Modeling of Emerging Infectious Diseases for Public Health Decision Support

Caitlin M. Rivers

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Genetics, Bioinformatics and Computational Biology

Bryan L. Lewis, Co-Chair
Stephen G. Eubank, Co-Chair
Kathleen A. Alexander
Jean-Paul Chretien

March 16, 2015
Blacksburg, Virginia

Keywords: epidemiology, zoonoses, emerging infectious diseases, infectious disease modeling
Copyright 2015, Caitlin M. Rivers

Modeling of Emerging Infectious Diseases for Public Health Decision Support

Caitlin M. Rivers

Emerging infectious diseases (EID) pose a serious threat to global public health. Computational epidemiology is a nascent subfield of public health that can provide insight into an outbreak in advance of traditional methodologies. Research in this dissertation will use fuse nontraditional, publicly available data sources with more traditional epidemiological data to build and parameterize models of emerging infectious diseases. These methods will be applied to avian influenza A (H7N9), Middle Eastern Respiratory Syndrome Coronavirus (MERS-CoV), and Ebola virus disease (EVD) outbreaks. This effort will provide quantitative, evidenced-based guidance for policymakers and public health responders to augment public health operations.

Dedication

For Isaac and Amayla.

Acknowledgments

Two of the three chapters of this dissertation are previously published. I would like to acknowledge my co-authors of those previously published works with a brief biography of each.

Stephen G. Eubank is deputy director of the Network Dynamics and Simulation Science Laboratory at Virginia Tech. He is a tenured professor in the Department of Population Health Sciences, and adjunct professor in the Department of Physics. His research is focused on studying diffusive processes on networks. Dr. Eubank oversaw the research direction for each of the three primary chapter (2-4) in this dissertation.

Bryan L. Lewis is a Public Policy Analyst at the Network Dynamics and Simulation Science Laboratory He is a computational epidemiologist interested in transmission dynamics of infectious diseases. As primary mentor, Dr. Lewis was integral to the planning and execution of every chapter in this dissertation.

Eric Lofgren is a postdoctoral fellow at the Network Dynamics and Simulation Science Laboratory. He is interested in healthcare-facility associated infections, and transmission dynamics of respiratory and enteric pathogens. Dr. Lofgren implemented the stochastic forecasting for Ebola virus disease in chapter 4.

Kristian Lum is a statistician, formerly of the Network Dynamics and Simulation Science Laboratory. She is interested in Bayesian population estimation and spatial quantile regression. According to her biography, she enjoys "finding problems and applications for which statistics is unexpectedly useful." Dr. Lum led the sensitivity analysis for H7N9 in chapter 2.

Madhav Marathe is director of the Network Dynamics and Simulation Science Laboratory, and a professor of Computer Science at Virginia Tech. He is interested in interaction-based modeling and simulation of large, complex biological, information, social and technical systems. Dr. Marathe guided the research direction of the Ebola project in chapter 4.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Public health is the cornerstone of modern society. Cities and towns are made possible by sanitation infrastructure, food quality control, and infectious disease management. Without sewer systems and waste management services to remove waste before it piles up in the streets, pests would proliferate and water supplies would be contaminated. Without building codes and commercial hygiene standards, we would be vulnerable to rodent-borne diseases like Plague and hantavirus. Without vaccinations and efficient systems to deliver them, pertussis, mumps, rabies, diphtheria, and more would spread quickly through our densely-populated communities. Smallpox, polio, and tuberculosis are diseases that contemporary developed communities hardly recognize. The successful control of those pathogens are public health achievements that have dramatically improved the foundation on which the rest of our modern lives are built.

As more of modern society has moved from the physical world to the digital, public health has adapted. Although infectious diseases cannot spread from computer to computer or phone to phone, information and strategy can. In the last 20 years, researchers have developed methods to take advantage of the opportunities that newly-connected world provides. Known broadly as computational epidemiology, the new discipline uses computer-based methods and data sources to surveil, detect and respond to events of public health concern [1, 2, 3]. It is a highly multidisciplinary field that draws from traditional public health, computer science, network science, data science, and geographic information science.

As the discipline has matured, several subfields have developed. Digital epidemiology uses data from individuals' 'digital footprint' to track "local and timely health information about disease and health dynamics in populations around the world" [3]. Sources include data generated by social networking sites, mobile phones, web search engines, and online news sites. Digital epidemiology has the advantage of leveraging timely information, like near-real time tweet streams, for surveillance purposes. Studies have shown that surveillance systems using these methods can detect outbreaks up to two weeks in advance of their traditional counterparts, which gives public health professionals an opportunity to implement interventions

to limit morbidity and mortality [4, 5]. Examples of digital epidemiology include Google's use of trends in search term data to track outbreaks of influenza like illness [6], and using Twitter data in a similar manner to track ILI outbreaks [4, 5], vaccine sentiments [7] and risky behaviors associated with the transmission of HIV [8].

Instead of curating 'digital exhaust' from individuals, open epidemiology uses population-level data from public sources, often published online for unrelated purposes. Sometimes that data is information about an outbreak itself, for example situation reports containing case counts, or case reports with information about the people infected. Sometimes more contextual information like population information from a published Census, or geographic distributions of suspected host animal is employed. These data are aggregated, synthesized, simulated or re-purposed to give insight into who is getting infected, where, and why, or they can be used to build and inform models.

As with digital epidemiology, one of the primary objectives of open epidemiology is to close the gap between when an outbreak is first recognized, and when effective public health interventions are designed and deployed. Traditional epidemiological methods rely on data collected through field investigations, surveys, hospital records, etc., which are limited in scope and timeliness. These data are constrained by the manpower available to do the collection, and our prior knowledge about what disease to look for and what those diseases look like. Although still integral to epidemiological investigations and response, manual data collection is too time consuming to be the sole means for understanding an outbreak. Using public data available online is much faster than gathering it on the ground, greatly shortening time to response.

In silico epidemiology, the third nascent subfield of computational epidemiology, combines the individual-level scale of digital epidemiology with the community-level scale of open epidemiology. Rather than gathering existing data, *in silico* epidemiology generates new data from artificial societies built to approximate the real world. These artificial societies are known as agent-based models, because every individual in the population, known collectively as agents, are represented in the model. Agents can transmit infections to one another, resulting in epidemics that are generated organically in a way that closely approximates the real transmission dynamics of an infectious disease. The advantage of *in silico* epidemiology is that researchers have complete information about the epidemic, which is impossible in a real-world setting. Every individual's disease status is known, be it uninfected, asymptomatic, mildly symptomatic, etc. The transmission path that describes how a person came to be infected is also known, so network analyses like identifying demographic characteristics of 'super spreaders' becomes possible. Because *in silico* societies so closely approximates real world dynamics, findings from *in silico* studies can be easily translated to public health applications like planning vaccine stockpiles or school closure policies.

The strength of computational epidemiology is that it extends our knowledge of the etiology, natural history, and dynamics of diseases, people and populations. Traditional epidemiological methods rely on data collected through field investigations, surveys, hospital records,

etc., and are limited in scope and timeliness.

1.1 Contribution of this work

To date, a majority of efforts in all branches of computational epidemiology have focused on well-characterized diseases like influenza-like illness [9, 10, 11, 12, 6, 13, 14]. The advantage of using an established disease are numerous. First, the availability of traditional public health data, e.g. incidence, is critical for validating the novel and often untrusted methods developed by computational epidemiologists. Without a benchmark, there is no way to evaluate the performance of a new method. Second, disease parameters like the incubation and infectious periods are often necessary for parameterizing models or developing new surveillance methods. These values are known and widely accepted for well-characterized diseases. Finally, the impact of diseases like influenza on morbidity and mortality is well studied, so the need for studying them further is widely understood.

For emerging infectious diseases, the data available to build and refine new models are scarce. Outbreaks are chaotic and constantly evolving, so data streams are inherently incomplete and messy. Furthermore, the data are not usually prepared and published online with research purposes in mind, so documentation is often poor and data formats are usually not machine readable. For these reasons, most modeling studies are conducted after an outbreak has completed and a clearer data picture has emerged. Although retrospective modeling studies are useful for informing future outbreak situations, waiting until the outbreak itself is over is a missed opportunity in terms of providing support to public health responders while the crisis is still in progress.

To give an example of the challenges, in the case of the Ebola outbreak in West Africa there are numerous steps data from each patient must make in order to appear in the epidemiological database. First, the patient must either seek medical care or be identified as sick in the community. That patient must then be properly identified as a suspected Ebola case. This requires that the clinic keep proper records of the number of new cases in their care. At periodic intervals, every clinic that receives suspected cases must send their tallies to the outbreak headquarters, who must in turn keep accurate records of the numbers received. Finally, these compiled numbers must be released to the wider audience. Each step is an opportunity for error to be introduced. Cases that do not enter the surveillance system, or are misdiagnosed; clinics that do not report summaries properly; clerical errors are all potential failings in the system.

Despite these challenges, there is an urgent need to extend computational epidemiology methods into the realm of emerging infectious diseases. In recent years, a number of new or re-emerging infectious diseases like Severe Acute Respiratory Syndrome (SARS), influenza H1N1, influenza H7N9, Middle East Respiratory Syndrome Coronavirus (MERS-CoV) have threatened public health worldwide. Although the emergence of novel human pathogens is

itself not new, the strength and speed with which new infections can sweep the globe is only beginning to be understood. Denser cities, a growing population, extensive global trade, a changing global climate, and cheap air travel have altered the global infectious disease dynamics. As the emergence of SARS in 2003 poignantly demonstrated, a brand new disease in China can spread to Toronto (and potentially a hundred other places, simultaneously) in a matter of hours [15, 16]. Detecting and understanding these events in a timely manner is more important now than ever before in the history of public health. The opportunities that computational epidemiology provides to extend the reach of public health prescience.

This work presents three examples of how computational epidemiology can assist public health officials with outbreak detection and response. The case studies presented occurred during the period from 2013-2015: avian influenza A(H7N9), Middle East Respiratory Syndrome Coronavirus, and Ebola virus disease. All are emerging zoonoses with some human to human transmission potential, high case fatality risk, and are serious concerns to regional or global public health. For each outbreak, a model was constructed entirely from publicly available data. The models were built while the outbreak was still in progress, and all were developed with the aim of developing insights useful for outbreak control.

1.2 Overview

1.2.1 Epidemiology modeling

Although computational epidemiology is in its infancy, its close cousin, the modeling of infectious diseases, is well established as a public health tool. The use of modeling in epidemiology dates back to the early 20th century, when Lowell Reed and Wade Hampton Frost developed a series of ordinary differential equations to represent the dynamics of an infectious disease in a fully susceptible population. The equations describe how people move between different states of the disease - susceptible, exposed, infectious, or removed. The compartmental model, as it is known, has spawned dozens of variations for diseases with different states, including exposed, hospitalized, vaccinated, and so on.

In a basic compartmental model, susceptible people are uninfected but able to be infected. Infectious people can pass the disease to others. People in the removed/recovered compartment are no longer infectious, and cannot become reinfected. The equations include a term for the probability of disease transmission for each unit of contact time for any two individuals in the population. The time it takes to move between compartments varies based on the natural history of the disease. An incubation period of 10 days means people will move from the S compartment to the I compartment at a rate of $\frac{1}{10}$.

$$\frac{dS}{dt} = -\beta_I SI \tag{1.1}$$

$$\frac{dI}{dt} = \beta_I SI - \delta I \tag{1.2}$$

$$\frac{dR}{dt} = \delta I \tag{1.3}$$

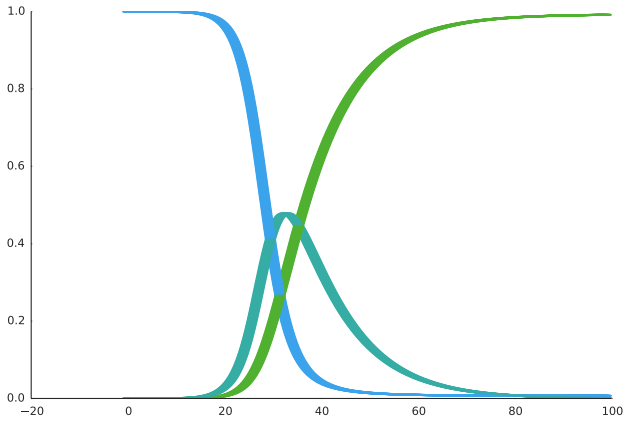

Figure 1.1: **Basic SIR curve dynamics.**

The states are inextricably interdependent. As infectious people cause new infections, the S compartment shrinks. Meanwhile as more and more people become infected, the growth rate of the epidemic accelerates. Eventually the S compartment will become so depleted that new infections dwindle and stop, as seen in Figure 1.1. The growth rate responds by slowing and then stopping. Changes in the S compartment appear as an inverse in the R compartment. Thus the equations describe how people in a population move through each of those states, and how that looks at the population level. These methods are used to simulate outbreaks and produce insights into the underlying natural system.

1.2.2 Models to aid in outbreak response

There are numerous ways to use modeling to improve outbreak response. Perhaps the most straightforward use is to forecast the number of expected new cases of a disease. Models are parameterized to matching existing data; a good model fit will accurately reproduce the historical trajectory of the disease. The model can then be used to look into the short term future to project the number of new cases expected if conditions remain the same. Outbreak responders can use forecasts to anticipate how many people will be needing medical care, so facilities, supplies and staff can be arranged accordingly. Forecasts can also be used as a

baseline to evaluate the current trajectory of the epidemic. Incidences that are significantly above or below the forecasted values may indicate that epidemic is changing.

In general, forecasts are only reliable in the short-term future. They describe the future if conditions remain the same, which is unrealistic over months or years. Epidemics are complex, dynamic processes. Although the biology of the pathogen, the host and the environment are fairly constant over the duration of a typical outbreak, human behavior is not. And it is human behavior that ultimately facilitates the transmission of the disease from one person to another. Changes in contact patterns, care-seeking behavior, surveillance, and treatment can all dramatically impact the course of an outbreak. Although changes in those processes can be represented in the models, it is very difficult to anticipate what exactly those changes will be, and when they will occur. Thus long-term forecasts have very high levels of uncertainty, and are generally considered unreliable.

Beyond forecasts, models can be used as platforms for asking what-if questions. For example, they can be used to test the effect of hypothetical interventions on the epidemic at the population level. Intervention assessments using models are extremely valuable, since experiments on humans are often unfeasible or unethical. Sometimes model experiments are the only way to explore public health scenarios. It is possible to measure the relative effects of social distancing compared to increased case-finding, for example, or the compound effects of implementing both. Models can also assess whether increasing adherence to a treatment regime would ultimately have upstream effects. Improved treatment adherence would clearly improve outcomes of the individual, but it may also reduce the number of new cases by shortening the duration of infection, or moderating the infectivity of an active case. These insights can help responders to understand where to direct their efforts for maximum effect given certain constraints.

Another what-if question asks what would happen if entirely new circumstances arise, for example if an infectious disease is introduced into a new population. Such a threat assessment would have been useful to anticipate the explosive spread of Ebola in the urban regions of West Africa. Ebola has previously been identified only in small, rural regions of central Africa, so its aggressive spread in a new location was unexpected. When it did emerge in a new place, few recognized how different the epidemiology would be in this new place. Modeling could have helped to identify that possibility earlier.

