obs}(t) = N(t)^{\frac{1}{t}}$$

Where $N(t)$ is the number of incident cases at generation time $t$. $R_{obs}$ is particularly useful when attempting to assess the effect of interventions on transmission dynamics of an outbreak in near-real time, as shown in Figure 4 (Majumder et al. 2015b).

---

The generation time associated with a given infectious disease may be defined as the average time between two consecutive cases in a chain of transmission (Vynnycky and Fine 2000).
Though $R_{0\text{eff}}$ and $R_{\text{obs}}$ make limited advances towards characterizing outbreaks as non-equilibrium states, they too are imperfect; like $R_0$, these values represent population averages. Because of this, estimates of the aforementioned reproduction numbers fail to capture in-population transmission heterogeneity (i.e. the fact that not all individuals that are infected cause the same number of secondary infections) and greatly impact population-level epidemic growth dynamics at aggregate (Larson 2007; Nigmatulina and Larson 2009; Teytelman and Larson 2012; Finkelstein et al. 2015). More explicitly, any given reproduction number is merely the mean of a random variable, which allows for the early extinction or sustained propagation of outbreaks irrespective of how high or low the
mean reproduction number in question is. This limitation is especially pertinent to the basic reproduction number, which is used extensively when modeling transmission dynamics during infectious disease outbreaks.

**III. $R_0$ and the (Faulty) Determination of Outbreak Trajectory**

Policy- and decision-making bodies – including the World Health Organization (WHO) and the Centers for Disease Control and Prevention (CDC) – typically report basic reproduction number estimates as unyielding constants of nature (Finkelstein et al. 2015). These estimates are often used to then establish presumably authoritative case count projections (i.e. outbreak trajectory) for the purposes of resource (i.e. intervention) allocation; however, accuracy – or even consensus across agencies – is not always guaranteed.

For example, in the case of the 2014-2016 West African Ebola outbreak, the WHO and the CDC arrived at wildly different case count projections; while the former predicted 20000 cumulative cases across the region by November 2014, the latter predicted a whopping 550000 cumulative cases in Sierra Leone and Liberia alone by January 2015 – just two months later (CDC 2014; WHO 2014). While the WHO estimate ultimately turned out to be reasonably accurate, the CDC estimate – which was derived from a deterministic SIR-type compartmental model (based off of a previously published SEIR model for Ebola by Chowell et al.) – significantly overshot reality (Chowell et al. 2004; Gaffey and Vibound 2018). Unfortunately, due to the predominance of SIR- and SIR-type models within the context of outbreak research (and the pervasiveness of variance in human populations), this particular kind of failing is all too common.
IV. In-Population Variance and Cross-Population Variability

One of the reasons that the CDC model for Ebola in West Africa failed as catastrophically as it did was due to the fact that it did not take into account in-population variance; both populations of interest were treated as if they were in-and-of-themselves homogeneous (i.e. in the model, all Liberians in any given compartment were treated as homogeneous, and all Sierra Leoneans in any given compartment were treated as homogeneous) (CDC 2014). However, the model did take into account some degree of cross-population variability by treating Liberia and Sierra Leone as two entirely separate contexts (i.e. in the model, Liberians were treated as markedly different from Liberians). Parameters for both models – Liberia and Sierra Leone, respectively – are shown in Table 1 below; notably, though incubation and infectious periods were held as constant across contexts, population size and the number of individuals initially infected were not – thus allowing for the treatment of the contexts as separate.

3 Though the CDC treated Liberia and Sierra Leone as separate populations when modeling expected disease propagation, it is important to note that the initial parameter estimates for the incubation and infectiousness periods (as well as the model itself) were based off of previous outbreaks of Ebola in Uganda and the Democratic Republic of the Congo (CDC 2014). In other words, the CDC used data from prior transmission of Ebola in populations that were vastly different from the West African context when parameterizing their model during the 2014-2016 outbreak. It is likely that this neglect of the local context (which could have been at least partially rectified through measured adjustment of the input parameters given cultural and infrastructural differences between the Central African and West African contexts) also played a contributing role in the model’s failure in West Africa.

4 Interestingly, in the technical appendix of CDC 2014, the CDC acknowledges that the period of infectiousness should perhaps be varied from location-to-location to take into account local customs (e.g. burial practices, etc.); however, they did not appear to make these adjustments prior to reporting case count projections for Liberia and Sierra Leone.
Table 1. Model parameters for Ebola in Liberia and Sierra Leone as reported by the CDC (CDC 2014).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Liberia</th>
<th>Sierra Leone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population Size</td>
<td>4294000</td>
<td>6092000</td>
</tr>
<tr>
<td>Number Initially Infected</td>
<td>9 persons</td>
<td>30 persons</td>
</tr>
<tr>
<td>Incubation Period</td>
<td>6.3 days</td>
<td>6.3 days</td>
</tr>
<tr>
<td>Infectious Period</td>
<td>6 days</td>
<td>6 days</td>
</tr>
</tbody>
</table>

This distinction – between in-population variance and cross-population variability – is critically important, and though they are inter-related, drivers of each of these two types of transmission heterogeneity operate at different scales. In-population variance (i.e. the fact that individuals are different from each other with respect to infectious disease transmission) is primarily motivated by individual-level differences in biology (e.g. sex, age, etc.) and behavior (e.g. frequency of hand-washing; choice of career; etc.). Meanwhile, cross-population variability (i.e. the fact that populations are different from each other with respect to infectious disease transmission) is dictated by differences in demographics (e.g. sex breakdown, age pyramid, etc.), as well as differences in culture (e.g. social prevalence of hugging and kissing) and infrastructure (e.g. public transportation availability; accessibility of healthcare services; etc.). Thus, in some ways, cross-population variability may be considered as a natural extension of in-population variance.

5Drivers of in-population variance may be generalizable across populations (even in the event of cross-population variability) if sufficiently robust physical science (and thus, essentially universal) mechanisms exist (by which the driver of interest propagates in-population variance) and if sufficient caution is used in interpretation across populations. For instance, if older age in Population A is found to be associated with higher likelihood of spread for Disease X (via Mechanism X), it is reasonable to expect (until proven otherwise) that older age will also be associated with higher likelihood of spread for Disease X in other populations.
Unfortunately, the CDC and the WHO frequently publish reproduction numbers that entirely ignore the importance of context; instead, both organizations opt to report general (as opposed to population-specific) ranges for many common diseases, thus effectively discounting cross-population variability (Table 2) (CDC and WHO 2010). However, it is worth noting that – during the 2014-2016 West African Ebola outbreak – the WHO and the CDC did in fact treat each affected country (i.e. Guinea, Liberia, and Sierra Leone) as distinct (and estimated different reproduction numbers for each country), as did other highly-cited modeling studies, which indicates that a culture shift may slowly be taking shape in the greater outbreak response community (Althaus 2014; CDC 2014; Fisman et al. 2014; WHO 2014).

Table 2. $R_0$ estimates as reported by the WHO and CDC for six well-known diseases.

<table>
<thead>
<tr>
<th>Disease</th>
<th>Reported $R_0$ Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diphtheria</td>
<td>6-7</td>
</tr>
<tr>
<td>Measles</td>
<td>12-18</td>
</tr>
<tr>
<td>Mumps</td>
<td>4-7</td>
</tr>
<tr>
<td>Polio</td>
<td>5-7</td>
</tr>
<tr>
<td>Smallpox</td>
<td>5-7</td>
</tr>
<tr>
<td>Rubella</td>
<td>5-7</td>
</tr>
</tbody>
</table>

Population B in the event that confounders that may impact the relationship between age and transmissibility are not significantly differential across populations. However, assuming that the age pyramid in Population B is different from the age pyramid in Population A, larger scale dynamics (such as the total number of expected cases of Disease X in Population A versus Population B) would be different, as well. For instance, if Population A and B are the same size, but Population B features fewer older individuals than Population A, then the expected number of total/Disease X cases would be less in Population B than in Population (even if older individuals act as the primary motivators of disease transmission in either population). This relationship between in-population variance and cross-population variability is discussed further in Chapter 4.
Nevertheless, treatment of human populations as *intrinsically* homogeneous (i.e. without consideration for in-population variance) remains common due to the ubiquity of the deterministic SIR model. Because of this, we concentrate primarily on the role of in-population variance on infectious disease outbreaks in this thesis—though we *do* highlight the relationship between in-population variance and cross-population variability in Chapters 4 and 5.

V. Modeling Infectious Disease Outbreaks

Though deterministic SIR- and SIR-type methods remain the most common approach to modeling infectious disease outbreaks, a number of stochastic methods may also be used for the purposes of *non-deterministic* case count projection (in which multiple possible “futures”—and the probability of their occurrence—are estimated). The three most common categories of such methods are as follows (Britton et al. 2015; Pellis et al. 2015; Riley and Trapman 2015):

- **Global:** Similar to the deterministic SIR-type models described above, global models assume random mixing (Britton et al. 2015). Though splitting the population into demographic subgroups (each with a different probability of infection given contact, $\rho$) can inject some heterogeneity into this kind of model, all infected individuals in a given subgroup have identical rates of disease-causing contacts, $C$.

- **Metapopulation:** Metapopulation models allow for individuals to have heterogeneous contact rates ($C$) based on sub-populations (e.g. households) within...
the model; namely, individuals in the same household may have one contact rate with each other, and another (lower) contact rate with other households (Britton et al. 2015). Probability of infection given contact \((p)\) may also be permitted to vary by sub-population; however, like global models, homogeneous mixing is assumed at the sub-population level (Ajelli et al. 2010).

- **Network:** As sub-populations in metapopulation models approach \(N = 1\) individuals, they essentially become network models (Riley and Trapman 2015). Typically, nodes in network models represent individuals and edges represent contacts, thus allowing in-population (i.e. individual-level) heterogeneity to be incorporated into the model for both \(C\) and \(p\) or some combination there-in (i.e. number of secondary transmissions caused by a given infectious agent) (Pellis et al. 2015). Such models, however, are plagued by the reality that accurate network information for real populations is hard to find (Keeling and Eames 2005). Because of this, synthetic networks are frequently used instead, which may or (more likely) may not represent real-life networks of interest. Nevertheless, creative explorations of synthetic networks can lend insights into the transmission heterogeneity we may see across individuals and populations in real life. Common model types that fall roughly under the purview of network methods – defined broadly by Britton et al. (2015) as “any individual-based epidemic model” – include (but are not at all limited to) analytical, Monte Carlo, and agent-based models:

  - Analytical methods often incorporate individual-level (i.e. in-population) variance through the derivation of either empirical or synthetic probability distributions (e.g. probability mass functions) for \(C\) and \(p\) or some
Chapter 1: Introduction & Framing

combination there-in, which may then be used to compute non-deterministic case count projections (Finkelstein et al. 2015; Teytelman and Larson 2013). Such non-deterministic case count projections include not only the probability associated with any given number of cumulative cases at time $t$ but also the probability of outbreak takeoff and the probability of early extinction. In many ways, analytical methods are the pre-cursor to more traditional network models, which use simulation to track the impact of individual-level (i.e. in-population) variance on outbreak trajectory, described below.

- Monte Carlo methods typically allow for the simulation of multiple outbreaks for any given disease, thus yielding multiple possible “futures” (i.e. non-deterministic case projections) for the disease of interest that are specific to the population of interest (Nigmatulina and Larson 2009; Teytelman and Larson 2013). Such methods include individual-level (i.e. in-population) variance by drawing from analytically-derived (synthetic or empirical) probability distributions for $C$ and $p$ or some combination there-in (i.e. number of secondary transmissions caused).

- Agent-based methods usually include individual-level (i.e. in-population) variance by specifying a wide range of parameters for each and every individual in the population of interest (Rahmandad and Sterman 2008; Crooks and Hailegiorgis 2014). Outbreaks are then simulated using Monte Carlo methods, through which randomness is often incorporated in the event that individual-level parameters are drawn from distributions (as opposed to
assigned explicitly as point estimates). Realistic parameterization of this particular kind of network model can be challenging due to the number of parameters that are typically employed and often requires highly-detailed individual-level data (e.g. individual census records) (Ajelli et al. 2010).

Unsurprisingly, deterministic models like SIR- and SIR-type models are intended to approximate the stochastic models described above; however, the validity of this approximation only holds if the population under consideration is sufficiently large (Nasell 2002). As a result, stochastic models – most of which are computationally more intensive – have been increasingly employed over the last 60 years, alongside advances in computing (Nasell 2002; Britton et al. 2015).

VI. Thesis Roadmap

In this thesis, we begin by comparing the performance of deterministic and stochastic methods when modeling infectious disease transmission in populations that exhibit in-population variance. Specifically, we start Chapter 2 with an explicit exploration of the limitations associated with SIR- and SIR-type models when modeling disease propagation in non-homogeneous populations, with a focus on how such models produce strictly deterministic case count projections. We then use both analytical and Monte Carlo simulation (heretofore referred to as “simulation”) methods to show that $R_0$ should be treated as the mean of a random variable – thus allowing for non-deterministic case count projections – when characterizing transmission dynamics of infectious disease in heterogeneous populations.
With this finding in tow, we move forward to Chapter 3, where we empirically investigate a real-world example of infectious disease transmission in a heterogeneous population. More specifically, we determine predictors of heterogeneous disease transmission within the context of MERS in South Korea (i.e. what types of individuals were more likely to pass on MERS and the why) through a combination of data- and review-driven approaches.

Then, in Chapter 4, we show how the underlying variance inherent to human populations (as exemplified during the 2015 South Korean MERS outbreak) can be used to better inform intervention deployment in low-resource settings. We pay special attention to the utility of variance-informed intervention deployment (in which those at greater risk of passing on the disease of interest are targeted exclusively) and compare the effectiveness of such a scheme against traditional (i.e. uniform) deployment methods (in which every individual in the effected population has an equal likelihood of being targeted).

Finally, in Chapter 5, we summarize our primary findings from Chapters 2, 3, and 4 and identify the real-world implications of such findings. We then recommend four inter-related policies, which aim to (1) normalize the treatment of $R_0$ as the mean of a random variable and (2) improve access to the data required to sufficiently capture population heterogeneity when modeling disease propagation in non-homogeneous populations.
Chapter 1: Introduction & Framing

References


Chapter 2: SIR Models and the Problem of Population Heterogeneity

I. Introduction

During an outbreak of an emerging infectious disease, mathematical modeling is typically used to estimate the number of cases expected at various points in time (i.e. generation time, $t$) for the purposes of informing on-the-ground decision-making and resource deployment. The most common approach to such modeling efforts involves the use of deterministic compartmental methods (in which individuals move over time from one compartment to another), and of these, the SIR (susceptible–infectious–recovered) model and its extensions are often considered the “gold standard” (Hethcote 1994).

Deterministic SIR-type models are typically parameterized using real-world data from an ongoing outbreak or from historical estimates of the basic reproduction number ($R_0$) associated with a given disease. A number of unrealistic assumptions are then employed in order to estimate an outbreak trajectory associated with the disease of interest; these assumptions may be summarized as follows (Teytelman and Larson 2012; Finkelstein et al. 2015):

1. The population of interest is assumed to be infinitely large and infection is assumed to occur via the Law of Mass Action, such that those belonging in the “infected” and the “susceptible” compartment are randomly mixed and the total number of “recovered” individuals is considered negligible early in an outbreak. This is rarely an accurate representation of reality, given that human populations are typically
Chapter 2: SIR Models and the Problem of Population Heterogeneity

non-randomly mixed (i.e. humans preferentially make contact with small subgroups – including family, friends, and colleagues – rather than at an equal rate with all members of an infinitely large population).