Models can also uncover features of a disease or system that are not directly observable, like detection biases. Age-structured models, for example, can represent the expected age distribution of a disease. If the observed age distribution deviates significantly from the expected, there may be either a detection bias in the surveillance system, or a bias in who gets infected. Both are insights useful for field epidemiologists attempting to control the disease. Similarly, models can identify steady state conditions where outbreaks can become endemic. This scenario occurs when the average number of secondary infections per case is around one. Identifying scenarios with endemic potential can help epidemiologists to take steps to prevent this from happening.

1.3 Bibliography

[1] Christopher L Barrett, Stephen Eubank, and Madhav V Marathe. An Interaction-Based Approach to Computational Epidemiology. pages 1590–1593, 2008.

[2] Bryan Lewis, Stephen Eubank, Chris Barrett, Madhav Marathe, and Keith Bissett. Simulation-Assisted Evaluation and Training of Public Health Decision Makers.

[3] Marcel Salathé, Linus Bengtsson, Todd J. Bodnar, Devon D. Brewer, John S. Brownstein, Caroline Buckee, Ellsworth M. Campbell, Ciro Cattuto, Shashank Khandelwal, Patricia L. Mabry, and Alessandro Vespignani. Digital Epidemiology. *PLoS Computational Biology*, 8(7):e1002616, July 2012. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1002616. URL `http://dx.plos.org/10.1371/journal.pcbi.1002616`.

[4] Harshavardhan Achrekar, Ross Lazarus, and West Cummings Park. Predicting Flu Trends using Twitter Data. In *The First International Workshop on Cyber-Physical Networking Systems*, pages 713–718, 2011. ISBN 9781424499205.

[5] Aron Culotta. Towards detecting influenza epidemics by analyzing Twitter messages. In *Proceedings of the First Workshop on Social Media Analytics*, pages 115–122, New York, New York, USA, 2010. ACM Press. ISBN 9781450302173. doi: 10.1145/1964858.1964874. URL `http://portal.acm.org/citation.cfm?doid=1964858.1964874`.

[6] Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–4, February 2009. ISSN 1476-4687. doi: 10.1038/nature07634. URL `http://www.ncbi.nlm.nih.gov/pubmed/19020500`.

[7] Marcel Salathé, Duy Q Vu, Shashank Khandelwal, and David R Hunter. The dynamics of health behavior sentiments on a large online social network. *EPJ Data Science*, 2(1): 4, April 2013. ISSN 2193-1127. doi: 10.1140/epjds16. URL `http://link.springer.com/10.1140/epjds16`.

[8] Sean D Young, Caitlin Rivers, and Bryan Lewis. Methods of using real-time social media technologies for detection and remote monitoring of HIV outcomes. *Preventive medicine*, 63:112–5, June 2014. ISSN 1096-0260. doi: 10.1016/j.ypmed.2014.01.024. URL `http://www.ncbi.nlm.nih.gov/pubmed/24513169`.

[9] Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. Twitter Catches The Flu : Detecting Influenza Epidemics using Twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1576, Edinburgh, Scotland, UK, 2011.

[10] Aron Culotta. Detecting influenza outbreaks by analyzing Twitter messages. In *KDD Workshop on Social Media Analytics*, number May, pages 1–11, 2010.

[11] Johannes Dreesman and Kerstin Denecke. A New Age of Public Health : Identifying Disease Outbreaks by Analyzing Tweets. *Public Health*.

[12] Nicholas Generous, Geoffrey Fairchild, Alina Deshpande, Sara Y Del Valle, and Reid Priedhorsky. Detecting epidemics using Wikipedia article views : A demonstration of feasibility with language as location proxy. *arXiv*, 2014.

[13] Elaine O Nsoesie and Bryan L Lewis. Agent-based Models for Influenza : Are they all the same? *In preparation.*

[14] Alessio Signorini, Alberto Maria Segre, and Philip M Polgreen. The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic. *PloS One*, 6(5):e19467, January 2011. ISSN 1932-6203. doi: 10.1371/journal.pone.0019467. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3087759\&tool=pmcentrez\&rendertype=abstract`.

[15] G. Chowell, P.W. Fenimore, M.a. Castillo-Garsow, and C. Castillo-Chavez. SARS outbreaks in Ontario, Hong Kong and Singapore: the role of diagnosis and isolation as a control mechanism. *Journal of Theoretical Biology*, 224(1):1–8, September 2003. ISSN 00225193. doi: 10.1016/S0022-5193(03)00228-5. URL `http://linkinghub.elsevier.com/retrieve/pii/S0022519303002285`.

[16] J O Lloyd-Smith, S J Schreiber, P E Kopp, and W M Getz. Superspreading and the effect of individual variation on disease emergence. *Nature*, 438(7066):355–9, November 2005. ISSN 1476-4687. doi: 10.1038/nature04153. URL `http://www.ncbi.nlm.nih.gov/pubmed/16292310`.

Chapter 2

Estimating Human Cases of Avian Influenza A(H7N9) from Poultry Exposure

Manuscript authors: Caitlin M. Rivers, Kristian Lum, Bryan L. Lewis, Stephen Eubank

Citation: Rivers, C., Lum, K., Lewis, B., & Eubank, S. (2013). Estimating Human Cases of Avian Influenza A (H7N9) from Poultry Exposure. PLoS Currents Outbreaks. doi:10.1371/currents.outbreaks.264e737b489bef383fbcbaba60daf928.

2.1 Forward

Although influenza is often perceived as a mild disease, it is a recurring public health threat. Most healthy adults experience a moderate disease, necessitating a few days at home to recover. Symptoms generally include fatigue, fever, body aches, and malaise. But for elderly populations, young children, pregnant women and immunocompromised people, influenza can be deadly. Every year, influenza causes an estimated 25 million cases and 35,000 deaths in the United States [1, 2]. Most of that activity occurs from December to March, which is peak season in the Northern hemisphere.

Sometimes something much worse than a regular flu season appears. Known as pandemic influenza, these aggressive strains can have attack rates as high as 20-40% [3]. Unlike seasonal influenza, pandemic influenza can cause severe disease in young and otherwise healthy people [4]. The 1918 H1N1 pandemic, for example, was one of the deadliest plagues in his-

tory. It killed an estimated 20-50 million people, more than the world war that was raging at the time [5, 3]. Pandemic influenza also circulated in 1958, 1968 and 2009, reminders of the virus's disastrous potential [6].

In the spring of 2013, a strain of influenza not previously seen in humans was identified in China. The global public health community was concerned about avian influenza A(H7N9) from the beginning. Because the pathogen was new to humans, the population was fully susceptible. H7N9 also had an unusually high case fatality risk; early estimates were around 20-50% [7, 8, 9]. The human to human potential of the disease was initially unknown. These risk factors for pandemic potential necessitated a swift and agile public health response.

Scientists quickly determined that H7N9 was likely of avian origin [10]. Settings where humans interacted with birds were likely opportunities for animal to human transmission. In China, many people do their grocery shopping at outdoor markets where live poultry are sold, among other things. These live poultry markets (LPM), or wet markets as they are also known, were identified as potential places of exposure. In April 2013, very soon after the first cases of H7N9 were diagnosed, Chinese public health officials closed the LPMs to minimize this exposure risk [11, 8]. The number of new cases quickly declined, providing support for that hypothesis, which was later confirmed by various studies [8, 12, 13, 14].

Despite this tidy sequence, not all of the pieces fit. A disproportionate number of H7N9 cases were reported to be occurring in elderly men - not the demographic thought to be frequenting the grocery markets most often. If live poultry market attendance were the sole risk factor for contracting H7N9, adult women would have been more heavily affected. We used this observation to generate the hypothesis that additional risk factors were present. Because risk factors must be known before effective public health interventions can be designed, identifying relevant risks was crucial.

In traditional outbreak investigations, the local public health department maintains a line listing by interviewing each patient. Each record is updated periodically with the patient's current status through followup visits or contact with medical care providers. Although this process is the cornerstone of epidemiology, it is very time consuming and requires significant manpower. Epidemiologists must spend time traveling to visit the patients, conducting the interview, and entering, cleaning, and maintaining the data before analyses can be conducted. For researchers outside the network of the health department, the data do not become available until much later, if at all.

During the emergence of H7N9, the data pipeline developed differently. It was one of the first outbreaks for which media reports were timely enough and detailed enough to build a line listing. Media reports on new cases included information about the patient's age, sex, location, dates of onset, and other relevant information. We took advantage of this development to maintain a comprehensive line listing that included a variety of demographic, clinical, and epidemiological information. Instead of waiting for data to be collected on the ground by field investigators, we fused the media-derived line listing with other publicly-available data sources to learn about the outbreak as it unfolded. We aggregated data

from the Chinese Census, and time use surveys to estimate the amount of time various demographic groups spent at markets where live birds were sold.

The line listing analysis and exposure assessment supported the hypothesis that although men age 65+ were at much higher risk for the disease than other demographic groups, they spent less time at the live poultry market - women ages 20-35 and 36-55 were primarily responsible for conducting the familys shopping. This evidence suggests that exposure to live birds is not the sole determinant in contracting H7N9. There is likely an additional immunological factor that makes older men more susceptible to the disease.

This work was published on May 9, 2013 in *PLoS Currents: Outbreaks*, a peer-reviewed journal for rapid communications relevant to infectious disease outbreaks.

2.2 Abstract

In March 2013 an outbreak of avian influenza A(H7N9) was first recognized in China. To date there have been 130 cases in human, 47% of which are in men over the age of 55. The influenza strain is a novel subtype not seen before in humans; little is known about zoonotic transmission of the virus, but it is hypothesized that contact with poultry in live bird markets may be a source of exposure. The purpose of this study is to estimate the transmissibility of the virus from poultry to humans by estimating the amount of time shoppers, farmers, and live bird market retailers spend exposed to poultry each day. Results suggest that increased risk among older men is not due to greater exposure time at live bird markets.

2.3 Introduction

On April 1 the first cases of avian influenza A(H7N9) in Chinese patients who became severely ill with an influenza-like illness were reported to the WHO [15]. Chinese health officials determined that a novel influenza was the source of the illness. Additional cases were soon diagnosed in other regions, including Beijing which is geographically distant. To date there have been 130 cases in eight provinces and two municipalities in China, and new cases continue to be diagnosed [7]. The novel influenza has not yet been found in any other countries, with the exception of a case imported to Taiwan from mainland China [16].

Virologists determined that the virus was an avian subtype, and that poultry were a possible reservoir [17]. Since many of the first human infections were found in people that had been exposed to poultry, epidemiologists hypothesized that the virus was transmitted to humans after close contact with birds [16, 9]. On April 5, 2013, Shanghai authorities ordered the closing of the citys live bird markets, which are food markets where live poultry are slaughtered and sold [11]. These markets were suspected sources of exposure for the novel

influenza, and their closing was meant to minimize human exposure to infected birds. The WHO notes that poultry as a source of exposure is still unconfirmed [15].

Among the 130 confirmed cases, 88 (68%) are men, 37 (28%) are women, and 5 (4%) are children. The mean age of the infected is 62 for men, and 61 for women; the median ages are 60.5 and 58, respectively. The age distribution of the influenza A(H7N9) trends much older than previous outbreaks; influenza A(H5N1) affected primarily people ages 20-30, and did not affect men more than women [15, 18]. Thirty-one (24%) of the infected have died; of the 24 deaths for whom demographic data are available, 18 are men (75%). It is still unknown why men are more affected than women, but the WHO suggests that patterns of exposure to live bird market may put certain demographic groups into more contact with the virus.

2.4 Methods

All source data and analyses are available on Figshare [19].

A line listing of the 130 confirmed cases was developed using data from a variety of sources, including the World Health Organization (WHO), HealthMap, and Flutrackers.com [9, 20, 21]. Data on the patients demographics, course of the disease, and possible exposures were aggregated from media reports and public health organization updates to form as complete of a case record as possible. The patients age, sex, and date of illness onset were used as unique identifiers to prevent duplication and data errors. Only cases with a reported onset date between March 13, 2013 and April 11, 2013 (7 days after the live bird markets were closed, assumed to be the incubation period) were used in the analysis. Cases with no available onset date that were announced before April 22, 2013 were included. Cases found in provinces other than Shanghai, Anhui, Zhejiang and Jiangsu, where a majority of cases were found during the analysis period, were also excluded, as were children under the age of 15. The final case count used in analysis is 83.

Data for estimating exposure time to poultry were collected from the National Bureau of Statistics of China [22]. Where possible, indicators were limited to the four analysis provinces and municipalities (Shanghai, Anhui, Zhejiang, Jiangsu). The population for each of the four regions was estimated from a 2005 survey of 1% of Chinese residents, stratified by age group and sex. An estimate of the number of people who visit a live bird market each day was derived from a 2008 time use survey of the proportion of the population that participated in the purchase of goods and services, also stratified by age group and sex [23]. Data were not available for people over the age of 75, so values from the 65-74 age group were applied.

Population counts and the proportion that shop each day were multiplied to estimate the number of people who shop daily. Various sources estimate that about 80% of Chinese residents shop at a market where live poultry are more likely to be present, rather than a supermarket [24, 25]. The reported analysis uses a uniform distribution of 80% live bird market attendance.

The time use survey also provided the number of minutes each day people spend shopping for goods and services. A native Chinese colleague estimated that a trip to the wet market takes approximately 15-30 minutes, which is about half what the time use survey estimates is spent on purchasing goods and services. The time use values were therefore halved to estimate the number of minutes that are spent specifically at the live bird market, rather than shopping for other things.

The number of exposure days was determined to be 30, derived from seven days (assumed to be the incubation period) before the outbreak began to take hold on March 13 until the markets closed on April 5. The number of people exposed at a live bird market each day was multiplied by the number of minutes spent shopping and by the number of exposure days in order to estimate the group exposure hours per day.

A similar procedure was followed to obtain the occupational exposure rate for agriculture workers and live bird market retailers in each of the four affected regions. The number of men and women in "farming, forestry, animal husbandry and fishing" and "wholesale and retail trade" occupations were taken from 2006 labor statistics provided by the National Bureau of Statistics of China [26]. There are an estimated 9.48 million poultry farming jobs in China, which is 15% of the number of workers in the occupational category as a whole. This multiplier was used to narrow down the number of occupational workers exposed to poultry specifically. Daily exposure minutes were derived from labor statistics on weekly working hours for both occupations. The estimated population exposed was then multiplied by the number of exposure minutes per day and the 30 exposure days to derive occupational group exposure hours.

For each case in the line listing, an exposure category was determined where possible. Seven cases were determined to be agricultural workers, and another four were retail (including food preparation) workers. The remaining 72 cases had no information about possible sources of exposure; these were presumed to be transient exposures, and were therefore categorized as shoppers. This estimate is supported by a Morbidity and Mortality Weekly Report that found 77% of confirmed H7N9 cases were exposed to live animals, "primarily chickens (76%) and ducks (20%)" [7]. An infection rate per hour of contact was estimated by dividing the number of cases for each demographic and exposure category by the group exposure hours.

The rate of infection per exposure hour for demographic group i is given by the equation:

$$rate_i = \frac{cases_i}{population_i * proportionwhovisitLBM_i * exposuretime_i}$$

2.5 Results

Men ages 55+ are disproportionately affected by avian influenza A(H7N9). Despite having a lower estimated exposure time to live bird markets, older men have a much higher infection

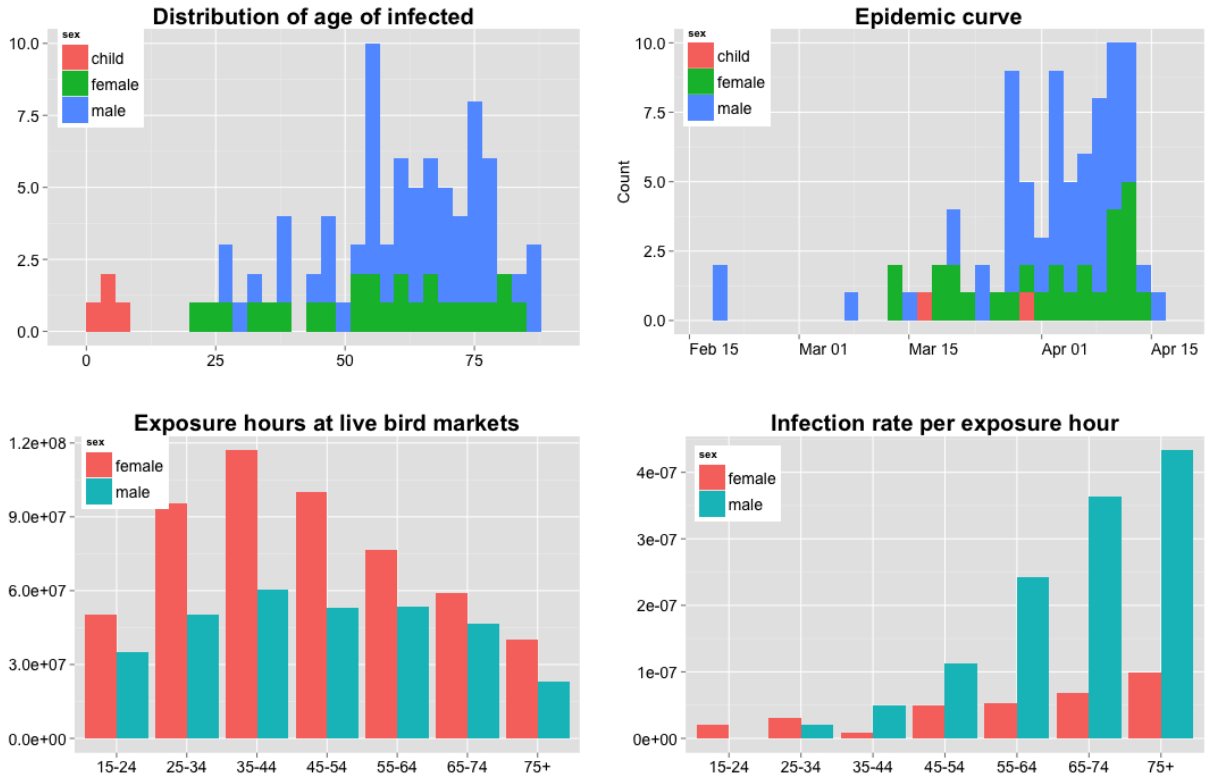

Figure 2.1: **Human infection rate per exposure hour to poultry**. Men ages 55+ are disproportionately affected by avian influenza A(H7N9). Despite having a lower estimated exposure time to live bird markets, older men have a much higher infection rate per exposure hour than other demographic groups.

rate per exposure hour than other demographic groups.

Among shoppers, estimated infection rates per hour increase with age, particularly among men. For example, the infection rate for men ages 65-74 is over five times as high as it is for women of the same age, and over eighteen times as high as for men ages 25-34 (see figure 1, bottom right panel). Shoppers under the age of 35 have very few infections ($n = 5$ adults, $n = 4$ children), despite having in some cases greater exposure hours than older demographics. Although the time use survey shows that older people spend more time shopping than younger people, there are fewer older people in the population overall. Among both men and women shoppers, the group exposure hours (the population at risk multiplied by the time spent exposed) are highest for those ages 35-44 (see bottom left panel of Figure 2.1).

Infection rates among occupationally-exposed people were estimated to be lower than among shoppers, likely because the amount of time spent in contact with poultry was overestimated. No data were available to refine these estimations. Nonetheless, the pattern remains the same

in that occupationally-exposed men have a higher infection rate than occupationally-exposed women.

A sensitivity analysis shows that these conclusions are quite robust. As noted, an estimated 80% of Chinese residents attend live bird markets [24, 25]. However, attendance rates within age and sex groups are unknown. To investigate whether it is possible that the lower number of cases for younger people is due to lower live bird market attendance rather than a decreased transmission rate for those age groups, we decreased bird market attendance rates for younger people to the extent possible while maintaining a marginal attendance rate of 80%. Attendance rates under this scenario are shown in the middle panel of Figure 2.3. Even under these extreme assumptions, the infection rate per exposure hour still strongly favors older men (see bottom panel of Figure 2.3).

In order for the infection rate for all age and sex groups to have been the same as that of the highest (assuming 100% live bird market attendance for the highest group), attendance rates for all other groups would have to be between 2-30%, with an average population-wide attendance rate of about 18%. This is inconsistent with the 80% figure reported. These findings are also supported by a study of the annual number of poultry exposures from live bird market visits and backyard poultry in four areas of China [18]. The poultry exposure data reported by Cowling et al yield very similar infection rates per exposure hours as is reported here. That study reported on the number of annual exposures to poultry in live bird markets and backyards in four regions of China. The survey was conducted in regions not reported on here, and findings were not stratified by sex. The provinces with the highest and lowest exposures were chosen to represent upper and lower bounds. Exposure values were scaled to match the time period of 30 days and applied to both male and female exposure categories. .

The present methodology can also be used for hypothesis-generating analyses. If we suppose that the infection rate per exposure hour is in fact uniform across all groups, and that men ages 75+ (for whom attack rates are the highest) have perfect case detection in which all infections are identified, we can estimate the percentage of undetected cases in other demographics. Under those assumptions, women ages 35-44 would have the lowest detection rate at around 2%, meaning that 98% of cases are not detected (see Figure 2.4). These results suggest that biases in case detection are not the sole cause of the age and sex distribution seen in this outbreak.

2.6 Conclusions

We have shown that the infection rates per exposure hour for avian influenza A(H7N9) are likely much higher among older populations, particularly men. It is implausible that these discrepancies are due to differences in the rates of market attendance or systematic under-reporting. These findings suggest that the age distribution of the outbreak is due

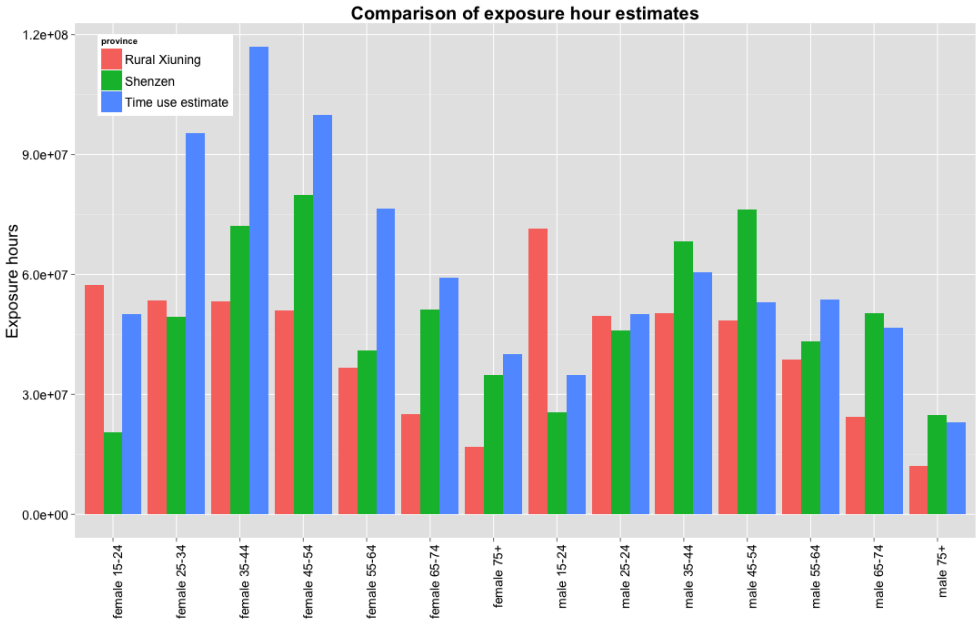

Figure 2.2: **Comparison of exposure estimates reported by Cowling et al.**

to an as-yet unknown epidemiological or immunological feature, and is not due to greater exposure to poultry among the older demographics. It should be noted that for many of the demographic groups, the case counts are quite low, resulting in instability in the rate calculations. It should also be noted that assumptions about the live bird market attendance and duration, and the incubation period, may affect results. That said, the overall conclusions have proven robust to changes in assumptions. Clearly, additional data is needed to confirm the existence of the striking differences in transmission rate by demographic group. Even without extensive epidemiological data, this analysis provides timely evidence that other factors may be contributing to the differential case detection of older men than simply market exposure, as has been hypothesized by the WHO and others.

2.7 Subsequent literature

Following the publication of this paper, additional research was published that provided support for our findings. A survey of almost 5,000 Chinese residents published 15 months after our study found that for some locations like Guangzhou and Shanghai, middle aged women were more likely to visit live poultry markets than elderly men, the high-risk H7N9 demographic [12]. However, LPM attendance by men and elderly men in particular was more prevalent across that survey than our preliminary results had suggested. However, the authors conclude that "our findings suggest that the higher risk for laboratory-confirmed

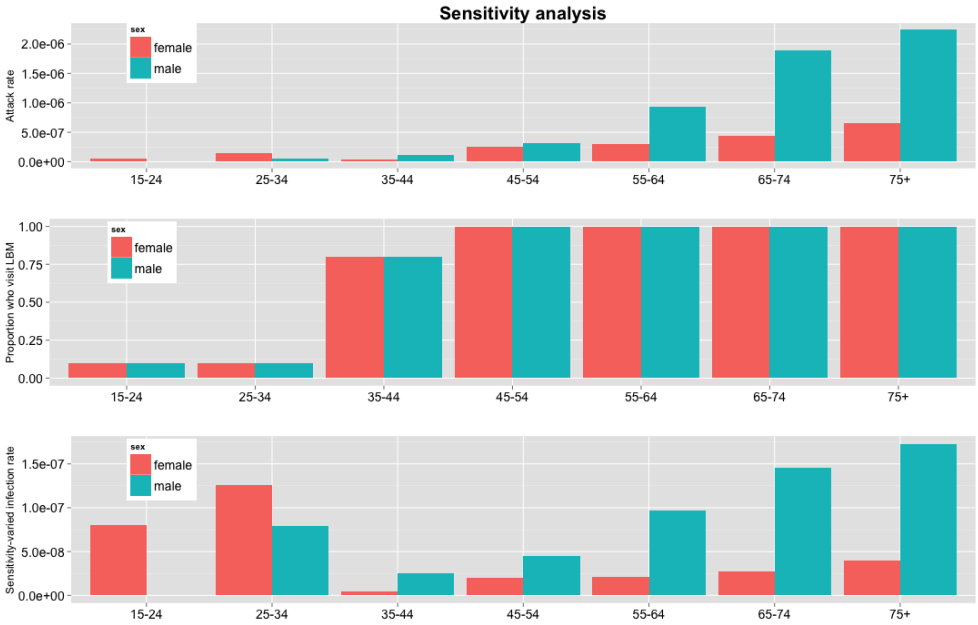

Figure 2.3: **Sensitivity analysis.** The hypothetical number of undetected cases can be estimated assuming that the infection rate per exposure hour is constant, using men ages 75+ as a reference group.

influenza A(H7N9) virus infection among men during the spring 2013 outbreak in the Yangtze River Delta might not be explained by sex differences in exposure but rather by increased susceptibility to serious disease after infection among men." A case-control study conducted in the spring of 2013 and published in June 2014 concluded that men were not significantly more likely to visit an LBM that women, providing further evidence that LBM attendance was not sufficient to explain the sex differences in infection [14].

A case-control study of risk factors for contracting H7N9 conducted in China identified chronic medical conditions; direct contact with poultry; and environmental-related exposures as major risk factors for contracting the disease [13]. The authors note that "a high proportion of elderly patients with severe influenza A(H7N9) virus infection may be due to decreased immune function caused by underlying chronic disease" rather than exclusively a function of greater exposure to live poultry. This directly supports our inference that live poultry market exposure is not the sole risk factor for infection.

Further support comes from research into the host immune response to infection. In a study comparing human H7N9, with H9N3, H5N1 and duck H7N9, the authors found that human H7N9 evokes a pro inflammatory cytokine response in healthy adult mice [27]. Additional work on H7N9 in a mouse model found that hemagglutinin distribution did not vary between young and middle aged mice, which suggests that viral replication in the host lung tissue is not responsible for the pulmonary damage suffered by patients with severe H7N9 infec-

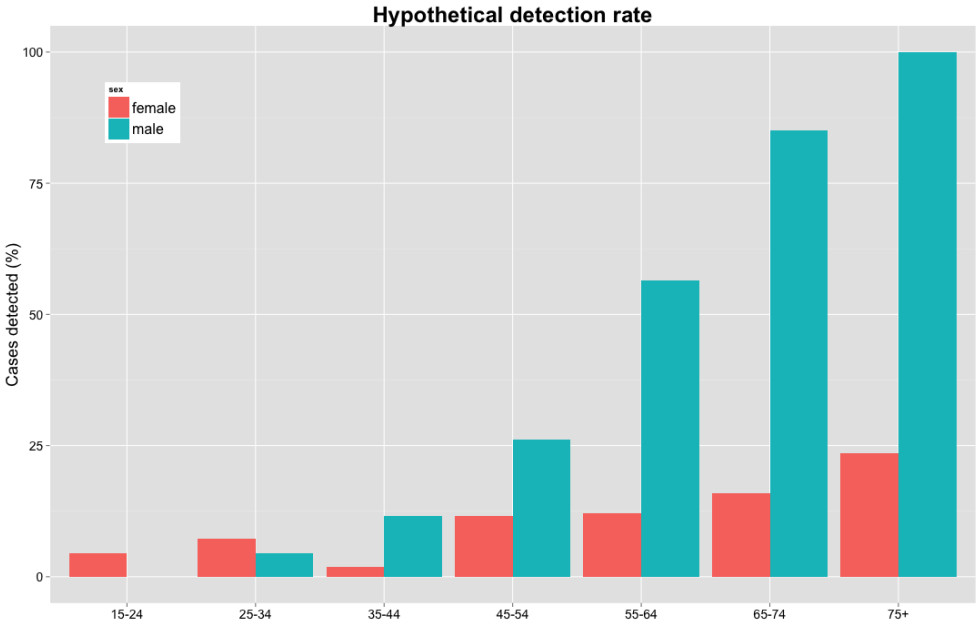

Figure 2.4: **Hypothetical detection rate using men ages 75+ as the reference group.** The hypothetical number of undetected cases can be estimated assuming that the infection rate per exposure hour is constant, using men ages 75+ as a reference group.

tion [28]. These findings suggest that dysregulated proinflammatory immune response is the underlying driver behind the age distribution of severe and fatal cases. Genetic susceptibility may provide an alternate, or perhaps contributing, explanation for disparate outcomes. A study by Wang et al found that rs12252-C genotype was associated with excessive cytokine response [29].