2. All individuals in the population are assumed to have an equal probability of contracting the disease at a rate specified by the transmission parameter $\beta$, which is a composite measure that considers both contact rates and probability of transmission; furthermore, the probability of recovering or dying at rate $\gamma$ is also assumed equal across the population. Given that $R_0$ – the average number of secondary cases an index case is expected to cause in a fully susceptible population – is equivalent to $\beta / \gamma$ (see Section II), the probability of transmitting the disease (given infection) is thus considered to be equal across the population as well. In other words, SIR models assume that populations exhibit zero variance with respect to outbreak dynamics; however, in reality, transmission of and susceptibility to a given disease are hardly ever homogeneous due to important contextual factors (e.g. heterogeneity of host immune response, individual connectivity, and hygiene practices, amongst others) (Kiss et al. 2006; Poland et al. 2007; Cantrell et al. 2009). This may be easily evidenced by the existence of super-spreaders for a number of emerging pathogens (e.g. MERS, see Chapter 3).

Given that the objective of this thesis centers on the role of in-population transmission heterogeneity during emerging infectious disease outbreaks, Chapter 2 will focus primarily on the limitations of classical definitions of $R_0$ and compartmental SIR-type models when
estimating case count projections and expected disease propagation in non-homogenous populations. As highlighted by Nigmatulina and Larson (2009), $R_0$ – which is, by default, an average quantity – “ignored important heterogeneity” in human populations; similarly, Teytelman and Larson (2012) note that, because deterministic compartmental models assume that “all people in a single compartment behave identically [with] random mixing”, they neglect in-population variance. After introducing these concepts briefly in the literature review component of each study, both of these seminal papers go on to show how incorporation of in-population heterogeneity – via analytical and simulation methods – can better represent reality during outbreaks of influenza. Here, we build off of this prior work by explicitly comparing the performance of deterministic SIR models against analytical and simulation methods within the context of a pre-defined synthetic disease.

To establish a baseline, we start with the SIR model itself and use it to estimate an outbreak trajectory for a synthetic disease with a pre-determined basic reproduction number (Section II); we then use analytical methods (Section III) and Monte Carlo simulation (Section IV) to model the same synthetic disease (i.e. using the same pre-determined basic reproduction number). Unlike SIR-type models, analytical and simulation approaches allow for the exploration of disease propagation in non-zero variance settings; thus, in Section V, we compare results across all three approaches and highlight the insufficiencies of SIR models when modeling infectious diseases when transmission is heterogeneous.
II. $R_0$ and The SIR Model

In this section, we review the relationship between the SIR model and the classical interpretation of the basic reproduction number (which, when taken together, yield deterministic case count projections). We then estimate the trajectory of a synthetic disease with a pre-determined basic reproduction number similar to that of recent emerging pathogens, including Ebola (i.e. $R_0 = 2$) (WHO 2014).

A. Defining $R_0$

The basic reproduction number ($R_0$) for diseases transmitted via human-to-human infection is often defined as the average number of secondary cases that an index case is expected to cause in a fully susceptible population. As it pertains to the SIR model, $R_0$ may be described as a product of three variables: $C$ (the average rate of contact between those who are susceptible and those who are infected), $p$ (the average probability of infection given contact between an individual that is infected and an individual that is susceptible), and $d$ (the average duration of an infected individual's infectiousness). Because possible heterogeneity across the population of interest with respect to $C$, $p$, and $d$ is not considered, $R_0$ is treated as a constant under the SIR modeling paradigm, and as a result, $R_0$ must be $>1$ at $t = 0$ for an outbreak to occur. This definition of $R_0$ may be derived directly from the SIR model itself, as is shown in Section II Part B below.

B. Deriving $R_0$ from the SIR Model

The deterministic SIR model may be defined by a series of ordinary differential equations, namely (Equations 1–3):

---

6When $R_0$ is treated as the mean of a random variable, this condition is no longer necessary (see Appendix I).
Chapter 2: SIR Models and the Problem of Population Heterogeneity

\[
\frac{dS}{dt} = -\beta \frac{SI}{N} \quad (1)
\]

\[
\frac{dI}{dt} = \beta \frac{SI}{N} - \gamma I \quad (2)
\]

\[
\frac{dR}{dt} = \gamma I \quad (3)
\]

Where \( N \), the total population, is equivalent to \( S + I + R \); \( \beta \) is the rate at which susceptible \((S)\) individuals become infected \((I)\) (per generation \( t \)); and \( \gamma \) is the rate at which infected individuals either die or recover \((R)\) (per generation \( t \)). Note that \( \beta \) is essentially the product of \( C \) and \( p \), and \( \gamma \) is essentially the inverse of \( d \).

Equations 1–3 can then be re-written to represent the fraction of individuals in each compartment (Equations 4–6):

\[
\frac{dS}{dt} = -\beta si \quad (4)
\]

\[
\frac{dI}{dt} = \beta si - \gamma i \quad (5)
\]

\[
\frac{dR}{dt} = \gamma I \quad (6)
\]

To derive \( R_0 \) from this series of equations, we first consider the fact that in order for an outbreak to begin, the rate of infection must be > 0; thus, Equation 5 may be re-written as follows (Equations 7 and 8):

\[
\beta si - \gamma i > 0 \quad (7)
\]
At the beginning of an outbreak, nearly everyone – except for the index case – is susceptible, such that $S \approx 1$. Solving Equations 7 and 8 such that $S = 1$, we arrive at the following expression (Equation 9):

$$\frac{\beta s i}{\gamma} > i$$

(Note that, when $t =$ generation time, then $\gamma = 1$ and $\beta = R_0 = Cp$.)

Because $\beta$ and $\gamma$ are equal across the population (i.e. each individual is prescribed the same $\beta$ and $\gamma$ values), $R_0$ is a constant under the SIR modeling paradigm. As a result, deterministic SIR-type models – by their very definition – only allow for the estimation of one possible “future” (i.e. case projection), as is shown in Section II Part C below.

C. Estimating Cumulative Incidence Using $R_0$

Early in an outbreak (and if the aforementioned assumptions associated with an SIR-type
model hold), the number of new cases \( I \) and the number of cumulative cases \( N \) expected at any given generation time \( t \) may be estimated as follows (Equation 10 and 11) (Fisman et al. 2013):

\[
I(t) = R_0^t
\]

\[
N = \sum_{t=0}^{\infty} R_0^t
\]

Thus, for our synthetic disease of interest (with \( R_0 = 2 \)), cumulative cases over time – as determined by an SIR-type model – is merely an exponential growth function (Figure 1).

---

7When the assumption of an infinitely large population is relaxed, more complex approaches to case projection may be used to take into account depletion of the susceptible compartment and the effects of said depletion on outbreak growth.
III. $R_0$ as the Mean of a Random Variable: An Analytical Approach

By their very definition, SIR-type models assume that the population affected is completely homogenous; as a result, the deterministic case projections ascertained from such models are limited in their insight when considering non-homogeneous, real-world populations where $R_0$ is the mean of a random variable (i.e. with non-zero variance and for which non-deterministic case count projections may be estimated) (Figure 1).

Human populations are heterogeneous, and the individuals that comprise said populations differ both in their behavior and in their biology – both of which impact risk of disease transmission (Kiss et al. 2006; Poland et al. 2007; Cantrell et al. 2009). Due to this heterogeneity, real-world outbreaks of diseases with $R_0 > 1$ frequently do not exhibit a period of early exponential growth in the real-world, despite what SIR-type models suggest; in fact, given sufficient variance, an outbreak may not take off at all (despite the presence of an index case) or early extinction may be just as likely (see Section III Part A).

Because deterministic SIR-type models do not allow for the treatment of $R_0$ as the mean of a random variable, other approaches must be utilized to evaluate the probability of outbreak takeoff or extinction. Thus, in this section, we use analytical methods to explore the potential ramifications of non-zero variance given a synthetic disease in which $R_0 = 2$ (i.e. identical to the one described in Section II).

In Section III Part A, we demonstrate that – through the inclusion of non-zero variance – an unlimited number of simple probability mass functions (PMFs) may accurately describe a synthetic disease with a mean of $R_0 = 2$ within the context of an infinitely large, randomly mixed population. Then, using a subset of these PMFs, we show in Section III Part B that each PMF exhibits its own unique probability of outbreak takeoff (by $t = 1$) and probability of early outbreak extinction (by $t = 2$). This subset of PMFs is further examined in Section IV, in which Monte Carlo simulation is employed to estimate minimum, maximum, and average incidence over time, as well as to produce a likelihood histogram associated with
total outbreak size at \( t = 5 \). A zero-variance example is also presented in both Section III and Section IV for direct comparison against the SIR-type model from Section II.

### A. Non-Zero Variance and the Probability of Outbreak Takeoff

An infinitely large number of simple probability mass functions (PMFs) may be used to define a synthetic disease with a mean of \( R_0 = 2 \). Because SIR-type models assume that \( R_0 \) is homogeneous across any given population of interest, we first start with a PMF in which there is zero variance (PMF A):

\[
P(N = n) = \begin{cases} 
1 & n = 2 \\
0 & n = 0 
\end{cases} \quad (PMF \ A)
\]

\[
R_0 = E[N] = 2(1) + 0(0) = 2
\]

Under an SIR-type (i.e. zero variance) paradigm, the probability of outbreak takeoff is 1. However, if we allow for the inclusion of variance, we can produce an infinitely large number of PMFs – all with a mean of \( R_0 = 2 \) – in which the probability of outbreak takeoff is <1. For example, if we split the mass centered at \( n = 2 \), we produce a PMF with half its mass centered at \( n = 4 \) and half its mass centered at \( n = 0 \) (PMF B):

\[
P(N = n) = \begin{cases} 
0.5 & n = 4 \\
0.5 & n = 0 
\end{cases} \quad (PMF \ B)
\]

\[
R_0 = E[N] = 4(0.5) + 0(0.5) = 2
\]

Like PMF A, PMF B also has a mean of \( R_0 = 2 \); however, within the context of PMF B, there is only a 50% chance of outbreak takeoff by \( t = 1 \) following the introduction of an index case at \( t = 0 \). From here, we can produce yet another PMF with a mean of \( R_0 = 2 \), in which a quarter of its mass is centered at \( n = 8 \) and three-quarters of its mass is centered at \( n = 0 \) (PMF C):

\[
P(N = n) = \begin{cases} 
0.25 & n = 8 \\
0.75 & n = 0 
\end{cases} \quad (PMF \ C)
\]

\[
R_0 = E[N] = 8(0.25) + 0(0.75) = 2
\]
Chapter 2: SIR Models and the Problem of Population Heterogeneity

\[ P(N = n) = \begin{cases} 
0.25 & n = 8 \\
0.75 & n = 0 
\end{cases} \quad (PMF \ C) \]

\[ R_0 = E[N] = 8(0.25) + 0(0.75) = 2 \]

Under the paradigm of PMF C, the probability of takeoff is only .25, even though \( R_0 = 2 \).

This same process may be continued ad infinitum, producing a generalizable PMF with the following form (PMF Gen):

\[ P(N = n) = \begin{cases} 
\frac{R_0}{x} & n = x \\
1 - \frac{R_0}{x} & n = 0 
\end{cases} \quad (PMF \ Gen) \]

\[ R_0 = E[N] = x \left(\frac{R_0}{x}\right) + 0(1 - \frac{R_0}{x}) = R_0 \]

Tabulation of PMFs A, B, C, etc. clearly demonstrates that as the underlying variance increases, the probability of outbreak takeoff decreases (Table 1). In other words, just because the \( R_0 \) associated with a given disease is \( >1 \) does not mean that an index case will definitively result in an outbreak (as SIR models suggest) in real-world, non-homogenous populations. Furthermore, in the event that an outbreak does takeoff, exponential growth is not necessarily guaranteed, either; in fact, as is shown in Section III Part B, the probability of early extinction is related to variance, too.
Table 1. Tabulation of $\sum np = R_0 = 2$ for various values of $n$ and $p$. Derivation of PMFs D, E, and F may be found in Appendix II. Note that as $n$ increases, variance increases.

<table>
<thead>
<tr>
<th>PMF</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>Gen</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R_0$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.5</td>
<td>0.25</td>
<td>0.125</td>
<td>0.0625</td>
<td>0.03125</td>
<td>$R_0/x$</td>
</tr>
<tr>
<td>2</td>
<td>0.75</td>
<td>0.5</td>
<td>0.25</td>
<td>0.125</td>
<td>0.0625</td>
<td>0.03125</td>
<td>$R_0/x$</td>
</tr>
<tr>
<td>2</td>
<td>0.875</td>
<td>0.5</td>
<td>0.25</td>
<td>0.125</td>
<td>0.0625</td>
<td>0.03125</td>
<td>$R_0/x$</td>
</tr>
<tr>
<td>2</td>
<td>0.9375</td>
<td>0.5</td>
<td>0.25</td>
<td>0.125</td>
<td>0.0625</td>
<td>0.03125</td>
<td>$R_0/x$</td>
</tr>
<tr>
<td>2</td>
<td>0.96875</td>
<td>0.5</td>
<td>0.25</td>
<td>0.125</td>
<td>0.0625</td>
<td>0.03125</td>
<td>$R_0/x$</td>
</tr>
</tbody>
</table>

B. Non-Zero Variance and the Probability of Early Extinction

The probability of early extinction (i.e. by $t = 2$) may be calculated analytically for any of the aforementioned PMFs by considering the probability associated with $n = 0$, which is constant across all generations times $t$. For example, for PMF C (where the probability associated with $n = 0$ is 0.75), the probability of extinction by $t = 2$ may be determined as follows (Equation 12):

$$P(\text{extinction by second generation}) = 0.75 + (1 - 0.75) \times 0.75^8 \approx 0.78 \quad (12)$$

As an extension of PMF Gen, Equation 12 may further be generalized for any value $n = x$ and any value $R_0$ via (Equation 13):

$$P(\text{extinction by second generation}) = \left(1 - \frac{R_0}{x}\right) + \left(\frac{R_0}{x}\right) \left(1 - \frac{R_0}{x}\right)^x \quad (13)$$
The latter half of the summation in Equation 13 represents the change in total probability of extinction between \( t = 1 \) and \( t = 2 \); tabulation of solutions for PMFs A–F demonstrate that this change (as a fraction of \( R_0/x \)) is directly related to variance (Table 2). In other words, just because an outbreak of a disease with \( R_0 > 1 \) takes off does not mean that its early growth will be exponential; in fact, when \( R_0 \) is the mean of a random variable, the likelihood of early extinction increases as the underlying variance increases.

Table 2. Probability of extinction (\( P_e \)) by \( t = 1 \) and \( t = 2 \) for PMFs A–F.

<table>
<thead>
<tr>
<th>( R_0 )</th>
<th>( n )</th>
<th>PMF</th>
<th>( P_e(1) )</th>
<th>( P_e(2) )</th>
<th>( [P_e(2)-P_e(1)]/(R_0/x) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2</td>
<td>A</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>B</td>
<td>0.5</td>
<td>0.53125</td>
<td>0.0625</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>C</td>
<td>0.75</td>
<td>0.77503</td>
<td>0.10011</td>
</tr>
<tr>
<td>2</td>
<td>16</td>
<td>D</td>
<td>0.875</td>
<td>0.88976</td>
<td>0.11807</td>
</tr>
<tr>
<td>2</td>
<td>32</td>
<td>E</td>
<td>0.9375</td>
<td>0.94542</td>
<td>0.12789</td>
</tr>
<tr>
<td>2</td>
<td>64</td>
<td>F</td>
<td>0.96875</td>
<td>0.97285</td>
<td>0.13108</td>
</tr>
</tbody>
</table>

\( R_0 \) is the mean of a random variable, \( x \) is the underlying variance. See Equation 13 for \( 1 - R_0/x \).
IV. $R_0$ as the Mean of a Random Variable: A Simulation Approach

When $R_0$ is treated as the mean of a random variable, there are palpable effects on probability of takeoff by $t = 1$ and early extinction by $t = 2$ (as is shown in Section III). Unlike the zero-variance case (i.e. SIR-type paradigm, PMF A), early exponential growth is not guaranteed simply because $R_0$ is >1. In this section, we build off of these initial findings by using Monte Carlo simulation to evaluate expected outbreak trajectory for a synthetic disease with $R_0 = 2$ (i.e. identical to the one described in Sections II and III) over the course of five generations for PMFs A–F.