In addition to various studies that support the findings of our paper, research into the robustness of the methodologies has emerged. Cowling et al compared the line listings maintained by five different groups, including ours, with the official Chinese CDC line listing [30]. The study compared age, sex, geographic location, health status on admission, dates of illness onset, hospital admission, and last known health status (e.g. death or discharge) of the publicly-sourced line listings with the official list. Overall conclusions from the study were that line listings from publicly available sources are able to produce "epidemic curves in different regions, estimated onset-to-admission distributions, onset-to-death distributions and impact of poultry market closure can very closely match the results from official data sources with little time lag." However, the authors note that certain information like when a case has recovered is less interesting for the media to report, and is therefore difficult to track using this method.

In April 2013, Chinese health officials moved to close live bird markets in order to limit exposure to H7N9. This intervention is credited with halting the outbreak - closures are

attributed to a 99% reduction in incident cases [8]. The dwindling of the outbreak and ultimately limited number of cases curbed the epidemiological analyses that could provide further insight into the effectiveness of our methodologies. However, the studies reviewed above provide important evidence suggesting that data fusion and other digital epidemiology methods may provide valuable clues in the early days of an outbreak, in advance of traditional field-based approaches.

2.8 Bibliography

[1] Estimates of Deaths Associated with Seasonal Influenza United States , 1976 2007. *Morbidity and Mortality Weekly Report*, 59(33), 2010.

[2] Noelle-Angelique M Molinari, Ismael R Ortega-Sanchez, Mark L Messonnier, William W Thompson, Pascale M Wortley, Eric Weintraub, and Carolyn B Bridges. The annual impact of seasonal influenza in the US: measuring disease burden and costs. *Vaccine*, 25(27):5086–96, June 2007. ISSN 0264-410X. doi: 10.1016/j.vaccine.2007.03.046. URL `http://www.ncbi.nlm.nih.gov/pubmed/17544181`.

[3] Jeffery K Taubenberger and David M Morens. 1918 Influenza: the mother of all pandemics. *Emerging Infectious Diseases*, 12(1):15–22, January 2006. ISSN 1080-6040. doi: 10.3201/eid1201.050979. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3291398\&tool=pmcentrez\&rendertype=abstract`.

[4] Stacey L Knobler, Alison Mack, Adel Mahmoud, and Stanley M Lemon. The Threat of Pandemic Influenza: Are We Ready? Workshop Summary. Technical report, National Academy of Sciences, Washington, D.C., 2005.

[5] Niall P. a. S. Johnson and Juergen Mueller. Updating the Accounts: Global Mortality of the 1918-1920 "Spanish" Influenza Pandemic. *Bulletin of the History of Medicine*, 76(1):105–115, 2002. ISSN 1086-3176. doi: 10.1353/bhm.2002.0022. URL `http://muse.jhu.edu/content/crossref/journals/bulletin\_of\_the\_history\_of\_medicine/v076/76.1johnson.html`.

[6] Flu.gov. Pandemic Flu History.

[7] Emergence of Avian Influenza A (H7N9) Virus Causing Severe Human Illness - China, February-April 2013. *Morbidity and Mortality Weekly Report*, 62(April), 2013.

[8] Hongjie Yu, Benjamin J Cowling, Luzhao Feng, Eric Hy Lau, Qiaohong Liao, Tim K Tsang, Zhibin Peng, Peng Wu, Fengfeng Liu, Vicky J Fang, Honglong Zhang, Ming Li, Lingjia Zeng, Zhen Xu, Zhongjie Li, Huiming Luo, Qun Li, Zijian Feng, Bin Cao, Weizhong Yang, Joseph T Wu, Yu Wang, and Gabriel M Leung. Human infection with avian influenza A H7N9 virus: an assessment of clinical severity. *The Lancet*,

6736(13):1–8, June 2013. ISSN 01406736. doi: 10.1016/S0140-6736(13)61207-6. URL http://linkinghub.elsevier.com/retrieve/pii/S0140673613612076.

[9] World Health Organization. Human infection with influenza A (H7N9) virus in China, 2013. URL http://www.who.int/csr/don/2013_04_01/en/.

[10] World Health Organization. Public health relevant virological features of Influenza A(H7N9) causing human infection in China. 2013.

[11] Shanghai closes poultry markets over flu, April 2013. URL http://www.upi.com/Health_News/2013/04/05/Shanghai-closes-poultry-markets-over-flu/UPI-68381365159114/.

[12] Influenza A Hn, Liping Wang, Benjamin J Cowling, Peng Wu, Jianxing Yu, Fu Li, Lingjia Zeng, Joseph T Wu, Zhongjie Li, Gabriel M Leung, and Hongjie Yu. Human Exposure to Live Poultry and Psychological and Behavioral Responses to. *Emerging Infectious Diseases*, 20(8), 2014.

[13] J Ai, Y Huang, K Xu, D Ren, X Qi, H Ji, A Ge, Q Dai, J Li, C Bao, F Tang, G Shi, T Shen, Y Zhu, M Zhou, and H Wang. Case-control study of risk factors for human infection with influenza A(H7N9) virus in Jiangsu Province, China, 2013. *Eurosurveillance*, 18(26), January 2013. ISSN 1560-7917. URL http://www.ncbi.nlm.nih.gov/pubmed/23827526.

[14] Bo Liu, Fiona Havers, Enfu Chen, Zhengan Yuan, Hui Yuan, Jianming Ou, Kai Kang, Kaiju Liao, Fuqiang Liu, Dan Li, Hua Ding, Lei Zhou, Weiping Zhu, Fan Ding, Peng Zhang, Xiaoye Wang, Jianyi Yao, Nijuan Xiang, Suizan Zhou, Ying Song, Hualin Su, Rui Wang, Jian Cai, Yang Cao, Xianjun Wang, Tian Bai, Jianjun Wang, Zijian Feng, Yanping Zhang, Marc-alain Widdowson, Qun Li, Public Health, Influenza Division, Zhejiang Province, Fujian Province, Henan Province, Hunan Province, Jiangxi Province, Jinan City, Shandong Province, Disease Control, and Changping District. Risk Factors for Influenza A(H7N9) Disease - China, 2013. *Clinical Infectious Diseases*, 2014.

[15] Yuzo Arima and Sirenda Vong. Human infections with avian influenza A(H7N9) virus in China: preliminary assessments of the age and sex distribution. *Western Pacific Surveillance and Response Journal*, 4(2):1–3, 2013. ISSN 2094-7313. doi: 10.5365/WPSAR.2013.4.2.005. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3762971\&tool=pmcentrez\&rendertype=abstract.

[16] World Health Organization. Frequently Asked Questions on human infection caused by the Avian Influenza A(H7N9) virus, 2014. URL http://www.who.int/influenza/human_animal_interface/faq_H7N9/en/.

[17] T Kageyama, S Fujisaki, E Takashita, H Xu, S Yamada, Y Uchida, G Neumann, T Saito, and Y Kawaoka. Genetic analysis of novel avian A (H7N9) influenza viruses isolated from patients in China , February to April 2013. *Eurosurveillance*, 18(15), 2013.

[18] B J Cowling, G Freeman, J Y Wong, P Wu, Q Liao, E H Lau, J T Wu, R Fielding, and G M Leung. Preliminary inferences on the age-specific seriousness of human disease caused by avian influenza A(H7N9) infections in China , March to April 2013. *Eurosurveillance*, 18(19), 2013.

[19] Caitlin Rivers, Kristian Lum, Bryan Lewis, and Stephen Eubank. Estimating human cases of avian influenza A(H7N9) from poultry exposure: Supplementary material. *Figshare*, May 2013. doi: doi:10.6084/m9.figshare.688050. URL `http://figshare.com/articles/Estimating\_human\_cases\_of\_avian\_influenza\_A\_H7N9\_from\_poultry\_exposure/688050`.

[20] New England Journal of Medicine H7N9 Map, 2014. URL `http://healthmap.org/h7n9/`.

[21] Multiple. Human Case List of Provincial / Ministry of Health / Government Confirmed Influenza A(H7N9) Cases, 2013. URL `www.flutrackers.com/forum/showthread.php?t=202713`.

[22] Population of provinces in China by age and sex, 2005. URL `www.stats.gov.cn/tjsj/ndsj/renkou/2005/html/0104.htm`.

[23] Participation rate: the purchase of goods and services, 2005. URL `http://www.stats.gov.cn/tjsj/qtsj/2008sjlydczlhb/P020091029588656311153.pdf`.

[24] Ella Lee. Outdated wet markets to clean up their act, August 2012.

[25] Susan Schwartz. Stores aim to win wet market custom, April 2012.

[26] Sector composition of urban employment by age and sex, 2006. URL `http://www.stats.gov.cn/tjsj/ndsj/laodong/2006/html/01-60.htm`.

[27] Chris Ka, Pun Mok, Hok Yeung, Chi Wai, Fun Sia, Maxime Lestra, and Malcolm Nicholls. Pathogenicity of the Novel A/H7N9 Influenza Virus in Mice. *mBio*, 4(4), 2013. doi: 10.1128/mBio.00362-13.Editor.

[28] Guangyu Zhao, Chenfeng Liu, Zhihua Kou, Tongtong Gao, Ting Pan, Xiaohong Wu, Hong Yu, Yan Guo, Yang Zeng, Lanying Du, Shibo Jiang, Shihui Sun, and Yusen Zhou. Differences in the pathogenicity and inflammatory responses induced by avian influenza A/H7N9 virus infection in BALB/c and C57BL/6 mouse models. *PloS One*, 9(3):e92987, January 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0092987. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3968029\&tool=pmcentrez\&rendertype=abstract`.

[29] Zhongfang Wang, Anli Zhang, Yanmin Wan, Xinian Liu, Chao Qiu, Xiuhong Xi, Yanqin Ren, Jing Wang, Yuan Dong, Meijuan Bao, Liangzhu Li, Mingzhe Zhou, Songhua Yuan, Jun Sun, Zhaoqin Zhu, Liang Chen, Qingsheng Li, Zhiyong Zhang, Xiaoyan

Zhang, Shuihua Lu, Peter C Doherty, Katherine Kedzierska, and Jianqing Xu. Early hypercytokinemia is associated with interferon-induced transmembrane protein-3 dysfunction and predictive of fatal H7N9 infection. *Proceedings of the National Academy of Sciences of the United States of America*, 111(2):769–74, January 2014. ISSN 1091-6490. doi: 10.1073/pnas.1321748111. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3896201\&tool=pmcentrez\&rendertype=abstract`.

[30] Eric H Y Lau, Jiandong Zheng, Tim K Tsang, Qiaohong Liao, Bryan Lewis, John S Brownstein, Sharon Sanders, Jessica Y Wong, Sumiko R Mekaru, Caitlin Rivers, Peng Wu, Hui Jiang, Yu Li, Jianxing Yu, Qian Zhang, Zhaorui Chang, Fengfeng Liu, Zhibin Peng, Gabriel M Leung, Luzhao Feng, Benjamin J Cowling, and Hongjie Yu. Accuracy of epidemiological inferences based on publicly available information: retrospective comparative analysis of line lists of human cases infected with influenza A(H7N9) in China. *BMC Medicine*, 12(88), January 2014. ISSN 1741-7015. doi: 10.1186/1741-7015-12-88. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4066833\&tool=pmcentrez\&rendertype=abstract`.

Chapter 3

Modeling Emergence Scenarios of Middle East Respiratory Syndrome Coronavirus

Manuscript authors: Caitlin M. Rivers, Bryan L. Lewis, Justin O'Hagan, Manoj Gambhir, James Hyman, Sara Del Valle

3.1 Forward

Middle East Respiratory Syndrome Coronavirus (MERS-CoV, or MERS) was identified in Saudi Arabia in September 2012. The first case was announced on the PRO-Med mailing list after a sample isolated from a patient with severe respiratory disease revealed a novel coronavirus [1]. In a 2012 case report published in the New England Journal of Medicine, the first clinical description of the new disease, the authors noted the virological and clinical picture of MERS was reminiscent of Severe Acute Respiratory Syndrome Coronavirus (SARS) [2]. SARS emerged in 2002, alarming the global public health community when a patient exported the disease from China to Canada by plane very soon after the disease was first recognized [3]. Although the international spread of infectious disease was hardly new, the ease and speed at which the novel disease was transported across the globe was alarming.

The similarities between MERS and SARS made MERS a disease to watch. Both SARS and MERS are coronaviruses of zoonotic origin, and both have a high case fatality risk; the WHO currently reports nearly 30% [4] of MERS patients die. Both also appear to spread easily in healthcare settings, putting healthcare workers at risk [3]. The exact mechanism of human to human transmission remains unknown, though droplet and contact transmission are suspected [5, 6, 7].

To date, much is unknown about the animal reservoir and transmission mechanisms for MERS. Many serological and epidemiological studies identify camels as a likely source of human infection. Seroprevalence in camels in Saudi Arabia and other Arabian Peninsula countries ranges from 6%-100% [8, 9, 10, 11, 12, 13]. Evidence of infection in camels in Saudi Arabia dates back as far as 1992 [13].

Camels outside the Arabian Peninsula may also harbor the virus. A 2013 serosurvey of camels in Egypt concluded that 94%-98% of the 110 camels tested had antibodies to MERS [14], though a similar study in Egypt yielded just a 4% positive proportion [15]. Camels in Spain, Nigeria, Tunisia, Ethiopia and elsewhere have also tested positive for antibody response [12, 16]. Even a serosurvey of camel samples dating back to the early 1980s in Sudan, Somalia and Egypt found evidence of MERS infection [17]. Serology of other domesticated animals like sheep, equids, goats, cattle and chickens has uncovered no evidence for infection [8, 12, 9, 18]. Given the long history of MERS presence in camels in regions outside the Arabian Peninsula, it is unclear why human infection has not occurred or been identified in those regions.

Epidemiological studies support serologic evidence implicating camels. A 2014 study collected viral samples from camels whose owner was infected with MERS-CoV. The viral genome of the infected camels was identical to that isolated from the human patient, evidence of camel to human transmission [19]. Similar findings were reported in a 2013 study of a Qatari camel-human cluster [10]. However, not all epidemiological studies have been conclusive. One study conducted at a healthcare center in Saudi Arabia found that only one of the 70 patients studied reported camel contact [20]. Another study of 26 index patients and 280 secondary contacts found that just two had contact with camels [21]. Seven of the secondary contacts showed evidence of asymptomatic MERS-CoV carriage, so the authors speculate that the true case burden may be higher than previously thought, and that young people in particular may have unrecognized illness. A 2015 serosurvey of Saudi Arabian people routinely exposed to MERS-infected camels found no evidence of camel to human transmission, leading the authors to conclude that "zoonotic transmission of this virus from dromedaries is rare" [8].

Data availability for this extended outbreak has been quite poor. A vast majority of recognized cases have occurred in Saudi Arabia, which makes their participation in international inquiry vital. Details about Saudi Arabia's surveillance and diagnostic procedures have not been made publicly available. Despite many pleas, no case control study or other rigorous epidemiological inquiry has been conducted [22, 23, 24, 25, 26, 27, 28]. From the outside perspective, it seems as though the Kingdom of Saudi Arabia has been reluctant to engage in the research needed to better understand the outbreak. The Saudi Arabian Ministry of Health announced the launch of a case control study in June 2014, but no results or further information on the project have since been announced [29].

The exception to the slow flow of data is that case reports for Middle East Respiratory Syndrome were released to the public soon after cases were diagnosed, though they were made available through the Kingdom of Saudi Arabia's Ministry of Health and the World

Health Organization rather than the media. Case reports included demographic information, location, case status, and dates of onset. This facilitated the curation of a line listing similar to the one described in Chapter 1. For this research, we collected data on the demographics, clinical timeline, and exposure source for all available cases. We used this data to construct and parameterize a compartmental model. A primary objective of this research was to provide insight into the emergence scenarios and transmission dynamics to aid planning response and control efforts, should the disease be found outside the Middle East.

The compartmental model is SEIR style, with additional compartments for hospitalized and deceased patients. There are also two susceptible compartments, high risk and low risk. Unlike most compartmental models, the MERS model does not assume homogenous mixing – it has separate sub-models for men and women, and four age groups for each gender. Each demographic group sub-model interacts with the other groups with varying frequency to reflect the segregated mixing of socially-conservative Saudi Arabia. This structure was used to determine the expected age and sex distribution of MERS cases, which served as a baseline from which to compare the observed distribution.

There were three model parameters for which values could not be determined, but each would provide insight into the emergence of the outbreak. The parameters are susceptibility in a high risk group that included agricultural workers, healthcare workers, and people in poor health; susceptibility in a normal risk group; and frequency of zoonotic introductions. We used a parameter sweep approach to isolate parameter spaces that would be plausible given biological priors and fit to observed data. Other model parameters were derived from the literature and from the line list assembled from publicly available case reports. This approach incorporates data fusion methods to learn about outbreaks earlier than would otherwise be possible.

Results suggest that a single spillover event with all subsequent cases being attributed to human to human transmission could not explain the observed case pattern. Thus we were able to conclude that ongoing zoonotic introductions into the human population are occurring. This finding is well supported by studies that indicate the human to human reproduction number is not high enough to support ongoing transmission [30, 31]. Although we were not able to put an upper bound on the number of spillover events, we did determine that introductions likely occur at least daily. Although wild animal exposures cannot be ruled out, this range indicates that domestic animals are likely a primary source of exposure. Epidemiological field work has not yet confirmed the animal reservoir, but there is strong evidence that dromedary camels may play a role. Our spillover estimates are compatible with this hypothesis.

Our analysis suggests that less than 20% of MERS cases are being detected, meaning a vast majority are going undiagnosed. These infectious people may be accessing the healthcare system but are not receiving appropriate infection control measures, or they may be circulating in their communities and putting others at risk. Additional findings suggest that a majority of missed cases are occurring in adult men, particularly those ages 40-64. We hy-

pothesize that this finding can be attributed to social norms in Saudi Arabia that put men at higher risk; men in this demographic are more likely to work in agricultural and healthcare settings than women, and are therefore presumably more likely to be exposed. This fails to support a previously suggested theory in the research community that posits the reason there are so few cases in women compared to men is because there exists a significant case detection bias attributable to the gender-segregation in Saudi Arabian society.

This work demonstrates the utility of combining traditional methods like mathematical modeling with emerging data fusion methodologies. Modeling efforts to do not need to wait for an outbreak to finish and for ground-truth data to be produced before producing scholarly insights. We show that rapid modeling efforts using real-time and publicly available data sources can uncover insights into the outbreak sooner than previously thought. Furthermore, this study highlights the importance of using computational epidemiology in the early days of an outbreak, when results can be used to design and target public health interventions while benefit can still be had.

This manuscript is being prepared for submission.

3.2 Abstract

Background: Middle East Respiratory Syndrome Coronavirus (MERS-CoV) is a novel human pathogen that emerged in Saudi Arabia in 2012. The disease's high mortality rate, propensity to affect healthcare workers, and severe respiratory symptoms have earned it comparisons to Severe Acute Respiratory Syndrome (SARS), which killed 916 people and infected thousands in 2002-2003.

Methods: An age and sex structured ordinary differential equation model with non-homogenous mixing was developed using publicly available data from media reports and the scientific literature.

Results: Findings suggest that MERS-CoV emerged multiple times in the human population, likely from a zoonotic source associated with older men. Between 86% and 96% of incident cases in the last year have gone undiagnosed, with the strongest detection bias found in adult and elderly men. Exposure to an animal source, likely a domesticated or agricultural animal, continues to be a serious risk factor for contracting the disease.

Conclusion: Because MERS-CoV is a brand new human pathogen, little is known about how it emerged, whether it may become a widespread epidemic, or how many people it could eventually infect. Insights derived from mathematical models can help provide insight into the dynamics of the disease to help inform public health efforts.

3.3 Introduction

As of June 11, 2014, MERS has infected 699 people worldwide, primarily in Saudi Arabia. The World Health Organization has tallied 209 deaths, bringing the case fatality risk to 30% [32]. An additional 113 cases and 282 deaths have been announced by the Saudi Arabian Ministry of Health, but details of those cases have not been released and have not been included here [33].

We compiled a line list of 707 cases, including 139 deaths, from publicly available data current as of Jun 4. Our case counts differ from the official tally due to differences in the way cases and deaths are announced and counted by public health agencies and the World Health Organization.

In our line list, over half of cases are in men (61%), three quarters of deaths (73%) and 66% of cases reported as critical are in men, suggesting a significant sex bias. The mean age of all infected patients is 48 (median = 47), but among patients reported as dead or in critical condition, the mean age is 56 (median = 56). Patients infected with MERS present with fever, chills, sore throat, and sometimes gastrointestinal symptoms [5, 6, 34]. Severe cases progress to respiratory and kidney failure, and ultimately death. The incubation period is thought to be around 5 days, but may range from 3-21 days [5]. The incubation period is thought not to be infectious [35]. SARS was able to be controlled, in part, because infected patients did not begin shedding the virus until symptoms began, making isolation of infectious patients more feasible and effective [36]. The similarity of disease dynamics between SARS and MERS suggests that effective control is possible.

Mathematical modeling using systems of ordinary differential equations (ODE) has played a major role in understanding infectious disease dynamics and the impacts of mitigation strategies [37, 38]. Previous models have been used to estimate the reproduction number, or the average number of secondary cases from each infected case, for a variety of diseases including SARS [39]. Similar approaches have been used to evaluate the effectiveness of possible interventions, like the use of face masks and school closures during H1N1 [40].

Very limited studies have used mathematical and computational approaches to understand the dynamics and impact of mitigation strategies. Breban et al. used Bayesian analysis to estimate the basic reproductive number and evaluate different scenarios to estimate the risk of a pandemic [30]. However, most of the published work on MERS has focused on epidemiological, demographic, clinical, and genetic aspects. For example, nascent MERS literature has identified dromedary camels as possible reservoirs for the virus [12, 10, 9, 8, 14, 13]. Reports from several authors have described a number of clusters in which MERS has successfully been transmitted from human to human, particularly in healthcare settings, though the mechanism for transmission has not yet been confirmed [6, 41, 34, 5, 42].

This effort seeks to understand the impact of age and general mixing on disease spread, and the impact of bias detection on the overall dynamics and spread of MERS.

3.4 Methods

3.4.1 Data sources

A line listing was collected from publicly available sources, including reports from the media, the World Health Organization [43], case reports published in the scientific literature, bulletins from Saudi Arabia's Ministry of Health [44], and other public line listings [45, 46]. Where possible, data collected include date of disease onset or report; patient age, sex and health status; country and town of origin; whether the patient has underlying medical conditions or is a healthcare worker; and whether the patient has had contact with another confirmed case. Cases originating outside of the Kingdom of Saudi Arabia (KSA) were excluded from the analysis. Selected cases originated in KSA between June 2012 and June 4, 2014, and had both age and sex information available. There were 570 cases that met the criteria for model building and parameterization.

Several model features such as the population and workforce structure were based on other sources of existing data. Data on the age and sex for the entire population was obtained from the 2004 Census in the Kingdom of Saudi Arabia [47]. The age distribution of agricultural workers was extracted from the 2012 Census in the KSA [48], and the gender distribution was generated to match overall worker distribution by sex in the country [49]. The demographic distribution of healthcare workers was estimated from the overall worker distribution in the country, with the sum total matching real data [50]. No data on the prevalence and distribution of comorbidities in the country was available, however it was noted that the number and rate of people with diabetes, cancer, chronic respiratory disease, and cardiovascular disease is approximately the same as in the United States [51]. Thus, data on the demographic distribution of people with comorbidities were generated from U.S. data on people who self-report being in poor or fair health in the United States National Health Interview Survey conducted by the Centers for Disease Control and Prevention [52].

Age	**Sex**	**Code**	**Population**	**High risk pop.**
Youth 0-19	M	YM	5,263,843	56,729
	F	YF	5,117,110	51,929
Young adult 20-39	M	YAM	3,932,720	395,252
	F	YAF	2,904,137	207,229
Adult 40-64	M	AM	1,757,565	501,636
	F	AF	1,245,076	249,866
Elderly 65+	M	EM	360,838	126,615
	F	EF	265,592	67,010

Table 3.1: Total population and high risk population of each demographic group.

3.4.2 Model description

We developed a mathematical model where the population is stratified by age and gender to reflect the strong demographic bias of observed cases. The population was divided into eight groups: four age groups for males and four for females. The age groups are: youth, ages 0-19 (denoted as YM and YF for youth male and female, respectively); young adults ages 20-39 (denoted as YAM and YAF); adults ages 40-64 (AM and AF); and elderly ages 65+ (EM and EF). We hypothesize that the strong demographic bias in observed cases is attributable in part to cultural norms that prohibit homogenous mixing [53]. Thus contact between age and sex groups was estimated by consulting experts familiar with the Saudi society, and assumes a strong segregation between males and females. The mixing matrix representing the probability of demographic group k coming in contact with group j (for $k = j =$ YM, YF, YAM, YAF, AM, AF, EM, and EF) is shown in Figure 3.1.

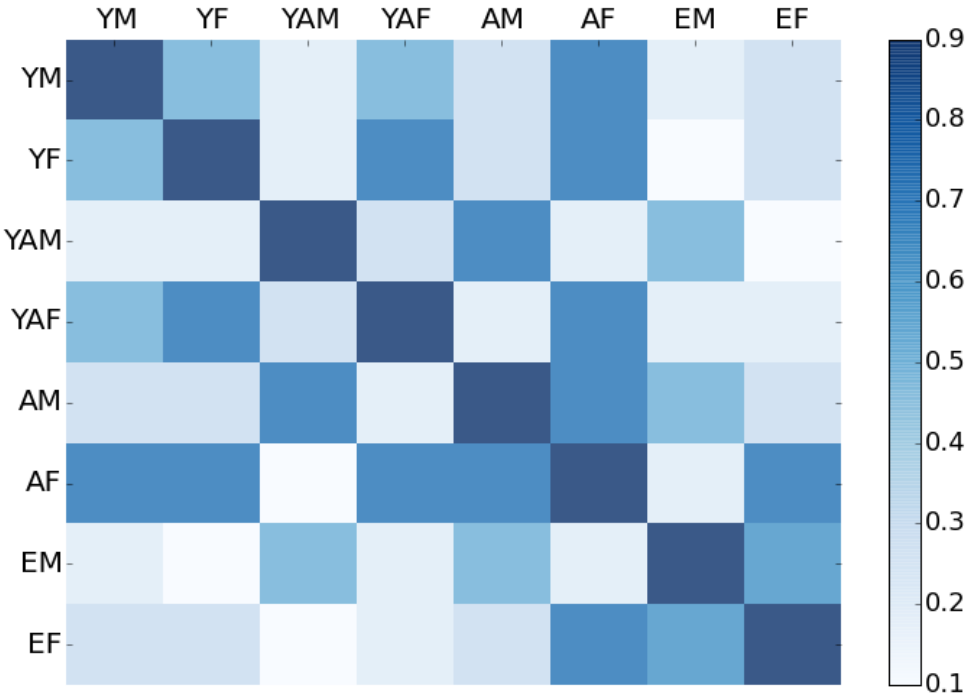

Figure 3.1: **Mixing matrix: Estimated contact patterns between demographic group k and group j.**

Individuals in the model are assigned to one of two susceptible populations: S_1, which represents the high-risk group and S_2, which represents the normal or low risk group.The high risk

group, S_1 includes agricultural workers, who are assumed to be at higher risk of contracting the disease from a zoonotic source; healthcare workers, who are heavily represented in the line listing; and people in poor or fair health who we assume to be more susceptible due to underlying comorbidities. There are a total of 1,656,266 people in the high risk group in the model. The normal or low risk group, S_2, includes all others in the population (n = 19,190,615).

After contact with an infectious person, susceptible individuals move to an exposed compartment at rate λ_1 or λ_2, where they are no longer susceptible, but are not yet infectious. In this model, λ is a function of the probability of contact with other age and sex groups as determined by the mixing matrix (α), the transmission probability per contact (β), and the fraction of contacts that are infected. Thus λ is given by:

$$\lambda_{1,2} = \begin{pmatrix} \text{Probability of} \\ \text{Contacts per} \\ \text{Unit Time} \end{pmatrix} \begin{pmatrix} \text{Infectivity} \\ \text{of the} \\ \text{Disease} \end{pmatrix} \begin{pmatrix} \text{Fraction of} \\ \text{Contacts that} \\ \text{are Infected} \end{pmatrix}$$

$$\lambda_{1,2} = \sum_{k=1} (\beta_1^{kj} \alpha^{kj} \frac{I^k + qE^k + lJ^K}{N^k}) \tag{3.1}$$

Exposed persons, E, move to an infectious compartment, I, where they are symptomatic and thus fully infectious. Infectious individuals can then either enter the diagnosed compartment, J, where their infectiousness is reduced to $\frac{1}{5}$ the level of fully infectious people, or they can

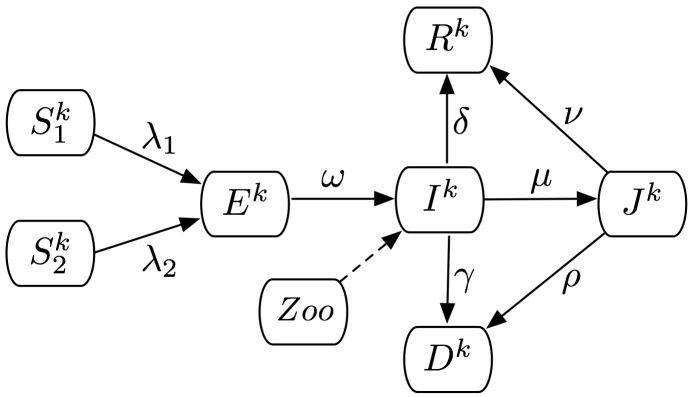

Figure 3.2: **MERS disease progression model.** The schematic shows the epidemiological progression for high-risk (S_1) and non high-risk (S_2) individuals. The arrows represent movement of individuals from one group to an adjacent one.

recover (R) or die (D). Diagnosed people can also either die or recover.

An additional feature of the model is the continuous introduction of new infectious individuals into the population, which represents zoonotic spillover. We assume that agricultural workers are at highest risk for contracting the disease from an animal source. Thus zoonotic introductions in the model are distributed across age and sex demographic groups proportional to the distribution of agricultural workers in Saudi Arabia. The rate at which these introductions occur is varied across different model scenarios.