In Section IV Part A, we describe the methods used to parameterize our simulation models for PMFs A–F (i.e. SIMs A–F); then, in Section IV Parts B and C, we report a number of findings (as ascertained from our simulations):

- In Part B, we report minimum, maximum, and average cumulative incidence over time to assess the effect of non-zero variance on the spread of possible “futures” (i.e. case projections) for SIMs A–F; and
- In Part C, we report the likelihood of extinction for $t = [1, 5]$, as well as the histogram associated with total outbreak size at $t = 5$ for SIMs A–F.

Comparison of findings across Sections II, III, and IV – as well as methodological limitations associated with each approach – are discussed in Section V.

A. Monte Carlo Simulation of PMFs

For direct comparison against the SIR model presented in Section II and the analytical approach presented in Section III, we used Visual Basic to create a simple Monte Carlo simulation model of an infinitely large, randomly mixed population. To reflect the 6 PMFs
of interest (i.e. A-F), 6 different versions of the model (i.e. SIM A–F) were parametrized – each with a different fraction of the population assigned to \( n = 0 \) and \( n = x \). For example, SIM C was parameterized such that 25% of the population would go on to infect 8 other individuals (i.e. \( n = x = 8 \)) if infected (or selected as an index case), while the remaining 75% of the population would go on to infect no one (i.e. \( n = 0 \)) if infected (or selected as an index case).

Index cases were selected at random at the start of each outbreak simulation, and a total of 5000 trial runs were simulated for each of the 6 SIMs (i.e. A–F). Each outbreak simulation was permitted to propagate for 5 generations\(^8\) before termination of the simulation sequence. Cases were modeled with an infectious period of 1 generation, and given the infinitely large and randomly mixed nature of the modeled population, re-infections were not permitted.

---

\(^8\)Trial runs were capped at 5 generations for direct comparison against the SIR model presented in Section II; however, the simulation sequence can be terminated beyond \( t = 5 \) (if needed).
B. Cumulative Incidence Over Time for SIMs

After running 5000 trials for each of the SIMs, we plotted average, maximum, and minimum cumulative incidence over time for SIMs B-F (Figure 2). The minimum cumulative incidence for all 5 SIMs was 0, and the average was nearly identical to that estimated by the zero variance (i.e. SIR-type) PMF B, max variance (i.e. SIR-type) PMF G max case (Figure 1). The SIR-type case (i.e. SIM A) was excluded from the plot due to absence of variance (i.e. average = maximum = minimum). Notably, the difference away from the average cumulative incidence curve increased with increases in underlying variance.

C. Extinction Over Time and Total Outbreak Size for SIMs

After plotting cumulative incidence over time, we determined the likelihood of extinction for each of the 6 SIMs over time (Table 3). Results were fairly consistent with those produced via the analytical approach (Table 2).
Table 3. Likelihood of extinction at \( t = \{1, 5\} \) for 6 different simulation models (i.e. SIMs A-F) as determined over the course of 5000 trial runs. Note that the likelihood of takeoff is equivalent to \( 1 - \) (likelihood of extinction at \( t = 1 \)).

| SIM = & A & B & C & D & E & F |
|-------|-------|-------|-------|-------|-------|-------|
| \( t \) & \[ \text{Likelihood of extinction by generation time, } t \] |
| 1     & 0    & 0.5082 & 0.7388 & 0.879 & 0.9392 & 0.9678 |
| 2     & 0    & 0.5388 & 0.7652 & 0.8892 & 0.9482 & 0.9732 |
| 3     & 0    & 0.549  & 0.7728 & 0.8932 & 0.951  & 0.9748 |
| 4     & 0    & 0.551  & 0.7764 & 0.8938 & 0.9516 & 0.9756 |
| 5     & 0    & 0.5514 & 0.7766 & 0.894  & 0.9518 & 0.9756 |

We then plotted a likelihood histogram for each of the SIMs to compare the spread associated with total outbreak size at \( t = 5 \) across SIMs (Figure 3). SIMs with higher degrees of underlying variance exhibited greater spread, as would be expected from Figure 2. Because of the high likelihood of extinction across SIMs B-F, 0 was excluded from the \( x \)-axis. Nevertheless, all histograms – with the exception of the zero-variance case (i.e. SIM A) – were right-tailed.
Chapter 2: SIR Models and the Problem of Population Heterogeneity

Figure 3. Likelihood histograms for outbreak size at t = 5 (SIMs A-F). Histograms show data from 5000 trial runs each. The index case was excluded from all counts.
V. Discussion

In this final section of Chapter 2, we compare findings across Section II, III and IV; review the methodological limitations associated with each approach (i.e. SIR, analytical, and simulation); and discuss real-world implications of the work presented in this chapter.

A. Comparison of Findings

In Section II, we used a simple SIR model to explore the classical interpretation of $R_0$ and to deterministically estimate the number of cumulative cases that would be expected under such a paradigm given a synthetic disease with $R_0 = 2$ (Figure 1). Then, in Section III and IV, we used analytical and simulation approaches to treat $R_0$ as the mean of a random variable and to explore the impact of said treatment on disease propagation of the same synthetic disease (i.e. with $R_0 = 2$). Through the inclusion of a zero variance example in both Section III (i.e. PMF A) and Section IV (i.e. SIM A), we were able to fully replicate our findings from the SIR model presented in Section II – namely early exponential growth following a 100% chance of outbreak takeoff given the introduction of an index case and a 0% chance of early extinction; we were then able to compare this zero variance case against the non-zero variance cases explored in Sections III (i.e. PMF B-F) and Section IV (i.e. SIMs B-F), as well.

In Section III, we found that – as the underlying variance increases away from 0 – the probability of outbreak takeoff decreases and the probability of early extinction increases (Tables 1 and 2). Both of these findings were further verified by the simulation models that
Chapter 2: SIR Models and the Problem of Population Heterogeneity

we implemented in Section IV.

Through our simulation models, we also discovered that – irrespective of the underlying variance – the average outbreak trajectory of a given disease with $R_0 = 2$ (over the course of 5000 simulated outbreaks) was roughly equivalent to the exponential growth function presented in Section II (Figures 1 and 2). However, the difference between the average outbreak trajectory and the maximum simulated outbreak trajectory was largest for the simulation model with the highest degree of underlying variance (i.e. SIM F). As a result, the horizontal spread of the likelihood histograms associated with final outbreak size at $t = 5$ varied substantially across simulation models (i.e. SIM F exhibited the greatest spread), even though the average final outbreak size was roughly equivalent across all SIMs.

B. Methodological Limitations

Given our results from Sections III and IV, it is evident that – in non-homogeneous populations that exhibit variance with respect to disease transmissibility – SIR models can provide little insight beyond the mean. In such instances, analytical and simulation approaches should be used preferentially in order to ascertain the probability of outbreak takeoff given the introduction of an index case and the probability of early extinction given outbreak takeoff, as well as the many possible “futures” (i.e. case projections) that might be expected in the event of continued transmission. This said, such approaches are not without limitations – especially in the form in which they are used here.
Like the SIR model presented in Section II, the analytical and simulation approaches presented in Sections III and IV assume that the populations affected by the synthetic disease of interest are infinitely large and randomly mixed. Given that human populations are clearly non-infinite and given that human beings tend to interact non-randomly (i.e. preferential interaction with family, friends, and colleagues), neither of these assumptions are realistic. With this in mind, it is worth noting that the simulation models developed in Section IV can be easily adjusted to represent a more realistic human population (i.e. non-infinite size with non-random mixing). However, because we wanted to compare results from our non-zero variance simulations (i.e. SIMs B-F) directly against those obtained from the SIR model presented in Section III and the analytical approach presented in Section IV, we chose to allow these assumptions (i.e. infinitely large population with random mixing) to hold.

Results obtained from Monte Carlo simulations are, however, beholden to the number of trials that are run. In Section IV, we opted to run 5000 trials for each of the 6 simulation models; because of this, low probability events (i.e. \( p < 1/5000 \)) likely were not adequately captured (e.g. a simulated outbreak via SIM F in which each and every individual infected between \( t = 0 \) and \( t = 5 \) infected 64 other individuals, resulting in a final outbreak size of approximately 1.1 billion). When necessary, analytical approaches can be used to assess the likelihood of such low probability events (Appendix III). However, for more complex PMFs such techniques may not be analytically tractable, and in such instances, simulation featuring an increased number of trials may be considered instead.
Thus, though analytical and simulation approaches pose their own challenges, they both allow for the treatment of $R_0$ as the mean of a random variable and should be used in complement when investigating outbreaks in non-homogenous populations.

C. Real-World Implications

Given that human populations are non-homogeneous, our findings suggest that SIR models frequently fail to properly estimate disease propagation during real-world outbreaks. Such a failure was recently evidenced during the 2014-2016 West African Ebola outbreak (CDC 2014). With this in mind, analytical and simulation methods may be a better alternative when attempting to estimate outbreak trajectory for the purposes of crisis management (e.g. planning and preparedness).

By treating populations as homogeneous, SIR models cannot capture the fact that people – and the populations that they comprise – are heterogeneous, not only in their biology but in their behavior, too (Kiss et al. 2006; Poland et al. 2007; Cantrell et al. 2009). Evidence of in-population transmission heterogeneity – namely, the fact that some infected individuals in a given population cause more secondary infections than others – during real-world outbreaks is proof that $R_0$ should be treated as the mean of a random variable (See Chapter 3).

In Chapter 3, we use data from one such outbreak – the 2015 South Korean MERS epidemic – to empirically explore the effects of in-population transmission heterogeneity on disease propagation, as well as its potential causes. Then, in Chapter 4, we investigate how the
variance inherent to human populations – and illustrated in Chapter 3 – can be used to our advantage within the context of outbreak mitigation and intervention deployment.
Appendix I. Outbreak Takeoff and Early Extinction When $R_0 < 1$ (See Section IV)

As will be further discussed in Chapters 3 and 4, when $R_0$ is treated as the mean of a random variable, it need not be $>1$ in order for an outbreak to occur. Consider the following PMF:

$$P(N = n) = \begin{cases} 0.25 & n = 2 \\ 0.75 & n = 0 \end{cases}$$

$$R_0 = E[N] = 2(0.25) + 0(0.75) = 0.5$$

Here, though $R_0 = 0.5$, there is a 25% chance that an outbreak will takeoff. Furthermore, using Equation 13 we find that the probability of early extinction by $t = 2$ is $<1$:

$$P(\text{extinction by second generation}) = (1 - 0.5/2) + (0.5/2)(1 - 0.5/2)^2 \approx .891$$

Thus, it is not only possible that an outbreak may occur when $R_0 < 1$; in fact, it is also possible that such an outbreak will not experience early extinction.
Appendix II. Derivation of PMFs D, E, and F (See Section III)

**PMF D:**

\[ P(N = n) = \begin{cases} 
0.125 & n = 16 \\
0.875 & n = 0 
\end{cases} \]

\[ R_0 = E[N] = 16(0.125) + 0(0.875) = 2 \]

**PMF E:**

\[ P(N = n) = \begin{cases} 
0.0625 & n = 32 \\
0.9375 & n = 0 
\end{cases} \]

\[ R_0 = E[N] = 32(0.0625) + 0(0.9375) = 2 \]

**PMF F:**

\[ P(N = n) = \begin{cases} 
0.03125 & n = 64 \\
0.96875 & n = 0 
\end{cases} \]

\[ R_0 = E[N] = 64(0.03125) + 0(0.96875) = 2 \]
Appendix III: Likelihood Estimation of a Low Probability Event (See Section V)

As noted in the text, Monte Carlo simulations are limited in their ability to capture low probability events due to constraints on the number of trials that can be feasibly simulated. However, the likelihood associated with such low probability events may be easily estimated analytically for simple PMFs.

Let us consider again the low probability event described in Section V:

"a simulated outbreak via SIM F in which each and every individual infected between $t = 0$ and $t = 5$ infected 64 other individuals"

Recall that the probability associated with $n = 64$ (See PMF F, Appendix II) is $p = 0.03125$. Thus, the probability of the aforementioned event is as follows:

$$P(event) = (0.03125)(0.03125^{64})(0.03125^{64^2})(0.03125^{64^3})(0.03125^{64^4}) \approx 0$$
References

Chapter 3: An Empirical Example of In-Population Transmission Heterogeneity

I. Introduction

As discussed in the work presented in Chapter 2, deterministic SIR-type models – which assume in-population homogeneity – often fail to properly capture disease propagation during real-world outbreaks. This is largely due to the reality that human populations are heterogeneous, and as a result, not all infected individuals cause the same number of secondary infections – even when the affected population is \( \sim \) 100% susceptible.

Examples of in-population transmission heterogeneity (i.e. the fact that some individuals in a given population cause more secondary infections than others) during real-world outbreaks (including the 2015 South Korean MERS outbreak, which we investigate at length in this chapter) provide ample evidence that \( R_0 \) should be treated as the mean of a random variable. Given that \( R_0 \) is the product\(^9\) of \( C \) (the average rate of contact between those who are susceptible and those who are infected) and \( p \) (the average probability of infection given contact between an individual that is infected and an individual that is susceptible), anything that may impact individual rates of contact or individual probabilities of infection may be considered potential drivers of in-population transmission heterogeneity. For example, individuals with pre-existing conditions – especially those that are hospitalized for said conditions – may have a higher individual

---

\(^9\)Within the context of diseases that are transmitted via human-to-human infection (See Chapter 2, Section II, Part A)
probability of infection than the average population due to their compromised immune systems. Similarly, during outbreaks in nosocomial (i.e. hospital) settings, healthcare workers may have a higher individual rate of contact than the average population due to the duties associated with their profession.

Though drivers of in-population transmission heterogeneity vary from disease to disease and from population to population, successful identification of such predictors – and thus, of individuals at high risk of transmitting the infection in question – offers opportunities for smarter decision-making during real-world outbreaks (e.g. risk-informed intervention deployment). However, in addition to identification of characteristics that are merely associated with higher risk of transmission, the development of potential mechanisms of action for these proposed predictors of in-population transmission heterogeneity is also imperative.

In this chapter, we explore the 2015 South Korean MERS outbreak as a case study of real-world in-population transmission heterogeneity. First, we focus on the identification of characteristics associated with in-population transmission heterogeneity during the outbreak, as well as mechanistic explanations for said associations and the impact of said heterogeneity on disease propagation (Section II). We then discuss how in-population transmission heterogeneity – and the underlying variance it represents – can be used to our advantage during infectious disease outbreaks in non-homogeneous populations (e.g.

\footnote{An edited version of this case study was published in the *Transactions for the Royal Society of Tropical Medicine and Hygiene* in October 2017 (Majumder et al. 2017).}
risk-informed intervention deployment) (Section III). Finally, data, findings, and recommendations from Sections II and III are used to model and compare uniform deployment of interventions (i.e. in which all individuals are equally likely to receive an intervention) versus risk-informed intervention deployment (i.e. in which variance is taken into consideration those that are most at risk for passing on the infection are exclusively targeted) (Chapter 4).
II. In-Population Transmission Heterogeneity of Middle East Respiratory Syndrome in South Korea, 2015

In this section, we empirically investigate in-population transmission heterogeneity of MERS in South Korea by reconstructing the network of infectivity associated with the outbreak and assessing what types of cases were more likely to pass on the virus (and why). Part A provides an overview of MERS-\textit{Coronavirus}, as well as a description of the specific research questions explored in the remainder of the section within the South Korean context. We then describe the data sources and methods used to address these research questions in Part B, followed by study results and conclusions in Parts C and D, respectively. These results and conclusions are then further discussed in Section III, where we consider how lessons learned from the 2015 South Korean MERS outbreak can be used to model the potential utility of risk-informed intervention deployment in heterogeneous populations.

A. Background

Middle East Respiratory Syndrome (MERS), which is caused by MERS-\textit{Coronavirus} and can be transmitted from camels to humans and from humans to humans, was first discovered in 2012 and has since resulted in more than 1700 cases and 700 deaths worldwide (Memish et al. 2014a; Raj et al. 2014; ProMED Mail 2017; WHO 2017a). The vast majority of cases to date have been reported out of Saudi Arabia, where MERS is endemic due to frequent interactions between humans and dromedary camels (Memish et al. 2014a; Raj et al. 2014, Saudi MOH 2017; WHO 2017a). That said, the World Health Organization (WHO)
estimates that about 75% of cases in Saudi Arabia are due to human-to-human transmission, though the vast majority of cases cause no secondary infections (Chu et al. 2005; Penttinen et al. 2013; Tsiodras et al. 2014; WHO 2017).

Universally, MERS-\textit{Coronavirus} transmission appears to be heterogeneous, with notable amplification in healthcare settings (i.e. nosocomial amplification) due in part to the occurrence of nosocomial super-spreading events (Chu et al. 2005; Hijawi et al. 2013; Penttinen et al. 2013; Tsiodras et al. 2014; Majumder et al. 2014). During such events, a small handful of individuals are responsible for the vast majority of cases, which typically results in a greater number of total infections than what would be expected given average estimates of the basic reproduction number associated with the disease (i.e. $R_0 < 1$) (Chu et al. 2005; Hijawi et al. 2013; Penttinen et al. 2013; Tsiodras et al. 2014; Majumder et al. 2014). This was exemplified during the 2015 MERS outbreak in South Korea, which was initiated by a local businessman returning home from a visit to the Middle East and resulted in nearly 200 nosocomial cases over the course of just two months (\textbf{Figure 1}) (Cowling et al. 2015).
Chapter 3: An Empirical Example of In-Population Transmission Heterogeneity

Throughout the entirety of the outbreak, the South Korean government provided a wealth of publicly available information on case demographics, including case contact details crucial for reconstructing a network of infectivity. However, it remains unclear whether or not certain types of cases were more likely to cause direct (i.e. human-to-human) secondary transmissions than others, and if so, why.

In the remainder of this section, we explore the following two research questions: (1) Can a network of infectivity for the 2015 South Korean MERS outbreak be reconstructed, and if so, can we deduce which demographic characteristics (i.e. predictors) were associated with direct disease transmission within the nosocomial setting? (2) Can knowledge gleaned from the network of infectivity and analyses of demographic characteristics lend insight
into potential mechanisms for in-population transmission heterogeneity – and more specifically, nosocomial amplification – of MERS-\textit{Coronavirus} around the world?

Two separate research methods are used to address Research Questions 1 and 2; thus, they are referenced as Analyses 1 and 2 (respectively) in Sections B (Methods) and C (Results). We then discuss our findings across both analyses in Section D (Study Conclusions).

\textbf{B. Methods}

\textit{Analysis 1.} Publicly available case data were manually curated from text-based MERS reports via the South Korean Ministry of Health and Welfare (MOHW) (MOHW 2017). Whenever possible, the MOHW-derived data set was cross-checked against relevant WHO Disease Outbreak News reports (DONs); in such instances, matching between data sources was conducted using age, sex, and date of reporting (WHO 2017c). It is worth acknowledging that the use of publicly available data poses unique challenges; though such data enable timely execution of preliminary epidemiological research for novel and emerging pathogens such as MERS-\textit{Coronavirus}, case information is restricted to protect patient privacy. This is particularly problematic with respect to potential behavioral confounders (e.g. hospital shopping\textsuperscript{11}, individual hygiene practices, etc.) for which data

\textsuperscript{11}Hospital shopping may be defined as a practice in which patients seek care at multiple facilities before settling on a treatment location. This particular practice was identified as an issue in the earliest days of the 2015 South Korean MERS outbreak due to the novelty of the infection (which made diagnosis challenging and resulted in “bed-hopping” across hospitals); however, it is worth noting that this practice did not seem to permeate beyond the first generation of cases (Appendix I).
were not available. Because of this, follow-up analyses should be conducted pending availability of additional case data from the South Korean Ministry of Health and Welfare.

The following variables were available for 100% of the 186 South Korean MERS cases: case class (categorical: healthcare worker, patient, visitor); gender (binary); age (continuous, normally-distributed, mean = 55 years old); comorbidity status (e.g. presence of pre-existing conditions) (binary); and case outcome (binary). Case contact details necessary for deducing direct (i.e. human-to-human) cause of infection (i.e. identification number of causative infectious case) were available for 77% (N = 144) of non-index cases. Direct transmission of MERS-Coronavirus was presumed if a given case – during his or her infectivity period – had close contact with an individual that went on to be diagnosed with MERS within 2 weeks after exposure; close contact was defined as being within 6 feet of or in the same room (or care area) as a confirmed case for a prolonged period of time without wearing recommended personal protective equipment (CDC 2017). Thus, case contact details from 144 cases were used to construct a network of infectivity for the outbreak (Appendix I) and to approximate the number of direct secondary transmissions caused by each of the 186 cases (i.e. identification numbers); these counts were used to generate a histogram of direct secondary transmissions caused per infectious case (Figure 2).
Both univariate chi-square analysis and multivariate logistic regression were then conducted against all 186 cases to further refine our understanding of the relationship between independent demographic characteristics and likelihood of human-to-human transmission (Tables 1 and 2). Each of the 186 cases in the data set was categorized as either a "human-to-human transmission agent" ($N = 29$) or not ($N = 157$). Human-to-human transmission agents were defined as cases that caused one or more direct secondary MERS-Coronavirus infections. Binary dummy variables were created for case class, and the "healthcare worker" class was treated as the reference category. Age was retained as a continuous variable and all other demographic characteristics (e.g., gender, comorbidity status, and case outcome) were retained as binary variables.
### Table 1. Characteristics associated with human-to-human transmission, univariate statistics.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Average Number of Direct Secondary Infections Caused</th>
<th>% Human-to-Human Transmission Agents</th>
<th>$\chi^2$</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Case Class</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Healthcare</td>
<td>0.03</td>
<td>3%</td>
<td>5.64</td>
<td>0.06</td>
</tr>
<tr>
<td>Worker</td>
<td>0.32</td>
<td>12%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Visitor</td>
<td>1.09</td>
<td>20%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patient</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td></td>
<td></td>
<td>0.67</td>
<td>0.41</td>
</tr>
<tr>
<td>Female</td>
<td>0.30</td>
<td>13%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>1.11</td>
<td>17%</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Age</strong></td>
<td></td>
<td></td>
<td>0.33</td>
<td>0.57</td>
</tr>
<tr>
<td>&lt;55 Years Old</td>
<td>0.99</td>
<td>14%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥55 Years Old</td>
<td>0.59</td>
<td>17%</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Comorbidity</strong></td>
<td></td>
<td></td>
<td>3.33</td>
<td>0.07</td>
</tr>
<tr>
<td>Status</td>
<td>0.58</td>
<td>13%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-Comorbid</td>
<td>1.77</td>
<td>27%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comorbid</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Case Outcome</strong></td>
<td></td>
<td></td>
<td>5.04</td>
<td>0.02</td>
</tr>
<tr>
<td>Recovered</td>
<td>0.82</td>
<td>13%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Deceased</td>
<td>0.58</td>
<td>28%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2. Characteristics associated with human-to-human transmission, multivariate statistics.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Adjusted Odds Ratio</th>
<th>95% CI</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Case Class</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Healthcare Worker</td>
<td>Reference</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Visitor</td>
<td>4.71</td>
<td>0.47 - 46.80</td>
<td>0.19</td>
</tr>
<tr>
<td>Patient</td>
<td>8.35</td>
<td>0.95 - 73.01</td>
<td>0.06</td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>Reference</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>1.08</td>
<td>0.45 - 2.58</td>
<td>0.87</td>
</tr>
<tr>
<td><strong>Age (Continuous)</strong></td>
<td>0.98</td>
<td>0.94 - 1.01</td>
<td>0.13</td>
</tr>
<tr>
<td><strong>Comorbidity Status</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-Comorbid</td>
<td>Reference</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comorbid</td>
<td>1.42</td>
<td>0.50 - 4.04</td>
<td>0.51</td>
</tr>
<tr>
<td><strong>Case Outcome</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recovered</td>
<td>Reference</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Deceased</td>
<td>2.96</td>
<td>1.03 - 8.48</td>
<td>0.04</td>
</tr>
</tbody>
</table>

**Analysis 2.** Results from Analysis 1 were used to inform a hypothesis-directed literature review, through which mechanistic explanations for the data-driven findings were developed. Search terms for the hypothesis-directed literature review can be found in Appendix II. For increased generalizability, the literature review was expanded beyond MERS-\textit{Coronavirus} to include the virologically related and considerably less novel SARS-\textit{Coronavirus}.\textsuperscript{12}

\textsuperscript{12}Like the 2015 MERS outbreak in South Korea, the 2003 SARS outbreak in Hong Kong (and related outbreaks) also exhibited early super-spreading that was mitigated despite the absence of pharmaceutical interventions (Li et al. 2004). This, combined with the virological similarities across pathogens, makes literature pertaining to SARS-\textit{Coronavirus} an ideal supplement to the smaller volume of literature currently available regarding MERS-\textit{Coronavirus}. 

65
Chapter 3: An Empirical Example of In-Population Transmission Heterogeneity

C. Results

**Analysis 1.** On average, cases caused 0.77 direct secondary transmissions (i.e. $R_0 = 0.77$) of MERS-Coronavirus throughout the course of the outbreak; however, significant heterogeneity existed across various demographic characteristics, and the vast majority of cases (84%) caused zero secondary infections (Table 1). In other words, not all cases were equally likely to cause secondary infections (i.e. in-population transmission heterogeneity), and some were far more likely to cause secondary infections than others (i.e. nosocomial super-spreading events).

The percentage of deceased cases that were human-to-human transmission agents (i.e. MERS cases that caused one or more direct secondary infections) was significantly greater than that of recovered cases ($\chi^2 = 5.04$, $p = 0.02$) (Table 1). Univariate statistics for all other demographic characteristics were insignificant ($p \geq 0.05$) (Table 1). After controlling for all demographic characteristics via multivariate regression, case outcome remained statistically significant ($p < 0.05$); cases who died were more likely to pass on the virus than those who survived, thus making case outcome a robust predictor for heterogeneous human-to-human transmission (Table 2). When compared to those who recovered, odds of being a human-to-human transmission agent were nearly 3 times higher for those who died from MERS (adjusted odds ratio: 2.96; 95% confidence interval [95%CI]: 1.03, 8.48) (Table 2).
The histogram of direct secondary transmission caused per infectious case was right-tailed with non-zero variance; though most (84%) \((N = 157)\) infectious cases caused zero direct secondary transmissions, a small fraction (16%) caused 1 or more and served as the primary propagators of nosocomial amplification (Figure 2). Meanwhile, 23\% \((N = 42)\) of cases had unknown causes of infection that could not be directly attributed to a human-to-human transmission agent, thus suggesting the possibility of indirect transmission as an additional source of infection.

**Analysis 2.** To develop mechanistic explanations for the findings from Analysis 1 – namely, that case outcome was a statistically significant predictor for heterogeneous human-to-human transmission of MERS-Coronavirus in South Korea and that causative agents for 42 cases were unknown – a two-pronged, hypothesis-directed literature review was utilized.

**Literature Review I: Clinical Manifestations.** A review of the relevant literature revealed that clinical manifestations – namely, deposition location of virus particles and viral load – might impact transmissibility of a variety of respiratory pathogens (Lee et al. 2009; Li et al. 2010; Thomas 2013; Bourouiba et al. 2014). Lower respiratory tract (LRT) infections often necessitate smaller particle size; thus, upon expulsion, the virus particles can stay suspended in the air for longer, where they may be redirected and dispersed by airflow (Thomas 2013; Bourouiba et al. 2014; Bourouiba 2016). Similarly, high viral load of respiratory pathogens has been shown to correlate with high rates of viral shedding into the surrounding environment (Pitzer et al. 2007; Lee et al. 2009; Li et al. 2010). Variance in viral load has been noted among MERS cases, and both lower and upper respiratory tract
MERS-"Coronavirus" infections have been observed (Drosten et al. 2013; Memish et al. 2013; Omrani et al. 2013; Memish et al. 2014b; Al-Gethamy et al. 2015). Thus, to mechanistically explain the relationship between case outcome and heterogeneous human-to-human transmission, the following hypothesis was further explored using the search terms specified in Appendix II: "LRT infections and/or higher viral load are more common in individuals who suffer from severe disease and/or fail to recover from MERS (or SARS)." Of the 12,635 total articles recommended by Google Scholar, 632 abstracts were considered. 12 articles were relevant to the aforementioned hypothesis (Table 3). All 12 articles found either LRT infections or higher viral load to be more common among severe and fatal MERS and SARS cases (Table 3). None of the articles reviewed nullified the aforementioned hypothesis.

Table 3. Clinical manifestations, summary of relevant articles.