Using the disease progression model in Figure 3.2 and Equation 3.4.2, we arrive at the following system of differential equations:

$$\frac{dS_1^k}{dt} = -(\lambda_1 S_1^k) \tag{3.2}$$

$$\frac{dS_2^k}{dt} = -(\lambda_2 S_2^k) \tag{3.3}$$

$$\frac{dE^k}{dt} = \lambda_1 S_1^k + \lambda_2 S_2^k - \omega E^k \tag{3.4}$$

$$\frac{dI^k}{dt} = \omega E^k - \mu I^k - \gamma I^k - \delta I^k \tag{3.5}$$

$$\frac{dJ^k}{dt} = \mu I^k - \rho J^k - \nu J^k \tag{3.6}$$

$$\frac{dD^k}{dt} = \gamma I^k + \rho J^k \tag{3.7}$$

$$\frac{dR^k}{dt} = \delta I^k + \nu J^k \tag{3.8}$$

There are a total of seven differential equations for each age and sex groups, totaling 56 equations. The model is initially seeded with just one infectious person belonging to the male aged 65+ demographic group, and the disease is simulated for 365 days to represent long time scale of the observed outbreak.

	Description	Units	Baseline	Range	Reference
S_1^k	Highly susceptible population of demographic group k	People	See Table 3.1	1-21 million	Saudi Arabia Census
S_2^k	Normal susceptible population of demographic group k	People	See Table 3.1	1-21 million	Saudi Arabia Census
N^k	Total population of demographic group k	People	1-21 million	See Table 3.1	Saudi Arabia Census
β^{kj}	Transmission rate between demographic groups k and j	1	See text	0-1	
α^k	Probability of contact with other demographic groups	1	See Figure 3.1	0-1	
q	Reduced infectivity due to incubation	1	.20	0-1	
l	Reduced infectivity due to diagnosis	1	.80	0-1	
ω	Incubation relative rate	Day	.20	0-1	Assiri et al, Cauchemez et al [5, 31]
μ	Diagnostic rate of infected	Day	Varied	0-1	
γ	Death rate for infectious	Day	.10	0-1	Estimated
δ	Recovery rate for infectoius	Day	.16	0-1	Estimated
ρ	Death rate for diagnosed	Day	.10	0-1	Assiri et al [5]
ν	Recovery rate for diagnosed	Day	.10	0-1	Assiri et al [5]

Table 3.2: Transfer rates in disease progression model.

Estimation of parameter values

Epidemiological parameters as seen in Table 3.2 were estimated from medical case reports published in the scientific literature. The incubation period is reported to be between 3 and 21 days, with a mean of around 5 days [5, 31]. The incubation period is deterministic in the model and was assumed to be $\omega = \frac{1}{5}$. Although the infectious period is unknown, case reports suggest the symptomatic period is between 5 and 27 days [5]. The time from onset of the infectious or symptomatic period to death in undiagnosed cases is unknown, thus, we estimate this value to be $\gamma = \frac{1}{10}$.

One of our goals was to explore the impact of diagnosis bias on the overall dynamics of MERS. For this purpose, we varied the diagnostic rate, μ, where $\mu = 0$ implies no diagnosis is occurring and $\mu = 1$ implies everyone is diagnosed. For these scenarios, the rate at which diagnosed individuals progress from infection to death is given by $\rho = \frac{1}{12}$.

Because the case fatality risk of a MERS infection is high ($> 30\%$), little is known about individuals who recover from the disease. However, we assume that among diagnosed cases, those who recover take longer to do so than those who progress from the infectious compartment to death. Thus, the infectious to recovery rate is given by $\delta = \frac{1}{9}$, and the diagnosed to recovery rate is $\nu = \frac{1}{10}$. We recognize that mild or asymptomatic cases likely progress at a different rate than severely infected individuals. However, the disease progression for those individuals is unknown, so the parameter is fixed.

Scenario choice

There are three parameters for which we have no available data: the probability of disease transmission given contact between a susceptible and an infectious individual (β), the variation in this value between high and low susceptible risk groups, and the frequency of zoonotic seeding that occurs. In order to identify possible ranges for these parameter values, we conducted a broad parameter sweep testing combinations of these variables. We then narrowed down the results using the following assumptions:

- We eliminate scenarios where the overall number of cases in the model is less than the number of observed cases, and greater than 25 times the number of observed cases.
- We eliminate scenarios where the number of cases per age group in the model is less than truly observed cases per age group.

Filtering model iterations using these assumptions reduces the number of scenarios with no diagnoses from 5,400 to 52 simulations, thereby isolating the plausible scenario space and yielding interpretable results.

3.5 Results

3.5.1 Model results

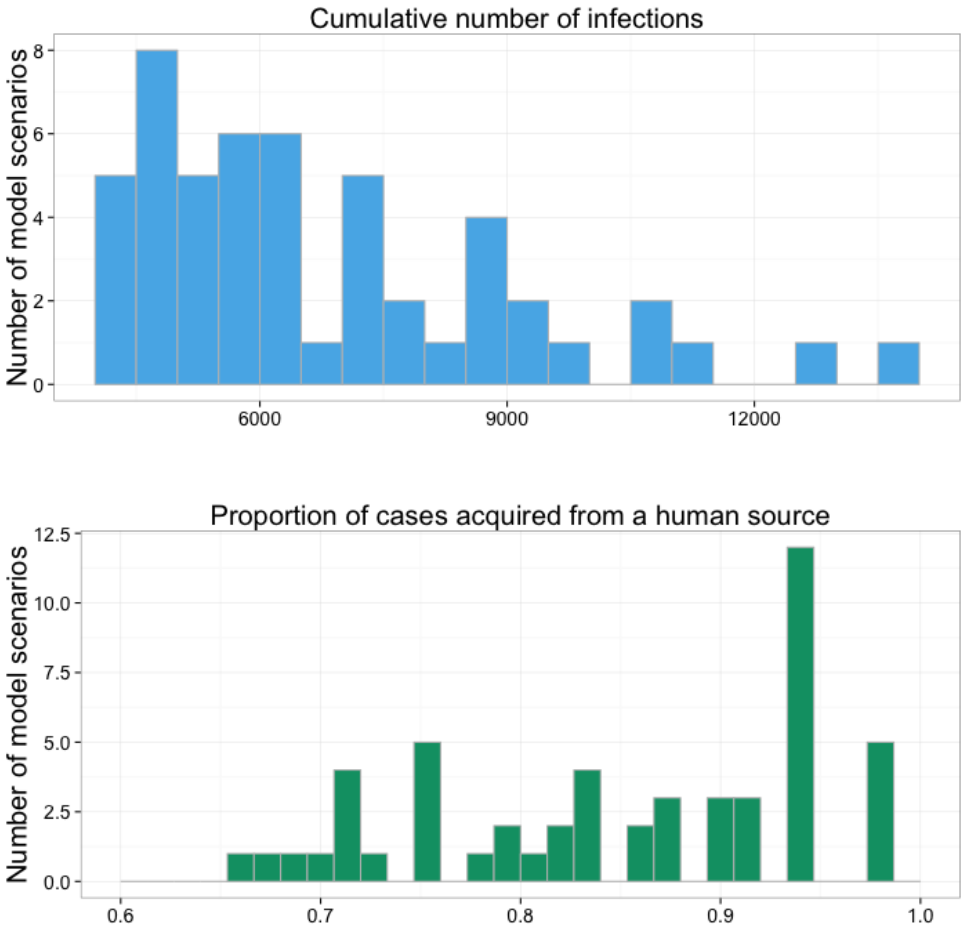

Figure 3.3: **Distribution of simulation results.**

Model results suggest total case counts range from 4,087 to 14,085 ($\mu = 6,912, \omega = 2,503$), as shown in Figure 3.3. Case counts at the time of this writing are over 700, which suggests the proportion of cases that have been identified ranges from 4% to 14%.

Analysis of the distribution of cases in the model by demographic group suggests that most cases are occurring in men, particularly those ages 40-64 and 65+, as shown in Figure 3.4. Men in these age groups are more likely to work in agriculture and healthcare professions, and are assumed to have frequent contact with others. The number of observed cases of MERS in men ages 40-64 is relatively low, which suggests that case detection among this

group should be improved. Men ages 19-39 are also more heavily affected than current observed cases suggests, though to a lesser degree than their older counterparts. Men under the age of 19 experience few cases in the model, which fits with the observed case distribution. One possible explanation is that young men are not employed in high-risk professions like healthcare and agriculture sectors and are generally in good health, which reduces their susceptibility.

Women in general have fewer cases than their male counterparts. The most heavily affected group is elderly women, who may be at higher risk because they are more likely to be in poor health, and have frequent contact with other high-risk groups like elderly men. Younger age groups experience relatively few infections, likely because women in Saudi Arabia generally do not work outside the home, and are therefore avoiding likely sources of occupational exposure.

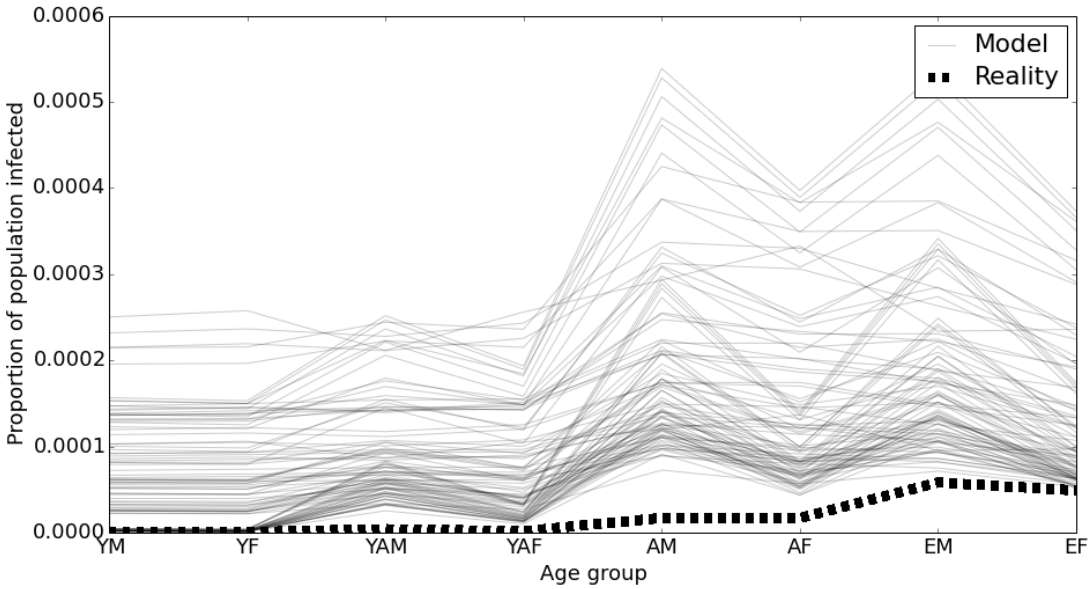

Figure 3.4: **Distribution of cases by age groups across 100 simulations.**

3.5.2 Spillover events

Among the 52 plausible simulations, the proportion of cases that originate from a zoonotic source ranges from less than 1% to 37%. We estimate that cases acquired from a zoonotic source occur at least daily. The frequency of spillover events in the model suggests contact with domesticated or agricultural animals are a possible infection source; a wild animal host would likely have a much lower spillover rate. Over half (54%) of cases in Saudi Arabia

report having contact with animals, which suggests they are zoonotically acquired. Thus model results indicate that there are a significant number of cases acquired from human to human contact. If this model-generated hypothesis is supported, it would increase estimates of the basic reproduction number substantially.

3.5.3 Impact of diagnosis

In order to evaluate the impact of effective case finding on the natural history of the outbreak, we added a diagnostic compartment to the disease progression model. Diagnosis reduced infectivity by 80% in the model. The rate at which cases transition from infected to diagnosed was varied from 0.1 to 1 in 0.1 intervals, and an additional .05 value. Infected cases could also die or recover without being diagnosed.

Multiple negative binomial regression shows that the diagnostic rate has a significant negative relationship with the total number of infections over the course of the outbreak. Each two unit increase in the proportion of cases diagnosed results in 1 fewer cases ($p < .01$, LCI=.43, UCI=.47 per unit), as seen in Figure 3.5. For example, a model configured with a β of 0.086 among the high-risk group and four zoonotic seeds each day produced 4,902 cases with 40% of cases identified, and 8,926 cases with 20% of cases identified. This finding suggests that effective case finding and isolation practices is critical to controlling the the outbreak.

Sensitivity analysis

In order to evaluate the possibility that the age and sex distribution we observe among plausible scenarios is not being driven entirely by our constructed mixing matrix, we ran a set of simulations using a homogenous mixing matrix such that each demographic group has the same probability of interacting with every other group. When we compared those to summary statistics from the constructed matrix simulations, we found that the results are robust to changes in the mixing matrix. Using the above example, with a β of 0.086 and four zoonotic introductions per day, the homogenous matrix produced slightly more cases than the constructed matrix (5,957 vs 4,902, respectively) at a 40% detection rate. The age distribution of cases under the homogenous matrix favors adult men, but to a lesser degree than the constructed matrix.

The results show that the constructed matrix yields more conservative results than the homogenous matrix. Even in the homogenous matrix, there is still a strong sex bias in the case distribution. This is likely because men are more likely belong to high risk groups in the Saudi Arabian population, and are thus at higher risk in the model.

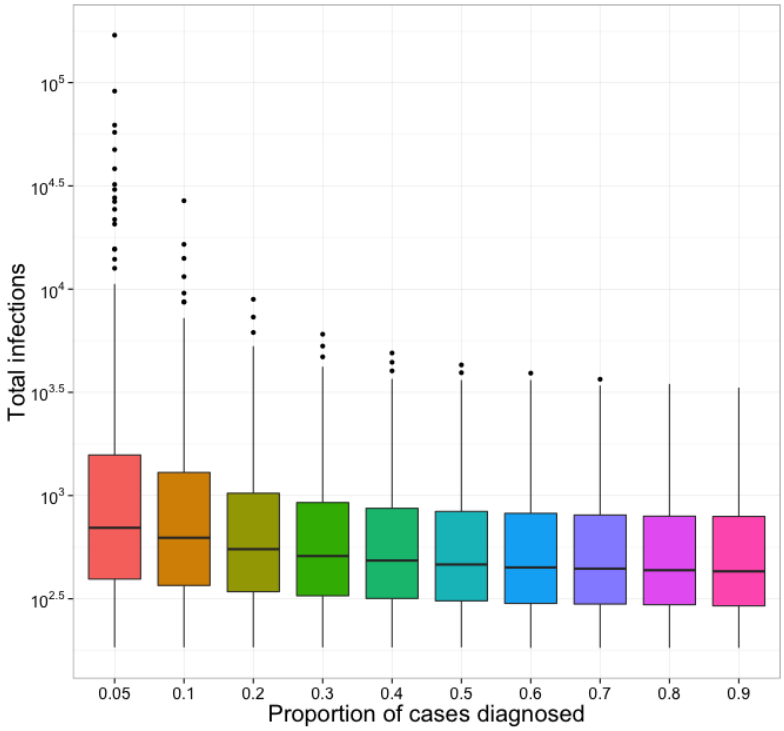

Figure 3.5: **Cumulative number of infections with variable diagnostic rate** Infections across 5,400 total model scenarios and 52 plausible model scenarios, with variable transmissibility values (β) and zoonotic seeding frequency. As the diagnostic rate increases, the cumulative number of infections decreases.