<table>
<thead>
<tr>
<th>Citation Number</th>
<th>Causative Agent</th>
<th>Number of Cases Studied</th>
<th>Relevant Finding</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>SARS-Coronavirus</td>
<td>218</td>
<td>Increased viral shedding in cases with severe disease</td>
</tr>
<tr>
<td>62</td>
<td>SARS-Coronavirus</td>
<td>323</td>
<td>Higher viral load in cases that failed to recover</td>
</tr>
<tr>
<td>13</td>
<td>SARS-Coronavirus</td>
<td>415</td>
<td>Increased viral shedding in cases with severe disease</td>
</tr>
<tr>
<td>15</td>
<td>SARS-Coronavirus</td>
<td>133</td>
<td>Higher viral load in cases that failed to recover</td>
</tr>
<tr>
<td>28</td>
<td>SARS-Coronavirus</td>
<td>154</td>
<td>Higher viral load in cases that failed to recover</td>
</tr>
<tr>
<td>16</td>
<td>SARS-Coronavirus</td>
<td>79</td>
<td>Higher viral load among cases that failed to recover</td>
</tr>
</tbody>
</table>
Chapter 3: An Empirical Example of In-Population Transmission Heterogeneity

<table>
<thead>
<tr>
<th></th>
<th>MERS-Coronavirus</th>
<th></th>
<th>Higher viral load in LRT in cases that failed to recover</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td></td>
<td>2</td>
<td>Higher viral load in LRT in case that failed to recover</td>
</tr>
<tr>
<td>42</td>
<td>MERS-Coronavirus</td>
<td>26</td>
<td>Increased viral shedding in cases with severe disease</td>
</tr>
<tr>
<td>50</td>
<td>MERS-Coronavirus</td>
<td>2</td>
<td>Higher viral load in case that failed to recover</td>
</tr>
<tr>
<td>36</td>
<td>MERS-Coronavirus</td>
<td>37</td>
<td>Higher viral load in LRT in cases that failed to recover</td>
</tr>
<tr>
<td>37</td>
<td>MERS-Coronavirus</td>
<td>102</td>
<td>Higher viral load in cases that failed to recover</td>
</tr>
<tr>
<td>38</td>
<td>MERS-Coronavirus</td>
<td>14</td>
<td>Higher viral load in LRT in cases that failed to recover</td>
</tr>
</tbody>
</table>

Literature Review II: Indirect Transmission via Environmental Contamination. Given that 23% of cases in the reconstructed infectivity network could not be directly attributed to a causative agent, it is possible that – in addition to heterogeneous human-to-human transmission of MERS-Coronavirus – environmental contamination may have also played a role in nosocomial amplification during the 2015 South Korean MERS outbreak via indirect transmission. A review of the relevant literature revealed that pathogen density and indoor confinement of patients might result in contamination of surfaces, fomites, and indoor ventilation systems by viral infections (Moser et al. 1979; Block et al. 1985; Butz et al. 1993; Boone and Gerba 2007). Moreover, recent studies have shown – via imaging, fluid dynamics analysis, and mathematical modeling – how respiratory emissions from patients can lead to both surface and long-term air contamination. In particular, this newer research – which focuses on examining the fundamental mechanics of disease transmission – demonstrates that respiratory droplet deposition and suspension in indoor spaces is
dependent on the coupling between host physiology and indoor environmental conditions (Bourouiba et al. 2014; Bourouiba 2016; Scharfman et al. 2016). Therefore, to mechanistically explain the potential role of environmental contamination in nosocomial amplification of MERS-\textit{Coronavirus} transmission, the following hypothesis was further explored using the search terms specified in Appendix II: “Indirect transmission of MERS-\textit{Coronavirus} (or SARS-\textit{Coronavirus}) via fomites and/or indoor ventilation systems has been posited or deemed probable in nosocomial settings.” Of the 4,098 total articles recommended by Google Scholar, 489 abstracts were considered. 14 articles were relevant to the aforementioned hypothesis (Table 4). All 14 articles posited or demonstrated either fomite or indoor ventilation system transmission of MERS- or SARS-\textit{Coronavirus} in nosocomial settings (Table 4). None of the articles reviewed nullified the aforementioned hypothesis.
Chapter 3: An Empirical Example of in-Population Transmission Heterogeneity

Table 4. Indirect transmission via environmental contamination, summary of relevant articles.

<table>
<thead>
<tr>
<th>Citation Number</th>
<th>Causative Agent</th>
<th>Environmental Contamination Studied</th>
<th>Relevant Finding</th>
</tr>
</thead>
<tbody>
<tr>
<td>29</td>
<td>SARS-Coronavirus</td>
<td>Indoor ventilation system</td>
<td>Probable transmission</td>
</tr>
<tr>
<td>32</td>
<td>SARS-Coronavirus</td>
<td>Fomite</td>
<td>Posited transmission</td>
</tr>
<tr>
<td>20</td>
<td>SARS-Coronavirus</td>
<td>Fomite</td>
<td>Posited transmission</td>
</tr>
<tr>
<td>31</td>
<td>SARS-Coronavirus</td>
<td>Both</td>
<td>Probable transmission</td>
</tr>
<tr>
<td>6</td>
<td>SARS-Coronavirus</td>
<td>Both</td>
<td>Posited transmission</td>
</tr>
<tr>
<td>36</td>
<td>SARS-Coronavirus</td>
<td>Indoor ventilation system</td>
<td>Probable transmission</td>
</tr>
<tr>
<td>69</td>
<td>SARS-Coronavirus</td>
<td>Indoor ventilation system</td>
<td>Probable transmission</td>
</tr>
<tr>
<td>14</td>
<td>SARS-Coronavirus</td>
<td>Indoor ventilation system</td>
<td>Probable transmission</td>
</tr>
<tr>
<td>63</td>
<td>MERS-Coronavirus</td>
<td>Fomite</td>
<td>Posited transmission</td>
</tr>
<tr>
<td>3</td>
<td>MERS-Coronavirus</td>
<td>Fomite</td>
<td>Posited transmission</td>
</tr>
<tr>
<td>22</td>
<td>MERS-Coronavirus</td>
<td>Fomite</td>
<td>Probable transmission</td>
</tr>
<tr>
<td>26</td>
<td>MERS-Coronavirus</td>
<td>Indoor ventilation system</td>
<td>Posited transmission</td>
</tr>
<tr>
<td>34</td>
<td>MERS-Coronavirus</td>
<td>Fomite</td>
<td>Probable transmission</td>
</tr>
<tr>
<td>56</td>
<td>MERS-Coronavirus</td>
<td>Fomite</td>
<td>Posited transmission</td>
</tr>
</tbody>
</table>
In this study, a combination of data-driven and review-driven approaches was utilized to propose potential mechanisms for nosocomial amplification of MERS-Coronavirus during the 2015 South Korean MERS outbreak.

For the data-driven components of this paper (Analysis 1), a potential network of infectivity was reconstructed using available case contact details from the 2015 South Korean MERS outbreak. 84% of infectious cases caused zero direct secondary infections, and overall, cases caused an average of 0.77 direct secondary transmissions (i.e. $R_0 = 0.77$) each. This high percentage of likely explains why the outbreak was so short-lived – persisting for barely 5 generations ($t = 13$ days) even in the absence of disease-specific pharmaceutical and non-pharmaceutical interventions (Park et al. 2016; Widagdo et al. 2017).

Direct cause of infection (i.e. identification number of causative infectious case) was deducible for 77% of non-index cases, from which a distribution of direct secondary infections caused per case (i.e. identification number) were approximated; the histogram generated from the available data was right-tailed, which is consistent with previous work regarding in-population transmission heterogeneity of respiratory pathogens and nosocomial super-spreading events (Figure 2) (Teytelman and Larson 2012; Finkelstein et al. 2015). The remaining 23% of cases were likely caused by undocumented contact.
between cases (i.e. direct transmission) or indirect transmission of MERS-\textit{Coronavirus} via environmental contamination.

Assuming that some portion of these cases were in fact due to undocumented direct transmission, results from both the univariate and multivariate statistical analyses were likely biased towards the null (Tables 1 and 2). However, this also suggests that the statistically significant dependent variable that did emerge (i.e. case outcome) is a robust predictor of heterogeneous human-to-human transmission – and thus, nosocomial amplification – within the context of the 2015 South Korean MERS outbreak.

Review of the relevant literature indicates that the mechanistic relationship between case outcome and heterogeneous human-to-human transmission of MERS-\textit{Coronavirus} may be due to the fact that LRT infections and higher viral load appear to be more common among severe and fatal coronavirus infections (Table 3). Furthermore, just as LRT infections and higher viral load among fatal cases may increase the probability of direct secondary infections, such clinical manifestations may also increase the probability of indirect transmission. Thus, under the assumption that the statistical results from Analysis 1 are applicable to other nosocomial outbreaks of MERS-\textit{Coronavirus}, interventions such as contact tracing should be prioritized for cases that fail to recover (post-mortem), as well as for cases that have a high risk of mortality, such as those with pre-existing conditions (i.e. the comorbid) (Majumder et al. 2015; Rivers et al. 2016).
While heterogeneous human-to-human transmission contributed substantially to nosocomial amplification during the 2015 South Korean MERS outbreak, it is possible that environmental contamination – as noted above – played a role as well. The relevant literature suggests that fomite and indoor ventilation system transmission may be the mechanism through which environmental contamination potentially amplifies disease transmission in hospital settings (Table 4). Though fomite transmission of viral pathogens is notoriously challenging to eliminate (e.g. via sanitization), simulation studies involving computational fluid dynamics suggest that low-cost solutions for the mitigation of indoor ventilation system transmission may exist (Jiang et al. 2009; Yam et al. 2011). Unfortunately, such simulations remain limited in terms of use for recommendations due to lack of calibration and excessive utilization of free parameters, as well as notable absences of physical modeling for critical small-scale processes such as mixing and pathogen re-suspension among others. However, recent validation studies – which have directly measured the role of violent respiratory events (e.g. sneezing and coughing) on indoor environmental contamination – show a new path ahead where pathogen loads, transport, and mixing from a range of patient types can be addressed (Bourouiba et al. 2014; Bourouiba 2016; Scharfman et al. 2016). Moving forward, such studies will likely be integral to the development and effective deployment of more robust solutions for low cost re-design of indoor ventilation systems and patient-specific pathogen containment strategies, which should be implemented and evaluated in hospitals that are likely to treat MERS cases in the future.
Cross-population generalizability of these findings to Saudi Arabia – where MERS-
Coronavirus is endemic and causes frequent nosocomial outbreaks – requires deeper
consideration given the additional complexity posed by both the prevalence of zoonotic
transmissions from dromedary camels to humans and the occurrence of household and
communal spread outside of the healthcare context. It is possible that the characteristics
associated with heterogeneous human-to-human transmission may differ under such
circumstances; as a result, the role of zoonotic cases and non-nosocomial human-to-human
disease transmission during hospital outbreaks of MERS must be further explored. A rich
privately-curated data set recently obtained from the Saudi Arabian Ministry of Health –
which includes information on demographic characteristics (e.g. case outcome, zoonotic
and human contacts, etc.) and clinical manifestations (e.g. presence or absence of LRT
infection, viral load, etc.) of cases that have sought care at government hospitals – may
allow us to investigate the possible relationship between camel-acquired cases and direct
secondary transmissions caused by said cases in the near future.

Though they are most certainly interconnected, direct transmission (i.e. human-to-human)
and indirect transmission (i.e. via environmental contamination) of MERS-Coronavirus
were treated as decidedly distinct in this paper. Prior mathematical modeling studies in
other disease contexts have shown that competition between the timescales of these two
transmission routes is key to determining the dominant observed effect on and nature of
the epidemic curve (Bourouiba et al. 2011; Bourouiba et al. 2014). Unfortunately, the
number of indirect secondary transmissions caused per case was not discernable from the
available data and could likely have only been determined via molecular epidemiology
methods (e.g. viral phylogenetic analysis). If possible, such methods should be employed during future investigations of nosocomial MERS outbreaks in Saudi Arabia (and elsewhere, as is applicable) to lend insight into the real-world interactions between direct and indirect transmission.

At present, it seems reasonable to posit that individuals who are more likely to cause direct secondary transmissions (i.e. human-to-human transmission agents) may also be more likely to cause indirect secondary transmissions (i.e. via environmental contamination) within nosocomial settings. However, high rates of human-to-human contact in healthcare settings suggest that direct transmissions are considerably more likely than indirect transmissions, which is further supported by the fact that direct causes of human-to-human infection were deducible for 77% of cases in our data set. Nevertheless, if human-to-human transmission agents are indeed more likely to experience LRT infections and higher viral loads in such settings, elongated mid-air suspension of virus particles and increased rates of viral shedding may result in considerable environmental contamination of indoor ventilation systems and fomites.

Ideally, thorough sanitization of all fomites and ventilation system components should be implemented to mitigate environmental contamination; however, doing so may be cost-prohibitive in low-resource healthcare settings. With this in mind, our study suggests that – in addition to closely monitoring contacts of likely human-to-human transmission agents (i.e. MERS cases that fail to recover and cases at high risk of death) during future nosocomial outbreaks of MERS – potential fomites with which they may have had contact
Chapter 3: An Empirical Example of In-Population Transmission Heterogeneity

should be sanitized preferentially. Furthermore, if solutions for the mitigation of indoor ventilation system transmission cannot be implemented hospital-wide, wards intended for biosecurity level 2 respiratory pathogens (i.e. MERS-\textit{Coronavirus}) should be given precedence.

Our study – and recommendations therein – are not without limitations, however. Due to our use of publicly available data, the dimensions of our data-driven analysis (Analysis 1) were determined by what the South Korean Ministry of Health and Welfare has shared. As a result, direct causes of MERS-\textit{Coronavirus} infection (i.e. identification number of causative infectious case) were not available for 23% of cases, and additional variables that may have been of interest – such as length of hospital stay and viral phylogenetic data – were either incomplete or unavailable for analysis. It is also possible that our results were confounded by potential behavioral factors (e.g. hospital shopping, hygiene practices, etc.); unfortunately, data for such behavioral factors were not available at time of analysis – though they likely played an important role in heterogeneous human-to-human transmission of MERS-\textit{Coronavirus} in South Korea. Furthermore, because MERS-\textit{Coronavirus} and MERS remain relatively novel to the existing literature, the review-driven components of our study (Analysis 2) were broadened to include SARS-\textit{Coronavirus} and SARS, which – though virologically and clinically similar – are not without their differences (Momattin et al. 2013; Al-Tawfiq et al. 2014). Thus, the mechanistic explanations proposed in Analysis 2 are tentative at best.
Chapter 3: An Empirical Example of In-Population Transmission Heterogeneity

The proposed mechanisms – including clinical manifestations and indirect transmission via environmental contamination – by which case outcome acts as a predictor for heterogeneous human-to-human transmission and nosocomial amplification must be studied more extensively in order to establish causality. In particular, indirect transmission (i.e. via fomites and/or indoor ventilation systems) must be demonstrably correlated with both environmental contamination and nosocomial amplification. Similarly, clinical manifestations (i.e. LRT infection and/or high viral load) must be demonstrably correlated with both case outcome and direct or – indirect – transmission in future nosocomial outbreaks of MERS-\textit{Coronavirus}.

It is also worth considering that case outcome may be partially confounded with length of hospital stay. Cases that failed to recover from MERS may have spent longer periods of time within the hospital setting, giving them greater opportunity to cause both direct and indirect transmission – perhaps via the mechanisms proposed above. When length of hospital stay was included in our multivariate logistic regression model and run against cases for which these data were available ($N = 166$), case outcome remained the strongest predictor for heterogeneous human-to-human transmission of MERS-\textit{Coronavirus} in South Korea, followed by length of hospital stay (Appendix III). Notably, the inclusion of length of hospital stay in the model moderately dampens the statistical significance of case outcome as a predictor, indicating that a confounding relationship between the two variables may indeed exist. This said, given sizeable missingness of length of hospital stay data (>10%), results from this secondary analysis should be treated as preliminary and will be further extended given availability of additional data.
Chapter 3: An Empirical Example of In-Population Transmission Heterogeneity

In addition to length of hospital stay, other potentially confounding variables exist as well. For example, MERS cases that failed to recover may have received more visitors on average due to the severity of their illness or been more likely to undergo aerosolizing procedures (e.g. bronchoscopy) in advanced stages of illness than those who recovered. Forthcoming work will explore these factors – namely, length of stay, visitor volume, and virus dissemination (including aerosolizing procedures) – and their impact on heterogeneous human-to-human transmission and nosocomial amplification of MERS-
Coronavirus in the future.
III. Further Discussion and Next Steps

As shown in Section II, Middle East Respiratory Syndrome in South Korea is a real-world example of in-population transmission heterogeneity that provides empirical evidence for the treatment of $R_0$ as the mean of a random variable. In fact, using the 2015 MERS outbreak as a case study, we show that the $R_0$ of a disease that is transmitted via human-to-human infection need not be $>1$ for an outbreak to occur if the underlying variance is sufficiently large. Furthermore, we demonstrate that it is possible to deduce predictors of heterogeneous human-to-human transmission when a network of infectivity from an outbreak of interest can be reconstructed, when demographic characteristics are available for the cases that occurred, and when mechanistic explanations can be ascertained from the existing literature.