3.5.4 Discussion

Patterns from model scenarios suggest that up to 96% of cases of MERS were undetected over the last year. Our model shows that a demographic-based detection bias exists; young men ages 19-39 and adult men ages 40-64 may be experiencing a heavier case burden than is currently recognized. This bias could be due to care-seeking behaviors of those demographic groups, or it could be that cases among young, healthy individuals are likely to be mild or asymptomatic, and thus not warrant care-seeking behavior or testing.

An additional finding is that adult and elderly women may be experiencing a higher proportion of human to human transmission than men, possibly because they are less likely to work in an agricultural setting with contact exposure to an animal host. Improved surveillance, particularly a serosurvey, would be useful for understanding the true prevalence if such a study has not yet been undertaken.

Our results show that the model assumptions are robust to variations, specifically that mixing

patterns whether segregated or homogeneous show similar disease dynamics.

The finding that human cases from a zoonotic source are happening frequently is consistent with the current hypothesis that domesticated or agricultural animals are a likely source of zoonotic exposure. Current evidence points to dromedary camels as a potential source. Even so, model results suggest that a majority of cases are acquired through human to human transmission. We found that by detecting and isolating more cases, the number of incident infections could be significantly reduced.

3.5.5 Conclusion

The research reported in this paper is motivated by concerns about the potential impact of wide spread of MERS-CoV around the globe. An age-gender structured ordinary differential equation model was used to understand the potential source of infections and detection bias. Overall, our model suggests that the proportion of cases that are being detected is likely less than half. Most of the undetected cases are occurring in men, particularly adult men ages 40+. Our analysis supports the existing understanding that the disease has spilled over from animals to humans multiple times, possibly in domesticated or agricultural animals. We conclude that deterministic epidemic modeling can provide insights into the disease dynamics and guide mitigation strategies.

Acknowledgements

Caitlin Rivers was supported by the US National Institute of General Medical Sciences of the National Institutes of Health under award number 2U01GM070694-09. Additional support provided by Defense Threat Reduction Agency Validation Grant HDTRA1-11-1-0016.

Sara Del Valle was supported by Los Alamos National Laboratory under the Department of Energy contract DE-AC52-06NA25396 and a grant from NIH/NIGMS in the Models of Infectious Disease Agent Study (MIDAS) program (U01-GM097658-01).

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

3.6 Bibliography

[1] Ali Mohamed Zaki. PRO/EDR Novel coronavirus - Saudi Arabia: human isolate, September 2012.

[2] Ali M Zaki, Sander van Boheemen, Theo M Bestebroer, Albert D M E Osterhaus, and

Ron a M Fouchier. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. *The New England Journal of Medicine*, 367(19), November 2012. ISSN 1533-4406. doi: 10.1056/NEJMoa1211721. URL `http://www.ncbi.nlm.nih.gov/pubmed/23075143`.

[3] Michael D Christian, Susan M Poutanen, Mona R Loutfy, Matthew P Muller, and Donald E Low. Severe acute respiratory syndrome. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 38(10):1420–7, May 2004. ISSN 1537-6591. doi: 10.1086/420743. URL `http://www.ncbi.nlm.nih.gov/pubmed/15156481`.

[4] Al Jawf, Al Taif, and Hafar Al Batin. Middle East respiratory syndrome coronavirus (MERS-CoV) summary and literature update as of 11 June 2014. (June):1–8.

[5] A. Assiri, A. McGeer, T. M. Perl, C.S. Price, A. A. Al Rabeeah, D .T. Cummings, Z. N. Alabdullatif, M. Assad, A. Almulhim, H. Makhdoom, H. Madani, R. Alhakeem, J. Al-Tawfiq, M. Cotten, S. Watson, P. Kellam, A. Zumla, and Z. A. Memish. Hospital Outbreak of Middle East Respiratory Syndrome Coronavirus. *New England Journal of Medicine*, 369(5):407–416, June 2013. ISSN 0028-4793. doi: 10.1056/NEJMoa1306742. URL `http://www.nejm.org/doi/abs/10.1056/NEJMoa1306742`.

[6] Ali S Omrani, Mohammad Abdul Matin, Qais Haddad, Daifullah Al-Nakhli, Ziad a Memish, and Ali M Albarrak. A family cluster of Middle East Respiratory Syndrome Coronavirus infections related to a likely unrecognized asymptomatic or mild case. *International Journal of Infectious Diseases*, 17(9):e668–72, September 2013. ISSN 1878-3511. doi: 10.1016/j.ijid.2013.07.001. URL `http://www.ncbi.nlm.nih.gov/pubmed/23916548`.

[7] Benoit Guery, Julien Poissy, Loubna El Mansouf, Caroline Séjourné, Nicolas Ettahar, Xavier Lemaire, Fanny Vuotto, Anne Goffard, Sylvie Behillil, Vincent Enouf, Valérie Caro, Alexandra Mailles, Didier Che, Jean-Claude Manuguerra, Daniel Mathieu, Arnaud Fontanet, and Sylvie van der Werf. Clinical features and viral diagnosis of two cases of infection with Middle East Respiratory Syndrome coronavirus: a report of nosocomial transmission. *Lancet*, 6736(13):1–8, May 2013. ISSN 1474-547X. doi: 10.1016/S0140-6736(13)60982-4. URL `http://www.ncbi.nlm.nih.gov/pubmed/23727167`.

[8] M G Hemida, R a Perera, P Wang, M a Alhammadi, L Y Siu, M Li, L L Poon, L Saif, A Alnaeem, and M Peiris. Middle East Respiratory Syndrome (MERS) coronavirus seroprevalence in domestic livestock in Saudi Arabia, 2010 to 2013. *Eurosurveillance*, 18(50):20659, January 2013. ISSN 1560-7917. URL `http://www.ncbi.nlm.nih.gov/pubmed/24342517`.

[9] C B Reusken, M Ababneh, V S Raj, B Meyer, A Eljarah, S Abutarbush, G J Godeke, T M Bestebroer, I Zutt, M a Muller, B J Bosch, P J Rottier, a D Osterhaus, C Drosten, B L Haagmans, and M P Koopmans. Middle East Respiratory Syndrome coronavirus

(MERS-CoV) serology in major livestock species in an affected region in Jordan, June to September 2013. *Eurosurveillance*, 18(50):20662, January 2013. ISSN 1560-7917. URL http://www.ncbi.nlm.nih.gov/pubmed/24342516.

[10] Bart L Haagmans, Said H S Al Dhahiry, Chantal B E M Reusken, V Stalin Raj, Monica Galiano, Richard Myers, Gert-Jan Godeke, Marcel Jonges, Elmoubasher Farag, Ayman Diab, Hazem Ghobashy, Farhoud Alhajri, Mohamed Al-Thani, Salih a Al-Marri, Hamad E Al Romaihi, Abdullatif Al Khal, Alison Bermingham, Albert D M E Osterhaus, Mohd M Alhajri, and Marion P G Koopmans. Middle East respiratory syndrome coronavirus in dromedary camels: an outbreak investigation. *The Lancet Infectious Diseases*, 14(2):140–5, February 2014. ISSN 1474-4457. doi: 10.1016/S1473-3099(13)70690-X. URL http://www.ncbi.nlm.nih.gov/pubmed/24355866.

[11] Five MERS- infected camels found in Kuwait, October 2014. URL http://www.kuna.net.kw/ArticleDetails.aspx?id=2381936\&language=en.

[12] Chantal B E M Reusken, Bart L Haagmans, Marcel a Müller, Carlos Gutierrez, Gert-Jan Godeke, Benjamin Meyer, Doreen Muth, V Stalin Raj, Laura Smits-De Vries, Victor M Corman, Jan-Felix Drexler, Saskia L Smits, Yasmin E El Tahir, Rita De Sousa, Janko van Beek, Norbert Nowotny, Kees van Maanen, Ezequiel Hidalgo-Hermoso, Berend-Jan Bosch, Peter Rottier, Albert Osterhaus, Christian Gortázar-Schmidt, Christian Drosten, and Marion P G Koopmans. Middle East respiratory syndrome coronavirus neutralising serum antibodies in dromedary camels: a comparative serological study. *The Lancet Infectious Diseases*, 13(10):859–66, October 2013. ISSN 1474-4457. doi: 10.1016/S1473-3099(13)70164-6. URL http://www.ncbi.nlm.nih.gov/pubmed/23933067.

[13] Abdulaziz N Alagaili, Thomas Briese, Nischay Mishra, Vishal Kapoor, Stephen C Sameroff, Emmie De Wit, and Vincent J Munster. Middle East Respiratory Syndrome Coronavirus Infection in Camels in Saudi Arabia. *mBio*, 2014. doi: 10.1128/mBio.00884-14.Editor.

[14] R a Perera, P Wang, M R Gomaa, R El-Shesheny, A Kandeil, O Bagato, L Y Siu, M M Shehata, a S Kayed, Y Moatasim, M Li, L L Poon, Y Guan, R J Webby, M a Ali, J S Peiris, and G Kayali. Seroepidemiology for MERS coronavirus using microneutralisation and pseudoparticle virus neutralisation assays reveal a high prevalence of antibody in dromedary camels in Egypt, June 2013. *Eurosurveillance*, 18(36), January 2013. ISSN 1560-7917. URL http://www.ncbi.nlm.nih.gov/pubmed/24079378.

[15] Daniel K W Chu, Leo L M Poon, Mokhtar M Gomaa, Mahmoud M Shehata, Ranawaka A P M Perera, Dina Abu Zeid, Amira S El Rifay, Lewis Y Siu, Yi Guan, Richard J Webby, and Mohamed A Ali. MERS Coronaviruses in Dromedary Camels, Egypt. *Emerging Infectious Diseases*, 20(6), 2014.

[16] Chantal B.E.M. Reusken, Lilia Messadi, Ashenafi Feyisa, Hussaini Ularamu, Gert-Jan Godeke, Agom Danmarwa, Fufa Dawo, Mohamed Jemli, Simenew Melaku, David

Shamaki, Yusuf Woma, Yiltawe Wungak, Endrias Zewdu Gebremedhin, Ilse Zutt, Berend-Jan Bosch, Bart L. Haagmans, and Marion P.G. Koopmans. Geographic Distribution of MERS Coronavirus among Dromedary Camels, Africa. *Emerging Infectious Diseases*, 20(7), July 2014. ISSN 1080-6040. doi: 10.3201/eid2007.140590. URL `http://wwwnc.cdc.gov/eid/article/20/7/14-0590\_article.htm`.

[17] Marcel A Müller, Victor Max Corman, Joerg Jores, Benjamin Meyer, Mario Younan, Anne Liljander, Berend-jan Bosch, Erik Lattwein, Mosaad Hilali, Bakri E Musa, and Set Bornstein. MERS Coronavirus Neutralizing Antibodies in Camels, Eastern Africa, 1983-1997. *Emerging Infectious Diseases*, 20(12):2093–2095, 2014.

[18] Benjamin Meyer, Ignacio García-bocanegra, Ulrich Wernery, Renate Wernery, Andrea Sieberg, Marcel A Müller, Jan Felix Drexler, Christian Drosten, and Isabella Eckerle. Serologic Assessment of Possibility for MERS-CoV Infection in Equids. *Emerging Infectious Diseases*, 21(1), 2015.

[19] Esam I Azhar, Sherif A El-Kafrawy, Suha a Farraj, Ahmed M Hassan, Muneera S Al-Saeed, Anwar M Hashem, and Tariq A Madani. Evidence for camel-to-human transmission of MERS coronavirus. *The New England Journal of Medicine*, 370(26), June 2014. ISSN 1533-4406. doi: 10.1056/NEJMoa1401505. URL `http://www.ncbi.nlm.nih.gov/pubmed/24896817`.

[20] Mustafa Saad, Ali S Omrani, Kamran Baig, Abdelkarim Bahloul, Fatehi Elzein, Mohammad Abdul Matin, Mohei a a Selim, Mohammed Al Mutairi, Daifullah Al Nakhli, Amal Y Al Aidaroos, Nisreen Al Sherbeeni, Hesham I Al-Khashan, Ziad a Memish, and Ali M Albarrak. Clinical aspects and outcomes of 70 patients with Middle East respiratory syndrome coronavirus infection: a single-center experience in Saudi Arabia. *International Journal of Infectious Diseases*, 29, December 2014. ISSN 1878-3511. doi: 10.1016/j.ijid.2014.09.003. URL `http://www.ncbi.nlm.nih.gov/pubmed/25303830`.

[21] Christian Drosten, Benjamin Meyer, Marcel a Müller, Victor M Corman, Malak Al-Masri, Raheela Hossain, Hosam Madani, Andrea Sieberg, Berend Jan Bosch, Erik Lattwein, Raafat F Alhakeem, Abdullah M Assiri, Waleed Hajomar, Ali M Albarrak, Jaffar a Al-Tawfiq, Alimuddin I Zumla, and Ziad a Memish. Transmission of MERS-coronavirus in household contacts. *The New England Journal of Medicine*, 371(9):828–35, August 2014. ISSN 1533-4406. doi: 10.1056/NEJMoa1405858. URL `http://www.ncbi.nlm.nih.gov/pubmed/25162889`.

[22] Frederick Hayden, Jeremy Farrar, and J S Malik Peiris. Towards improving clinical management of Middle East respiratory syndrome coronavirus infection. *Lancet Infectious Diseases*, 14(7), 2014.

[23] Neil M Ferguson and Maria D Van Kerkhove. Identification of MERS-CoV in dromedary camels. *The Lancet Infectious Diseases*, 14(2):93–4, February 2014. ISSN 1474-4457.

doi: 10.1016/S1473-3099(13)70691-1. URL http://www.ncbi.nlm.nih.gov/pubmed/24355867.

[24] Rita de Sousa, Chantal Reusken, and Marion Koopmans. MERS coronavirus: data gaps for laboratory preparedness. *Journal of Clinical Virology*, 59(1):4–11, January 2014. ISSN 1873-5967. doi: 10.1016/j.jcv.2013.10.030. URL http://www.ncbi.nlm.nih.gov/pubmed/24286807.

[25] The WHO MERS-CoV Research Group. State of Knowledge and Data Gaps of Middle East Respiratory Syndrome Coronavirus (MERS- CoV) in Humans. *PLoS Currents Outbreaks*, 2013. doi: 10.1371/currents.outbreaks.0bf719e352e7478f8ad85fa30127ddb8. Abstract.

[26] Robert Roos. Gush of MERS cases sparks speculation about causes, April 2014.

[27] David N Fisman and Ashleigh R Tuite. The epidemiology of MERS-CoV. *The Lancet infectious diseases*, 3099(13):13–14, November 2013. ISSN 1474-4457. doi: 10.1016/S1473-3099(13)70283-4. URL http://www.ncbi.nlm.nih.gov/pubmed/24239325.

[28] Anna Petherick. MERS-CoV: in search of answers. *The Lancet*, 381(9883):2069, June 2013. ISSN 01406736. doi: 10.1016/S0140-6736(13)61228-3. URL http://linkinghub.elsevier.com/retrieve/pii/S0140673613612283.

[29] Saudi Arabia recruits patients for vital MERS virus studies, June 2014. URL http://www.arabianbusiness.com/saudi-arabia-recruits-patients-for-vital-mers-virus-studies-556152.html.