With this in mind, data like those shared by the South Korean MOHW are of the utmost necessary for further exploration of non-zero variance during real-world outbreaks of diseases beyond MERS. Such data allow us to assess disease-specific drivers of in-population transmission heterogeneity, which can then be used to mechanistically explain more specific, highly inter-related phenomena – such as nosocomial amplification and nosocomial super-spreading events. By better understanding what types of individuals are most at risk of passing on an infection of interest, interventions may (in theory) be targeted towards such individuals – especially in low-resource settings where uniform deployment may be cost-prohibitive and in non-homogeneous populations where the vast majority of
individuals are unlikely to act as human-to-human transmission agents (e.g. MERS in South Korea).

However, given the paucity of data sources similar to those presented in Section II from the 2015 South Korean MERS outbreak, attempts to validate the actual utility of a risk- or variance-informed intervention deployment paradigm – especially within the context of a well-described heterogeneous population – have been limited. Thus, in Chapter 4, we combine the analytical and simulation methods presented in Chapter 2 with the data, results, and conclusions presented here in Chapter 3 to compare traditional (i.e. uniform) approaches to intervention deployment against approaches that are risk-informed. As opposed to using purely synthetic data as we did in Chapter 2, we use data from MERS in South Korea to more realistically depict the underlying variance that exists during real-world outbreaks of infectious disease. Thus, our overarching findings – with respect to the utility of risk-informed versus uniform intervention deployment – from Chapter 4 are likely generalizable to all non-homogeneous populations in which predictors (or proxy indicators) of heterogeneous human-to-human transmission for any given disease of interest are known.
Appendix I: Network of Infectivity for the 2015 South Korean MERS Outbreak

Please navigate to https://www.sciencenews.org/article/anatomy-south-korean-mers-outbreak for interactive and downloadable network of infectivity for the 2015 South Korean MERS outbreak. This work was conducted in collaboration between Maimuna S. Majumder and *Science News Magazine*. 
Appendix II: Search Terms for Hypothesis-Directed Literature Review

Literature Review I: Clinical Manifestations

- “MERS-Coronavirus” OR “Middle East respiratory syndrome” OR “MERS-CoV” AND “lower respiratory” (678 results; 5% considered)
- “MERS-Coronavirus” OR “Middle East respiratory syndrome” OR “MERS-CoV” AND “viral load” (506 total results; 5% considered)
- “MERS-Coronavirus” OR “Middle East respiratory syndrome” OR “MERS-CoV” AND “viral shedding” (195 total results; 5% considered)
- “SARS-Coronavirus” OR “Severe acute respiratory syndrome” OR “SARS-CoV” AND “lower respiratory” (4830 total results; 5% considered)
- “SARS-Coronavirus” OR “Severe acute respiratory syndrome” OR “SARS-CoV” AND “viral load” (5200 total results; 5% considered)
- “SARS-Coronavirus” OR “Severe acute respiratory syndrome” OR “SARS-CoV” AND “viral shedding” (1220 total results; 5% considered)

Literature Review II: Indirect Transmission Via Environmental Contamination

- “MERS-Coronavirus” OR “Middle East respiratory syndrome” OR “MERS-CoV” AND “indoor” AND “ventilation” (53 total results; 100% considered)
- “MERS-Coronavirus” OR “Middle East respiratory syndrome” OR “MERS-CoV” AND “indoor” AND “air” (107 total results; 100% considered)
Chapter 3: An Empirical Example of In-Population Transmission Heterogeneity

- “MERS-Coronavirus” OR “Middle East respiratory syndrome” OR “MERS-CoV” AND “fomite” (139 total results; 100% considered)
- “SARS-Coronavirus” OR “Severe acute respiratory syndrome” OR “SARS-CoV” AND “indoor” AND “ventilation” (1230 total results; 5% considered)
- “SARS-Coronavirus” OR “Severe acute respiratory syndrome” OR “SARS-CoV” AND “indoor” AND “air” (2220 total results; 5% considered)
- “SARS-Coronavirus” OR “Severe acute respiratory syndrome” OR “SARS-CoV” AND “fomite” (349 total results; 5% considered)
Appendix III: Supplementary Table

*Table Supp.* Characteristics associated with human-to-human transmission, multivariate statistics (including length of hospital stay data for $N = 166$ cases)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Adjusted Odds Ratio</th>
<th>95% CI</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Case Class</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Healthcare Worker</td>
<td>Reference</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Visitor</td>
<td>1.97</td>
<td>0.16 - 24.71</td>
<td>0.60</td>
</tr>
<tr>
<td>Patient</td>
<td>5.93</td>
<td>0.66 - 53.39</td>
<td>0.11</td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>Reference</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>0.85</td>
<td>0.33 - 2.17</td>
<td>0.73</td>
</tr>
<tr>
<td><strong>Age (Continuous)</strong></td>
<td></td>
<td>0.94 - 1.01</td>
<td>0.22</td>
</tr>
<tr>
<td><strong>Comorbidity Status</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-Comorbid</td>
<td>Reference</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comorbid</td>
<td>2.18</td>
<td>0.72 - 6.63</td>
<td>0.17</td>
</tr>
<tr>
<td><strong>Case Outcome</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recovered</td>
<td>Reference</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Deceased</td>
<td>3.51</td>
<td>0.95 - 12.90</td>
<td>0.06*</td>
</tr>
<tr>
<td><strong>Length of Hospital Stay</strong></td>
<td>1.07</td>
<td>0.99 - 1.14</td>
<td>0.06*</td>
</tr>
</tbody>
</table>

*Rounded up from $p = 0.059$ (case outcome) and $p = 0.063$ (length of hospital stay), respectively.
Chapter 3: An Empirical Example of In-Population Transmission Heterogeneity

References


Chapter 3: An Empirical Example of In-Population Transmission Heterogeneity


Chapter 4: In-Population Heterogeneity and Deployment of Interventions

I. Introduction

During infectious disease outbreaks, interventions are often deployed uniformly such that all members of the effected population have an equal likelihood of receiving the resource of interest (e.g. vaccines, soap for hand-washing, etc.). Much like the SIR-type models that are typically used to study such outbreaks, this approach to intervention deployment assumes that all individuals have an equal probability of passing on the infection of interest and that transmission of the disease is homogenous. However, as we demonstrated in Chapter 3, real-world populations are heterogeneous – and as a result, some individuals are more likely to transmit a given disease than others (i.e. in-population transmission heterogeneity).

In Chapter 3, we focused specifically on characterizing in-population transmission heterogeneity within the context of MERS in South Korea due to a wealth of publicly available information that allowed us to reconstruct the network of infectivity associated with the outbreak and to identify possible predictors for heterogeneous human-to-human transmission of MERS-Coronavirus within the South Korean nosocomial context. Given the novelty of MERS-Coronavirus in South Korea, no known pharmaceutical or non-pharmaceutical interventions were deployed over the course of the outbreak; nevertheless, the outbreak died out after only 5 generations because most cases went on to infect no one – even though a handful of cases (i.e. nosocomial super-spreaders) infected more than 20
other individuals. However, lessons learned about drivers of heterogeneous transmission risk may allow for the risk-informed intervention deployment during future outbreaks of MERS in nosocomial settings – especially as pharmaceutical interventions designed specifically for mitigating MERS proceed down the development pipeline (Du et al. 2016).

Though another outbreak of MERS in South Korea is unlikely to occur, Saudi Arabia – where the coronavirus is endemic – has experienced several nosocomial outbreaks since discovery of the virus in 2012 (Chu et al. 2005; Hijawi et al. 2013; Penttinen et al. 2013; Tsiodras et al. 2014; Majumder et al. 2014). However, caution must be used when seeking to generalize findings from South Korea to Saudi Arabia, due to the considerable contextual differences between the two countries as described in Chapter 3 (i.e. cross-population variability). Nevertheless, because we were able to identify a physical (i.e. natural science) mechanism by which the primary predictor of heterogeneous human-to-human transmission (i.e. case outcome) affected transmissibility (e.g. outcome-dependent invasion of the lower respiratory tract and viral load, etc.), it is reasonable to generalize findings from South Korea to other contexts (with sufficient mindfulness of cross-population variability) until data similar to those presented in Chapter 3 are available for Saudi Arabia and any other country that may experience a nosocomial outbreak of MERS in the near future.

One of the chief recommendations posited in Chapter 3 is the use of risk-informed intervention deployment for future outbreaks of MERS, in Saudi Arabia and elsewhere, that specifically targets individuals that are at higher risk of passing on the infection (i.e. those
who die from MERS or are at a higher risk of dying, such as those who are co-morbid). However, formal assessments of the utility associated with such a risk-informed intervention deployment scheme – as compared to the more traditional (i.e. uniform) schemes used today – have been limited due to the sparsity of data similar to those that were shared publicly during the 2015 South Korean MERS outbreak.

Modeling the possible benefits of risk- or variance-informed intervention deployment is critical to incentivizing data-sharing during outbreaks, and such data-sharing is critical to determining predictors of human-to-human in-population transmission heterogeneity for emerging infectious diseases beyond MERS. When these predictors are paired with sufficiently robust mechanisms of operation (as shown in Chapter 3), they can then be used to deploy risk-informed interventions in other populations that are likely to experience an outbreak of the disease of interest.

Thus, here in Chapter 4, we use an extension of the analytical and simulation methods presented in Chapter 2 to compare the effectiveness of risk-informed intervention deployment against the effectiveness of uniform intervention deployment within the context of human-to-human disease transmission in heterogeneous populations. Instead of using synthetic data as we did in Chapter 2, we use data from the 2015 South Korean MERS outbreak to more accurately capture the underlying variance of real-world infectious disease outbreaks. With this in mind, the high-level findings we present in this chapter – namely, whether or not a risk-informed approach to intervention deployment should be pursued as an alternative to the traditional (i.e. uniform) approach – are designed to be
generalizable across all non-homogeneous populations in which predictors (or proxy indicators) of heterogeneous human-to-human transmission (and the physical mechanisms by which they operate) for a given disease of interest are known.

We begin with a brief overview of our data (from Chapter 3) and methods (from Chapter 2) in Section II, after which we present our results in Section III. Finally, we discuss our findings – and their real-world implications – in Section IV, which serves as the foundation for our policy recommendations in Chapter 5.
II. Methods

In this section, we outline the methods used to assess the effectiveness of risk-informed intervention deployment in heterogeneous populations. Because we are using MERS in South Korea as an example of one such heterogeneous population, we start by first deriving a probability mass function (PMF) from the data presented in Chapter 3 of the outbreak of interest (Part A). We then use this PMF to parameterize three different simulation models (Part B), each of which describe a different intervention scenario:

- No intervention deployed (i.e. NO INT)
- Uniform intervention deployment (i.e. INT 1)
- Risk-informed intervention deployment (i.e. INT 2)

Beyond the purpose of parameterizing our simulation models, we also use this PMF to analytically assess the probability of takeoff and the probability of early extinction in the no-intervention scenario. These results are later compared (See Section IV) against those obtained from the NO INT simulation model as a form of cross-validation between analytical and simulation approaches.

A. Analytical Approach: Defining a PMF for the 2015 South Korean MERS Outbreak

Using the network of infectivity that we reconstructed in Chapter 3 for the 2015 South Korean MERS outbreak and assuming that the individuals who were infected during the outbreak were representative of the greater population to which they belonged (i.e.
including those who were *not* infected in addition to those who were), we first derived a probability mass function as follows (PMF ROK):

\[
P(N = n) = \begin{cases} 
1/186 & n = 43 \\
1/186 & n = 28 \\
1/186 & n = 23 \\
1/186 & n = 9 \\
1/186 & n = 6 \\
1/186 & n = 4 \\
1/186 & n = 3 \\
6/186 & n = 2 \\
16/186 & n = 1 \\
157/186 & n = 0 
\end{cases} \quad (PMF ROK)
\]

\[
R_0 = E[N] = 43(1/186) + 28(1/186) + 23(1/186) + 9(1/186) + 6(1/186) + 4(1/186) \\
+ 3(1/186) + 2(6/186) + 1(16/186) + 0(157/186) \approx 0.77
\]

Then, using an analytical method similar to the one introduced in Chapter 2, we used this PMF to estimate the probability of takeoff by generation 1 (i.e. \(t = 1\)) and early extinction by generation 2 (i.e. \(t = 2\)) in a no-intervention scenario (See Section III).

**B. Simulation Approach: Modeling Uniform vs. Risk-Informed Intervention Deployment**

Following the derivation of PMF ROK in Section II Part A, we used Visual Basic to create a Monte Carlo simulation model of an infinitely large, randomly mixed population (as in Chapter 2). We parameterized this base model to the specifications of PMF ROK such that \(~84.4%\) of the population (i.e. \(157/186\)) was assigned to \(n = 0\), \(~8.6%\) of the population (i.e. \(16/186\)) was assigned to \(n = 1\), and so on. 

96
To reflect the 3 intervention scenarios of interest, 3 separate versions of the model (i.e. scenarios) were developed (i.e. NO INT, INT 1, INT 2), each of which featured the following adjustments:

- **No intervention deployed (i.e. NO INT):** This scenario was simulated specifically as a benchmark by which to measure the effectiveness of uniform (i.e. INT 1) and risk-informed (i.e. INT 2) intervention deployment strategies.

- **Uniform intervention deployment (i.e. INT 1):** Between \( t = 0 \) and \( t = 1 \), approximately 10% of all individuals in the population (excluding the index case\(^{13}\)) were randomly deployed an intervention that was 100% effective\(^{14}\), thus re-assigning their \( n \)-values from \( n = n \) to \( n = 0 \). For instance, if an individual was assigned an \( n = 1 \) at the beginning of the outbreak simulation but was then selected to receive the intervention, the \( n \)-value associated with that individual was re-assigned to \( n = 0 \). In this scenario, all individuals in the population had an equal likelihood of receiving the intervention (i.e. 10%) - irrespective of risk.

- **Risk-informed intervention deployment (i.e. INT 2):** Between \( t = 0 \) and \( t = 1 \), approximately 62% of all comorbid individuals in the populations (excluding the index case) were randomly deployed an intervention that was 100% effective, thus

\(^{13}\)This choice reflects decision-making during emerging infectious disease outbreaks, when intervention deployment occurs strictly *after* identification of an index case (as opposed to steady-rate deployment of interventions in response to endemic infectious diseases). In the future, we plan to test how time of deployment (i.e. staggered deployment, later initiation, etc.) impacts the early results presented in this thesis.

\(^{14}\)The percent effectiveness associated with the simulated intervention can easily be adjusted to reflect the effectiveness associated with real-world interventions.
Chapter 4: In-Population Heterogeneity and Deployment of Interventions

re-assigning their \( n \)-values from \( n = n \) to \( n = 0 \). As discussed in Chapter 3, case outcome was the single best predictor of heterogeneous human-to-human transmission of MERS-Coronavirus in South Korea; however, given that deployment of most pharmaceutical and non-pharmaceutical interventions (with the exception of surveillance-related interventions, like contact tracing) occurs prior to a case outcome being known, comorbidity status was used as a proxy indicator\(^{15}\) instead (Majumder et al. 2015). Because ~16.1\% of the population of interest was considered to be comorbid (See Chapter 3), deployment of interventions to 62\% of all comorbid individuals was deemed equivalent to deployment of interventions to 10\% of the population in its entirety\(^{16}\). Thus, in this scenario, individuals at higher risk of spreading the infection were exclusively targeted to receive the intervention; however, the rate of deployment at the population-level was equivalent to the rate of deployment in INT 1, allowing for comparison across both scenarios.

For all 3 versions of the model (i.e. scenarios), index cases were selected at random at the beginning of each outbreak simulation, and a total of 5000 trial runs were simulated for each scenario. Cases were modeled with an infectious period of 1 generation\(^{17}\), and

\(^{15}\)Previous work by Majumder et al. (2015), which used the same data set as the one cited in Chapter 3 of this thesis, found that comorbidity status was the strongest predictor of case outcome during the 2015 South Korean MERS outbreak; thus, it is used as a proxy indicator of case outcome here.

\(^{16}\)Considering that the product of 62\% and 16.1\% is approximately 10\%.

\(^{17}\)As in Chapter 2, we assumed that the infectious period was homogeneous across the population (i.e. 1 generation), even though it is likely that this time interval varies from one individual to the next. When and if further data from the 2015 South Korean MERS outbreak becomes available, we plan to re-configure our model in order to incorporate this added layer of in-population variance.
considering the infinitely large and randomly mixed nature of the modeled population, re-infections were not permitted. Given that the 2015 South Korean MERS outbreak lasted for only 5 generations, each outbreak simulation was programmed to terminate propagation at $t = 5$.

After running 5000 trials for each of the aforementioned scenarios, we plotted average, maximum, and minimum cumulative incidence over time (See Section III). We also reported rates of extinction over time and produced likelihood histograms to compare spread associated with total outbreak size at $t = 5$ for all 3 scenarios (See Section III). Finally, to validate our simulation approach, we compared results from NO INT to those obtained via the analytical approach defined in Section II Part A and to the real-life outbreak trajectory associated with MERS in South Korea (See Section IV).
Chapter 4: In-Population Heterogeneity and Deployment of Interventions

III. Results

Here, we present results obtained via the analytical approach defined in Section II Part A and the simulation approach defined in Section II Part B. We discuss these results further in Section IV.

A. Analytical Approach

Using PMF ROK, we determined that the probability of takeoff by $t = 1$ would be approximately 16% in a no-intervention scenario, as follows:

$$P\left(\text{takeoff by first generation}\right) = 1 - \left(\frac{157}{186}\right) \approx 0.16$$

Similarly, we found that the probability of extinction by $t = 2$ would be approximately 95%:

$$P(\text{extinction by second generation}) = \left(\frac{157}{186}\right) + \left(\frac{16}{186}\right)\left(\frac{157}{186}\right)^1 + \left(\frac{6}{186}\right)\left(\frac{157}{186}\right)^2 + \left(\frac{1}{186}\right)\left(\frac{157}{186}\right)^3 + \left(\frac{1}{186}\right)\left(\frac{157}{186}\right)^4 + \left(\frac{1}{186}\right)\left(\frac{157}{186}\right)^6$$

$$+ \left(\frac{1}{186}\right)\left(\frac{157}{186}\right)^9 + \left(\frac{1}{186}\right)\left(\frac{157}{186}\right)^{23} + \left(\frac{1}{186}\right)\left(\frac{157}{186}\right)^{28} + \left(\frac{1}{186}\right)\left(\frac{157}{186}\right)^{43} \approx 0.95$$

Notably (and as was first theorized in Chapter 2 Appendix), both the probability of takeoff and the probability of propagation beyond $t = 2$ was non-negligible – despite the fact that $R_0 < 1$ for PMF ROK.
Chapter 4: In-Population Heterogeneity and Deployment of Interventions

B. Simulation Approach

Building off of our analytical findings in Part A, our simulation approach allowed us to explore the likelihood of extinction beyond $t = 2$ – not only for the no-intervention scenario (i.e. NO INT) but for the uniform and risk-informed intervention deployment scenarios (i.e. INT 1 and INT 2), too (Table 1).

Table 1. Likelihood of extinction at $t = [1, 5]$ for 3 different simulation models (i.e. NO INT, INT 1, and INT 2) as determined over the course of 5000 trial runs. Note that the likelihood of takeoff is equivalent to $[1 - ($likelihood of extinction at $t = 1$)].

<table>
<thead>
<tr>
<th>$t$</th>
<th>NO INT</th>
<th>INT 1</th>
<th>INT 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.8524</td>
<td>0.8386</td>
<td>0.8452</td>
</tr>
<tr>
<td>2</td>
<td>0.9522</td>
<td>0.9474</td>
<td>0.955</td>
</tr>
<tr>
<td>3</td>
<td>0.974</td>
<td>0.9754</td>
<td>0.983</td>
</tr>
<tr>
<td>4</td>
<td>0.9838</td>
<td>0.986</td>
<td>0.9896</td>
</tr>
<tr>
<td>5</td>
<td>0.988</td>
<td>0.993</td>
<td>0.9936</td>
</tr>
</tbody>
</table>

Because likelihood of early extinction was very high even in the no-intervention scenario, there were only marginal differences over time across all three scenarios. Nevertheless, INT 2 exhibited a slightly higher rate of extinction by $t = 5$ as compared to INT 1.

However, the differential effects associated with each scenario on maximum and average cumulative incidence over time were considerable (Figures 1 and 2). Though the minimum cumulative incidence for all 3 scenarios was 0, INT 2 exhibited both a smaller maximum and smaller average than both INT 1 and NO INT (Figures 1 and 2).
Figure 1. Maximum and minimum cumulative incidence for $t = [0, 5]$ for 3 different intervention scenario simulations (i.e. NO INT, INT 1, INT 2). Each simulation was run 5000 times, and the index case was excluded from all counts.

Figure 2. Average and minimum cumulative incidence for $t = [0, 5]$ for 3 different intervention scenario simulations (i.e. NO INT, INT 1, INT 2). Each simulation was run 5000 times, and the index case was excluded from all counts.
Chapter 4: In-Population Heterogeneity and Deployment of Interventions

Figure 3. Likelihood histograms for outbreak size at $t = 5$ (NO INT, SIM 1, SIM 2). Histograms show data from 5000 trial runs each, and max outbreak size for NO INT, SIM 1, and SIM 2 were $N = 623$, $N = 323$, and $N = 253$, respectively. The index case was excluded from all counts.

As a result, the likelihood histogram for INT 2 – which depicted the horizontal spread associated with total outbreak size at $t = 5$ across 5000 trial runs – exhibited less spread than INT 1, and the no-intervention scenario exhibited the greatest spread of all (Figure 3). Nevertheless, all three histograms were right-tailed, and given the high likelihood of extinction across scenarios, 0 was excluded from the x-axis for all three plots.
IV. Discussion

In this chapter, we used data from the 2015 South Korean MERS outbreak – paired with analytical and simulation methods – to explore the utility of risk-informed intervention deployment in realistically heterogeneous human populations. We compared the effectiveness of such a deployment scheme (i.e. INT 2) against the traditional uniform deployment scheme that is used most commonly today (i.e. INT 1), as well as against a scenario in which no interventions are deployed (i.e. NO INT) whatsoever.

First, we modeled the NO INT scenario using both analytical and simulation methods. We found that estimates for likelihood of outbreak takeoff (i.e. ~16% versus ~15%, respectively) and likelihood of early extinction (i.e. ~95% versus ~95%, respectively) were nearly identical across both approaches. Furthermore, we found that the real-life outbreak trajectory of MERS in South Korea (See Chapter 3) – which yielded a total of 186 cases by \( t = 5 \) – was within the bounds of the simulated outbreaks produced by the NO INT scenario\(^{18}\), where the minimum total case count at \( t = 5 \) was 0 and the maximum at \( t = 5 \) was 693. Both of these findings suggested that the simulation method used to compare the utility of risk-informed intervention deployment versus the utility of uniform intervention deployment.

\(^{18}\)Given that the \( R_0 \) that was used to parameterize this base (i.e. NO INT) simulation model (via PMF ROK) was <1, it is worth noting that an SIR-type model would not have been able to predict a MERS outbreak as big as the one that took place in South Korea. However, an SIR-type model would have predicted an average cumulative incidence curve roughly equivalent to the average cumulative incidence curve produced by the NO INT simulation model in Figure 2. See Appendix I for additional details.
deployment would be sufficiently robust, despite the assumptions outlined in Section II Part B.

Results from our simulated outbreaks across scenarios NO INT, INT 1, and INT 2 suggest that risk-informed intervention deployment would likely result in smaller total outbreak size – both on average and at maximum – than uniform intervention deployment would in non-homogenous populations with heterogeneous transmission risk. Though the average total outbreak size (at $t = 5$) associated with INT 1 was 19% smaller than the NO INT scenario, INT 2 yielded an average total outbreak size that was nearly 25% smaller than the INT 1 scenario (and thus, nearly 40% smaller than the NO INT scenario). Similarly, the maximum total outbreak size (at $t = 5$) associated with INT 1 and INT 2 were 53% and 63% smaller (respectively) than the NO INT scenario. Clearly then, though the uniform intervention deployment scheme that is typically used to combat infectious disease outbreaks today is certainly better than nothing, risk-informed deployment would allow for greater reductions in total outbreak size – especially in limited resource settings where availability of interventions is scarce.

Our simulation models are not without limitations, however. As in Chapter 2, our results are contingent upon the number of trials that we ran per scenario. In Section II Part B and Section III Part B, we chose to run 5000 trials for all 3 intervention scenarios; as a result, low probability events (i.e. $p < 1/5000$) were likely excluded from the potential “futures” (i.e. outbreak trajectories) characterized by the each of the 3 models (e.g. a simulated outbreak via NO INT in which each and every individual infected between $t = 0$ and $t = 5$)
infected 43 other individuals, resulting in a final outbreak size of approximately 150 million). However, analytical approaches similar to those presented in Section II Part A and Section III Part A can be used to assess the likelihood of such low probability events when necessary (Appendix II). It is also worth noting that simulation featuring an increased number of trials may also be an appropriate solution for the assessment of likelihoods associated with low probability events – especially when faced with more complex, analytically intractable PMFs.

In addition to this aforementioned inability to capture low probability events, our simulation models are further limited by the fact that disease propagation in a decidedly unrealistic population is both infinitely large and randomly mixed. Considering that human populations are in fact finite in size and feature non-random patterns of contact (i.e. preferential interaction with family, friends, and colleagues), both of these assumptions are impractical. However, it is worth noting that if additional data on contact structure (i.e. a network of connectivity as opposed to a network of infectivity) had been available and if a clearer definition of the population bounds (i.e. population size as opposed to outbreak size) associated with MERS in South Korea had been known, we could have easily re-configured our model to reflect a finite population with non-random mixing. In the event that such data become available in the future, such adjustments can and will be deployed to more accurately characterize the 2015 South Korean MERS outbreak.

However, by implementing a simulation model that assumed an infinitely large population with random mixing, we were able to increase the generalizability of our initial findings.
beyond MERS in South Korea. By refraining from further specifying the network of connectivity and population size associated with our simulation models, our overarching finding – namely, that risk-informed intervention deployment may be more effective at reducing final outbreak size following the identification of an index case than uniform intervention deployment would be – is likely applicable to any non-homogeneous population in which one or more generalizable predictors of heterogeneous human-to-human transmission for the disease of interest (as identified from previous outbreaks) are known.

As noted in Section I, it is imperative that we be cautious when generalizing predictors of heterogeneous (in-population) human-to-human transmission\(^\text{19}\) across populations due to issues of cross-population variability (i.e. the fact that populations are different from each other in ways that are meaningful within the context of infectious disease transmission [e.g. demographic structure, culture and customs, etc.]). Predictors must be paired with physical mechanisms of operation before generalization, and when possible, predictors should be generalized from populations that are demographically and culturally similar to the population of interest.

With this mind, it is clear that methods that allow us to model non-homogeneous populations (e.g. analytical and simulation methods, as shown here and in Chapter 2) and

\(^{19}\)This is less of a concern for diseases that occur with some regularity in any one given population (e.g. seasonal flu); in such instances, if predictors of heterogeneous human-to-human transmission are determined in Year X for Population A, said findings may be easily generalized to Year Y and Year Z for Population A under the assumption that the population of interest has not changed considerably.
data that allow us to identify predictors of heterogeneous (in-population) human-to-human transmission (e.g. data from the 2015 South Korean MERS outbreak, as shown here and in Chapter 3) are essential for variance-informed investigation and mitigation of outbreaks, for any and all infectious diseases. Unfortunately, most outbreak research remains firmly entrenched in the use of SIR-type models that do not allow for the treatment of $R_0$ as the mean of a random variable; furthermore, access to and availability of data similar to those obtained from the 2015 South Korean MERS outbreak (i.e. data that give us insight into in-population transmission heterogeneity in non-homogeneous populations) is scarce. Thus, in our concluding chapter (See Chapter 5), we propose a number of possible policy recommendations address these inter-linked issues.
Appendix I: SIR Model vs. Simulation Model Cumulative Case Estimates (See Section IV)

Because $R_0 = 0.77$ during the 2015 South Korean MERS outbreak (See Chapter 3), an SIR-type projection model (See Chapter 2) would have failed to predict an outbreak the size of what actually took place (i.e. 186 total cases by $t = 5$). However, the average cumulative incidence curve produced by the NO INT simulation model (See Section III Part B) is roughly equivalent to estimates produced by an SIR-type model, as shown below (Figure S).

![Cumulative incidence over time, comparing SIR model estimates for $R_0 = 0.77$ to average curve produced by NO INT](image)

*Figure S. Cumulative incidence over time, comparing SIR model estimates for $R_0 = 0.77$ to average curve produced by NO INT.*
Appendix II: Likelihood Estimation of a Low Probability Event (See Section IV)

Because the number of trials that are conducted in any given simulation model are constrained by the user, such models – including the ones mentioned in this chapter – are limited in their ability to capture low probability events. However, as noted in the text, the likelihood associated with such low probability events may be easily estimated analytically for simple PMFs.

Let us consider again the low probability event described in Section IV:

"a simulated outbreak via NO INT in which each and every individual infected between $t = 0$ and $t = 5$ infected 43 other individuals"

Recall that the probability associated with $n = 43$ (See PMF ROK) is $p = 1/186$. Thus, the probability of the aforementioned event is as follows:

$$P(event) = \left(\frac{1}{186}\right)\left(\frac{1}{186^{43}}\right)\left(\frac{1}{186^{43^2}}\right)\left(\frac{1}{186^{43^3}}\right)\left(\frac{1}{186^{43^4}}\right) \approx 0$$
Chapter 4: In-Population Heterogeneity and Deployment of Interventions

References


Chapter 5: Conclusions and Policy Recommendations

I. Conclusions

In Chapter 2 of this thesis, we showed that – despite current modeling conventions (i.e. deterministic SIR-type methods\textsuperscript{20}) that mandate otherwise – $R_0$ should be treated as the mean of a random variable when modeling infectious disease outbreaks in non-homogeneous human populations. Thus, we determined that deterministic SIR-type methods should be treated as quick-and-dirty measures that are limited to the estimation of averages when attempting to characterize disease transmission in heterogeneous populations.

Then, in Chapter 3, we explored a real-world example of infectious disease transmission in a heterogeneous human population (i.e. MERS in South Korea). Through this empirical investigation, we discovered that – when given a sufficiently robust data set – it is possible to determine predictors of heterogeneous human-to-human transmission of infectious diseases.

Finally, in Chapter 4, we studied how predictors of heterogeneous (in-population) human-to-human transmission can be used to better inform intervention deployment across

\textsuperscript{20} As illustrated in Chapter 2, SIR-type models should not – in general – be treated as a gold-standard, irrespective of convention. However, such models do have their time and place (i.e. when populations can be assumed to be relatively homogeneous or when data required to parameterize models that do allow for the incorporation of non-zero variance are limited), as is discussed further in Appendix I.
populations, given sufficiently robust physical mechanisms of operation. Using a combination of analytical and simulation methods, we showed that variance-informed intervention deployment (where those who are at higher risk of passing on the infection of interest are preferentially targeted with interventions) is likely more effective than traditional (i.e. uniform) intervention deployment (where all individuals have the same likelihood of receiving an intervention, irrespective of risk) at reducing disease transmission over the course of an outbreak.

Considering these high-level findings, it is evident that methods that allow for the inclusion of variance and data that allow for the identification of disease-specific predictors of heterogeneous human-to-human transmission are critical assets within the realm of outbreak research – both for the purposes of studying outbreaks as they happen, but for the purposes of responding to them, as well. Unfortunately, however:

(1) Current "best practices" in the mathematical modeling community center around the use of SIR-type models, and

(2) Patient-level data (which are imperative for assessing predictors of heterogeneous human-to-human transmission) – even during infectious disease outbreaks – remain challenging to obtain.

Unsurprisingly, these two problem areas are intimately connected. Perhaps the greatest roadblock to rectifying Problem Area (1) is not so much the intractability of alternatives to SIR-type models (e.g. analytical and simulation models, as demonstrated in this thesis), but rather the scarcity of the data types required (i.e. data that allow us to capture population
heterogeneity) for proper parameterization of such models (i.e. Problem Area (2)). Thus, in Section II of this chapter, we start by proposing two policy recommendations (Proposals 1 and 2) that are designed to improve data availability during outbreaks, thus presenting researchers with the opportunity to shift towards modeling methods that allow for the treatment and reporting of $R_0$ as the mean of a random variable (through which non-deterministic case count projections may be estimated). Then, in Section III, we propose two additional policy recommendations (Proposals 3 and 4) that are designed to promote the treatment and reporting of $R_0$ as the mean of a random variable via suggested changes in the language used to report $R_0$ by policy-driving agencies like the CDC and the WHO.

Overall, the aim of Proposals 1–4 is to normalize the inclusion of transmission heterogeneity in the modeling of infectious disease outbreaks. Though we have focused primarily on in-population heterogeneity thus far in this thesis, Proposals 2 and 4 consider cross-population heterogeneity, as well – specifically to address modeling of outbreaks that spread beyond a single population (e.g. Ebola in West Africa). Finally, in Section IV, we conclude with a summary of Proposals 1–4.
II. Data Availability During Infectious Disease Outbreaks

As acknowledged in Section I, methods that aim to consider and describe the effects of non-zero variance on outbreak trajectories require data that can be used to describe in-population heterogeneity (e.g. networks of infectivity, demographic characteristics of cases, etc.). Unfortunately, these data – which are, in their most traditional form, patient-level data sets – are rarely shared in full by hospitals and other such facilities due to concerns regarding patient privacy and data ownership. Thus, in this section, we propose two policy recommendations (Proposals 1 and 2) that aim to address this constraint.

Proposal 1: Open and Semi-Open Sharing of Traditional Data Sources

During infectious disease outbreaks, we propose that hospitals should allow for limited (i.e. adequately anonymized) access to patient-level data, either openly (i.e. publicly-available for download) or semi-openly (i.e. by request). To protect patient privacy, the number of characteristics (i.e. variables) available per individual in the data set can be restricted (as was the case during the 2015 South Korean MERS outbreak), thus reducing the probability of identifiability to the point of negligibility. To ensure that data ownership considerations are sufficiently addressed, open data sets should feature a preferred citation that links directly to the hospital or hospitals that have collected and curated the data; however, in the event of semi-open data sharing, protocols\(^\text{21}\) that ensure that proper credit is granted – especially within the context of low-resource settings – should be instated, as well. Such

\(^{21}\)In collaboration with the Chatham House, additional work by M.S. Majumder has contributed towards formalization of such a protocol (Chatham House 2017).
protocols should include items that are designed to empower researchers who are local to the low-resource setting in question, as well as their institutions. For instance, local researchers who are familiar with the data collection and curation process should be invited to co-author any studies that are published using the shared data.

Proposal 2: Integration of Publicly-Available Non-Traditional Data Sources

Even with improved data sharing protocols, open and semi-open traditional data sources will by default be limited due to patient confidentiality, identifiability, and privacy issues; indeed, it was for this reason that the data set we used to explore MERS in South Korea in Chapter 3 was so limited in the number of variables available for analysis.

Because of this limitation, we propose that non-traditional data sources – such as social media22 – be used to augment our understanding of both in-population and cross-population heterogeneity during infectious disease outbreaks. Given the current prevalence of Internet use around the world, social media and other such non-traditional digital disease data sources provide a compelling, publicly-available means by which researchers can monitor the well-being of both individuals and (more generally) populations (Brownstein et al. 2009; Gardy and Loman 2018). Furthermore, unlike traditional hospital data sources, such data can provide insight into the complete population affected by a given outbreak, as opposed to simply those who get sick.

22Beyond social media, news media and search trend data can also offer utility in capturing the role of in-population and cross-population heterogeneity in infectious disease outbreaks. See Appendix I for further information.
For instance, networks of real-life connectivity can be reconstructed using geo-coded social media data (e.g. Twitter). Data regarding individual demographic characteristics can also be gleaned from social media (e.g. Twitter) profiles (e.g. age, sex, occupation, etc.), which – when paired with wellness data (e.g. tweets about general health, disease-specific symptoms, etc.) – can then be used to determine predictors of heterogeneous (in-population) human-to-human disease transmission, thus enabling risk-informed intervention deployment. Furthermore, Twitter data may also be used to determine whether or not proposed predictors of heterogeneous human-to-human transmission of a given disease (or transmission in general) vary across populations (i.e. cross-population variability) by sorting users by location. This use-case is particularly valuable during outbreaks that are not geographically constrained and span multiple populations with varying demographic and societal (i.e. cultural and infrastructural) structures.

Despite these aforementioned benefits that argue in favor of integrating non-traditional data sources when modeling infectious disease outbreaks, it is important to acknowledge that social media data are not well-regulated, and ethical considerations regarding the use of these data for public health research remain relatively unresolved. Though the data researchers have access to via platforms like Twitter are generally public (e.g. a Twitter user’s public tweets), this does not guarantee that consumers of social media are truly aware that their data are public, nor does it guarantee that they truly agree with the use of their data for outbreak research.
Thus, the premise of consent within the context of using social media data for outbreak research remains contentious. With this in mind, we propose that confirmation of consent – and proof of confirmation – be mandatory for any and all outbreak research conducted using social media data and act as a core tenant of integrating social media data into modeling of transmission heterogeneity in and across populations. Such a policy can be enforced by academic journals (e.g. refusing opportunities for publication for researchers who cannot provide evidence of confirmed consent, similar to the mechanism that is currently in place for informed consent within the context of human subjects research), as well as social media platforms themselves (e.g. serving a dialogue box to consumers to remind them of the status of their data [public or private]).
III. Reporting $R_0$ as the Mean of a Random Variable

If Proposals 1 and 2 are realized, researchers will be empowered with the ability to properly parameterize models that allow for the treatment of $R_0$ the mean of a random variable (e.g. analytical and simulation methods, which produce non-deterministic case count projections). However, beyond availability of data, a cultural shift in the outbreak modeling community that prioritizes the reporting of $R_0$ as the mean of a random variable (as opposed to a constant with zero variance) will also be necessary to incentivize researchers to transition away from SIR-type modeling methods and towards methods that better capture population heterogeneity. Thus, in this section, we recommend two additional proposals that explore the role policy-driving agencies like the CDC and the WHO can play in spear-heading this cultural shift through changing the language that they use to report $R_0$.

Proposal 3: Acknowledging the Underlying Variance Associated with $R_0$ in a Single Population

At present, when agencies like the WHO and the CDC report $R_0$ mean estimates during an outbreak of a specific disease in a specific population, they often do so without also noting the existence of underlying (in-population) variance. Unfortunately, by reporting $R_0$ as a “constant of nature” in this way (i.e. reporting merely a mean estimate), policy-making bodies rob the individuals who comprise the population in question of their agency to change the course of an outbreak through the uptake of both pharmaceutical and non-pharmaceutical interventions (Finkelstein et al. 2015).
As noted by Finkelstein et al. (2015), small individual-level behavior changes (e.g. social distancing) can drastically decrease the transmissibility of viral respiratory diseases such as seasonal influenza. However, when organizations like the WHO and the CDC report point estimates of the $R_0$ associated with flu, they effectively fail to acknowledge the fact that individual-level actions (and thus, in-population variance) can rapidly decelerate (or accelerate) the growth of an outbreak in a given population.

Thus, we propose that policy-driving agencies like the CDC and the WHO should explicitly acknowledge the underlying variance associated with $R_0$ when reporting mean estimates for a given disease in a given population. In addition to empowering individuals to use individual-level behavior change to help mitigate the outbreak in question more rapidly, we believe that compliance with such a proposal would also incentivize the use of modeling methods that allow researchers to actually report measures of in-population variance (e.g. analytical and simulation methods) in addition to acknowledging that said variance exists.

**Proposal 4: Acknowledging Context-Dependence When Reporting $R_0$ Across Populations**

In addition to acknowledging in-population *variance* when reporting $R_0$ estimates for a given disease in a given population, we also propose that policy-driving organizations like the CDC and the WHO should also acknowledge cross-population *variability* when reporting $R_0$ estimates for diseases that occur over multiple populations. As first mentioned in Chapter 1 and as illustrated again here, both the CDC and the WHO currently report $R_0$
estimates for several common infectious diseases without noting the context-dependency of the reported ranges (Table 1) (CDC and WHO 2010).

Table 1. $R_0$ estimates as reported by the WHO and CDC for six well-known diseases and proposed changes.

<table>
<thead>
<tr>
<th>Disease</th>
<th>Reported $R_0$ Estimate</th>
<th>Proposed Changes to Reporting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diphtheria</td>
<td>6–7</td>
<td>$R_0 = 6$ in Population A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$R_0 = 7$ in Population B</td>
</tr>
<tr>
<td>Measles</td>
<td>12–18</td>
<td>$R_0 = 12$ in Population C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>...</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$R_0 = 18$ in Population D</td>
</tr>
<tr>
<td>Mumps</td>
<td>4–7</td>
<td>$R_0 = 4$ in Population E</td>
</tr>
<tr>
<td></td>
<td></td>
<td>...</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$R_0 = 7$ in Population F</td>
</tr>
<tr>
<td>Polio</td>
<td>5–7</td>
<td>$R_0 = 5$ in Population G</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$R_0 = 6$ in Population H</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$R_0 = 7$ in Population I</td>
</tr>
<tr>
<td>Smallpox</td>
<td>5–7</td>
<td>$R_0 = 5$ in Population J</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$R_0 = 6$ in Population K</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$R_0 = 7$ in Population L</td>
</tr>
<tr>
<td>Rubella</td>
<td>5–7</td>
<td>$R_0 = 5$ in Population M</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$R_0 = 6$ in Population N</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$R_0 = 7$ in Population O</td>
</tr>
</tbody>
</table>

Because these diseases are common around the world, it is likely that the $R_0$ ranges cited by the CDC and the WHO in Table 1 include estimates from multiple populations – each with its own demographic and societal (i.e. cultural and infrastructural) structure. The fact that many of the ranges reported – most notably, for measles and mumps – are so wide indicates that cross-population variability (e.g. social prevalence of hugging and kissing in
Population A versus Population B) likely plays an important role in the transmissibility of the diseases listed.

Much in the same way that in-population variance (e.g. the fact that Person A frequently engages in hugging and kissing and Person B does not) can change the dynamics of an outbreak within a single population, cross-population variability can change the dynamics of an outbreak across multiple populations. This is especially pertinent in situations where an outbreak spreads beyond a single population, as was the case during the 2014-2016 West African Ebola outbreak.

As a result, we propose that the population from which any given estimate of $R_0$ is derived should be reported explicitly alongside the estimates themselves (as shown in Table 1 under “Proposed Changes in Reporting”), including as much detail as is feasible (e.g. geographic location [i.e. country or state], age groups affected, year of outbreak, etc.). Furthermore, calling back to Proposal 3, in-population variance should be acknowledged whenever possible (e.g. “Though Population A exhibited a mean $R_0$ of 6 during an outbreak of diphtheria in Year X, it appears that individuals who were vaccinated were much less likely to get sick than those who were not.”)

---

23 As noted in Chapter 1, the WHO and the CDC did in fact report country-specific reproduction numbers during the 2014-2016 West African Ebola outbreak, which suggests that a cultural shift in acknowledging cross-population variability is currently underway.
IV. Summary of Recommendations

In order to capture in-population transmission heterogeneity of infectious diseases, we must employ modeling methods that allow for the treatment and reporting of $R_0$ as the mean of a random variable (e.g. analytical and simulation methods). To parameterize such models, access to traditional hospital data is imperative during infectious disease outbreaks; however, we must design and implement protocols that protect patient privacy and ensure sufficient acknowledgment of those who own the data – especially in low-resource settings (e.g. local researchers and institutions) (Proposal 1). In addition to open and semi-open sharing of hospital data (which are by default limited for the purposes of protecting patient confidentiality and anonymity), ethical integration of social media data (i.e. integration that protects consumers) should be considered as a means of augmenting traditional data sources for both in-population and cross-population modeling of infectious disease outbreaks (Proposal 2).

Though improved data availability via Proposals 1 and 2 is likely essential to a community-wide shift towards modeling methods that explicitly allow for the treatment and reporting of $R_0$ as the mean of a random variable, we also believe that policy-driving agencies like the CDC and the WHO can help create a culture that is more conducive to such a shift by changing the language used to describe $R_0$. More specifically, publications that report estimates of $R_0$ for a given disease in a given population should include an acknowledgment of the role that in-population variance can play in curbing or fueling an outbreak of the disease in question (Proposal 3). Similarly, publications that report estimated ranges of $R_0$
for a given disease across multiple populations should specify the populations from which the estimates were derived and acknowledge the role that cross-population variability can play in the population-level dynamics of the disease in question (Proposal 4).
Appendix I. Other Sources of Non-Traditional Data and Their Uses

As noted in Section III Proposal 4, social media data can lend insight into in-population variance (i.e. the primary focus of this thesis); however, other non-traditional data sources – including news media and search trend data – can help us glean insight into cross-population variability, too. For example, spatial heterogeneity of search interest in disease-related topics across multiple populations can help us predict heterogeneous uptake of interventions, etc. given an outbreak of a single disease in more than one population. Similarly, local news media and their coverage of local sentiment regarding interventions that have been deployed can help us assess the perceived effectiveness of such interventions (and thus, the likelihood of their uptake in adjacent populations).

However, these non-traditional data sources are not limited exclusively to these applications; indeed, perhaps the most common application of these data sources simply involves the simple re-construction of epidemic curves (i.e. incidence over time, which is required to parameterize even the simplest of models, including SIR). This particular application has demonstrated considerable utility in ultra-data-scarce settings, in which traditional data sources (e.g. hospital data) that would typically be used to construct an epidemic curve are unavailable due to diagnostic challenges (as was the case during the 2015-2016 Central and South American Zika epidemic, due to cross-reactivity between ZIKV and other related filoviruses) or due to resource constraints that lead to reporting delays (as was the case during the 2017-2018 Yemeni cholera outbreak) (Majumder et al. 2016; Nishiura et al. 2017). In these instances, SIR-type models (which require significantly
fewer parameters to run than simulation models, for example) are likely sufficient for the characterization of expected transmission dynamics – as long as the limitations of such methods (and the non-traditional data used to re-construct the epidemic curve) are clearly outlined.
Chapter 5: Conclusions and Policy Recommendations

References