[30] Romulus Breban, Julien Riou, and Arnaud Fontanet. Interhuman transmissibility of Middle East respiratory syndrome coronavirus: estimation of pandemic risk. *Lancet*, 382(9893):694–9, August 2013. ISSN 1474-547X. doi: 10.1016/S0140-6736(13)61492-0. URL http://www.ncbi.nlm.nih.gov/pubmed/23831141.

[31] S. Cauchemez, C. Fraser, M. D Van Kerkhove, C. Donnelly, S. Riley, A. Rambaut, V. Enouf, S van der Werf, and N. M Ferguson. Middle East respiratory syndrome coronavirus: quantification of the extent of the epidemic, surveillance biases, and transmissibility. *The Lancet Infectious Diseases*, 3099(13), November 2013. ISSN 14733099. doi: 10.1016/S1473-3099(13)70304-9. URL http://linkinghub.elsevier.com/retrieve/pii/S1473309913703049.

[32] World Health Organization. Middle East respiratory syndrome coronavirus (MERS-CoV) - update. Technical report, 2014. URL http://www.who.int/csr/don/2014_01_27mers/en/index.html\#.

[33] News - Update in Statistics Ministry of Health Institutes New Standards for Reporting of MERS-CoV.

[34] Ziad A. Memish, Alimuddin I. Zumla, Rafat F. Al-Hakeem, Abdullah A. Al-Rabeeah, and Gwen M. Stephens. Family Cluster of Middle East Respiratory Syndrome Coronavirus Infections. *New England Journal of Medicine*, 368(26):2487–2494, May 2013. ISSN 0028-4793. doi: 10.1056/NEJMoa1303729. URL http://www.nejm.org/doi/abs/10.1056/NEJMoa1303729.

[35] World Health Organization. WHO Risk Assessment: Middle East respiratory syndrome coronavirus (24 April). Technical report, 2014.

[36] G. Chowell, P.W. Fenimore, M.a. Castillo-Garsow, and C. Castillo-Chavez. SARS outbreaks in Ontario, Hong Kong and Singapore: the role of diagnosis and isolation as a control mechanism. *Journal of Theoretical Biology*, 224(1):1–8, September 2003. ISSN 00225193. doi: 10.1016/S0022-5193(03)00228-5. URL http://linkinghub.elsevier.com/retrieve/pii/S0022519303002285.

[37] Roy M. Anderson and Robert M. May. *Infectious Diseases of Humans Dynamics and Control*. Oxford University Press, Oxford, 1991.

[38] Herbert W Hethcote. The Mathematics of Infectious Diseases. *Society for Indusrial and Applied Mathematics*, 42(4):599–653, 2000.

[39] Gerardo Chowell, Carlos Castillo-Chavez, Paul W Fenimore, Christopher M Kribs-Zaleta, Leon Arriola, and James M Hyman. Model parameters and outbreak control for SARS. *Emerging infectious diseases*, 10(7):1258–63, July 2004. ISSN 1080-6040. doi: 10.3201/eid1007.030647. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3323341\&tool=pmcentrez\&rendertype=abstract.

[40] Samantha M Tracht, Sara Y Del Valle, and Brian K Edwards. Economic analysis of the use of facemasks during pandemic (H1N1) 2009. *Journal of Theoretical Biology*, 300:161–72, May 2012. ISSN 1095-8541. doi: 10.1016/j.jtbi.2012.01.032. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3307882\&tool=pmcentrez\&rendertype=abstract.

[41] Ziad A Memish, Matthew Cotten, Simon J Watson, Paul Kellam, Alimuddin Zumla, Rafat F Alhakeem, Abdullah Assiri, Abdullah a Al Rabeeah, and Jaffar A Al-Tawfiq. Community Case Clusters of Middle East Respiratory Syndrome Coronavirus in Hafr Al-Batin, Kingdom of Saudi Arabia: A Descriptive Genomic study. *International Journal of Infectious Diseases*, 23:63–8, June 2014. ISSN 1878-3511. doi: 10.1016/j.ijid.2014.03.1372. URL http://www.ncbi.nlm.nih.gov/pubmed/24699184.

[42] U. Buchholz, M. A. Müller, A. Nitsche, A. Sanewski, N. Wevering, F. Bonin, and C. Drosten. Contact investigation of a case of human novel coronavirus infection treated in a German hospital. *Eurosurveillance*, 18(8), 2013.

[43] World Health Organization. Global Alert and Response Disease Outbreak News, 2014. URL http://www.who.int/csr/don/en/.

[44] Kingdom of Saudi Arabia Ministry of Health. Novel Coronavirus, 2014. URL `http://www.moh.gov.sa/en/CoronaNew/PressReleases/Pages/default.aspx`.

[45] 2012-2014 List of MoH/WHO MERS nCoV Announced Cases, 2014. URL `http://flutrackers.com/forum/forum/novel-coronavirus-ncov-mers-2012-2014/`.

[46] Andrew Rambaut. MERS-Cases line list, 2014. URL `https://github.com/rambaut/MERS-Cases`.

[47] Census Highlights. Technical report, Kingdom of Saudi Arabia Central Department of Statistics & Information, 2004.

[48] Employed persons (15 Years and Over) By Age Group and Main Occupation Groups, 1433H-2012D. Technical report, Kingdom of Saudi Arabia Central Department of Statistics & Information, 2012.

[49] Labour Force (15 Years and Over) By Age Group and Sex, 1433H-2012D. Technical report, Kingdom of Saudi Arabia Central Department of Statistics & Information, 2012.

[50] Statistical Book for the Year 1433. Technical Report February, Kingdom of Saudi Arabia Ministry of Health, 2012.

[51] Noncommunicable diseases. Technical Report m, World Health Organization, 2010.

[52] Margaret Lethbridge-Cejku, Jeannine S Schiller, and Luther Bernadel. Summary health statistics for U.S. adults: National Health Interview Survey, 2002. Technical Report 222, National Center for Health Statistics, July 2002. URL `http://www.ncbi.nlm.nih.gov/pubmed/15791763`.

[53] Andrew C. Mills. Saudi Arabia: An Overview of Nursing and Health Care. *Focus on Critical Care*, 13(1):50–56, 1986.

Chapter 4

Modeling the impact of interventions on an epidemic of Ebola in Sierra Leone and Liberia

Manuscript authors: Caitlin M. Rivers, Eric T. Lofgren, Madhav Marathe, Stephen Eubank, Bryan L. Lewis

Citation: Rivers, C. M., Lofgren, E. T., Marathe, M., Eubank, S., & Lewis, B. L. (2014). Modeling the impact of interventions on an epidemic of Ebola in Sierra Leone and Liberia. PLoS Currents Outbreaks. doi: 10.1371/currents.outbreaks.4d41fe5d6c05e9df30ddce33c66d084c.

4.1 Forward

In March 2014, the World Health Organization announced what was to become the most severe outbreak in modern history. A cluster of 49 cases of Ebola virus disease (EVD) was identified in Guinea, thousands of miles west of where it had ever been seen before [1]. Every known EVD prior occurred in relatively rural, isolated communities in central Africa. The number of cases was usually limited to a few dozen, or a few hundred at most. For this reason, the appearance of Ebola in West Africa was largely overlooked as an epidemic threat. Even when the outbreak moved into Conakry, the capital city of Guinea, few recognized the danger to come.

The public health infrastructure in the region poorly equipped to handle the immense demands of an out-of-control outbreak. Across Liberia, Guinea and Sierra Leone, healthcare facilities were overwhelmed by an influx of patients. Severe and persistent shortages of per-

sonal protective equipment endangered workers. Contact tracing teams were hampered by poor road conditions and other difficulties. Strikes were implemented periodically by various critical workers to protest a lack of pay and unsafe working conditions.

By late summer, the outbreak had spread to neighboring Liberia and Sierra Leone, and case counts were soaring. Capital cities for each of the three affected countries were heavily affected, with widespread and intense transmission. An Ebola patient flew from Liberia to Lagos, Nigeria, introducing the disease to the most populated city in Africa. And despite growing panic, the global public health community was slow to acknowledge the severity of the situation, and even slower to coordinate a response. For months, the situation became more and more dire. International aid organizations finally began to deploy a response around September, but by then many months of exponential growth had yielded an outbreak out of control.

By October, the trajectory of the epidemic in the three countries began to diverge. Liberia, which had long had the most explosive growth, seemed to experience a slowing of new infections. Meanwhile Sierra Leone's case count continued to grow aggressively, and reports on the ground were grim. Guinea, always the slower-growing of the three, maintained a consistent course. It eventually became evident that the downturn in Liberia was real, although it is not yet known what is driving this turn of events.

Throughout the crisis, modeling has been key to better understanding the course of the outbreak. The first model published established the reproduction number to be well above the critical threshold of 1, and as high as 2.53 in Sierra Leone [2]. Numerous subsequent models agreed with that range (see for example [3, 4, 5, 6, 7]). Reproduction numbers are useful for understanding the epidemic potential, and were critical to communicating to policymakers the risk of the outbreak's growth.

A different use of models attempted to understand the hypothetical impact of various interventions. In addition to our study (the focus of this chapter), one study found that the number of treatment beds in Montserrado was not sufficient to meet demand[8]. Models were also developed to help public health officials outside of West Africa understand the risk of importation. A model published in early September predicted that Ghana, United Kingdom, Nigeria, Gambia and Ivory Coast were at highest risk of receiving an imported case via air travel [3]. As of January 2015, the United Kingdom is the only of those listed countries that has experienced a case (Nigeria had at that point already experienced an importation). A similar modeling effort by Bogoch et al determined Ghana, Senegal, the UK and France to be at highest risk. (Senegal did indeed experience an importation prior to the publication of that study, though not via air traffic) [9]. Other models assessing risk to China and Australia were also developed, giving guidance to their local governments [10, 6].

Contribution of this work

This project is interesting for several reasons. First, it is an example of using publicly available data to build models useful to public health responders. Instead of waiting for complete and cleaned data to be released after the outbreak ends, our models were built before much was known about the outbreak. We structured the model using a previously published Ebola model, and identified parameters from the literature where possible. For missing parameters, we used computational inference techniques to isolate the plausible parameter spaces. These decisions allowed us to have a working Ebola model in a matter of days, instead of weeks or months as might have otherwise been the case. This timeliness made the model useful while the outbreak was still unfolding.

This work is an example of using modeling capabilities to support decision makers in their public health response. Our ongoing working relationship with the Defense Threat Reduction Agency at the Department of Defense meant that our results were disseminated to people who could use them in hours or days, instead of months as would be the case with the traditional publication process. Furthermore, that relationship enabled the decision makers to make specific requests of the modelers, which guided us to questions that were more useful to operational response.

This research was published in *PLoS Currents Outbreaks* in October 2014.

4.2 Abstract

Background: An Ebola outbreak of unparalleled size is currently affecting several countries in West Africa, and international efforts to control the outbreak are underway. However, the efficacy of these interventions, and their likely impact on an Ebola epidemic of this size, is unknown. Forecasting and simulation of these interventions may inform public health efforts.

Methods: We use existing data from Liberia and Sierra Leone to parameterize a mathematical model of Ebola and use this model to forecast the progression of the epidemic, as well as the efficacy of several interventions, including increased contact tracing, improved infection control practices, the use of a hypothetical pharmaceutical intervention to improve survival in hospitalized patients.

Findings: Model forecasts until Dec. 31, 2014 show an increasingly severe epidemic with no sign of having reached a peak. Modeling results suggest that increased contact tracing, improved infection control, or a combination of the two can have a substantial impact on the number of Ebola cases, but these interventions are not sufficient to halt the progress of the epidemic. The hypothetical pharmaceutical intervention, while impacting mortality, had a smaller effect on the forecasted trajectory of the epidemic.

Interpretation: Near-term, practical interventions to address the ongoing Ebola epidemic may have a beneficial impact on public health, but they will not result in the immediate halting, or even obvious slowing of the epidemic. A long-term commitment of resources and support will be necessary to address the outbreak.

4.3 Introduction

West Africa is currently experiencing an unprecedented outbreak of Ebola, a viral hemorrhagic fever. On March 23, 2014 the World Health Organization announced through the Global Alert and Response Network that an outbreak of Ebola virus disease in Guinea was unfolding [11, 12, 1]. Ebola is generally characterized by sporadic, primarily rural outbreaks, and has not been seen before in West Africa, or in an outbreak of this size.

As of October 5, 2014, the World Health Organization has reported 8,033 cases of Ebola virus disease in Sierra Leone, Liberia, Guinea, Nigeria and Senegal, with sporadic cases occurring outside West Africa [13]. Considerable attention has been focused on preventing the outbreak from spreading further, either within Africa or intercontinentally. In principle many of the measures to contain the spread of Ebola, such as intensive tracing of anyone in contact with an infected individual and the use of personal protective equipment (PPE) for healthcare personnel treating infected cases, are straightforward [14]. However, implementing those interventions in a resource poor setting in the midst of an ongoing epidemic is far from simple, and subject to a great deal of uncertainty.

Mathematical models of disease outbreaks can be helpful under these conditions by providing forecasts for the development of the epidemic that account for the complex and non-linear dynamics of infectious diseases and by projecting the likely impact of proposed interventions before they are implemented. This in turn provides policy makers, the media, healthcare personnel and the public health community with timely, quantifiable guidance and support [15, 16, 17, 3].

We use a mathematical model to describe the development of the Ebola outbreak to date, provide short term projections for its future development, and examine the potential impact of several interventions, namely increased contact tracing, improved access to PPE for healthcare personnel, and the use of a pharmaceutical intervention to improve survival in hospitalized patients.

4.4 Methods

Outbreak data

A time series of reported Ebola cases was collected from public data released by the World Health Organization, as well as the Ministries of Health of the afflicted countries. These data sets do not include patient-level information, but rather laboratory confirmed, suspected or probable cases of the disease, which is thought to represent the best available estimate of the current state of the epidemic. A curated version of this data is available at https://github.com/cmrivers/ebola.

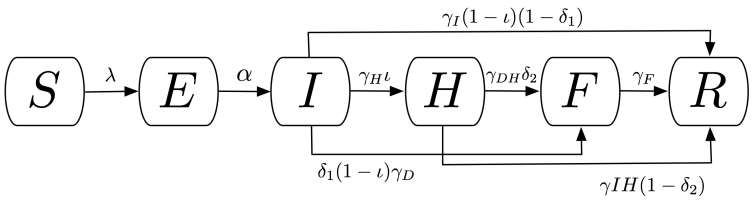

Figure 4.1: **Compartmental flow of a mathematical model of the Ebola Epidemic in Liberia and Sierra Leone, 2014.** The population is divided into six compartments: Susceptible (S), Exposed (E), Infectious (I), Hospitalized (H), Funeral (F) indicating transmission from handling a diseased patients body, and Recovered/Removed (R). Arrows indicate the possible transitions, and the parameters that govern them. Note that λ is a composite of all β transmission terms described in Table 4.1

.

A compartmental model was used to describe the natural history and epidemiology of Ebola, adapted from Legrand et al which was previously used to describe the 1995 Democratic Republic of Congo and 2000 Uganda Ebola outbreaks [18]. Briefly, the population is divided into six compartments, as shown in Figure 4.1. Susceptible individuals (S) may become Exposed (E) after contact with an infectious individual and transition in turn to the Infectious (I) class after the disease's incubation period, thereafter capable of infecting others. A proportion of these individuals may be Hospitalized (H). Both untreated patients in I and hospitalized patients in H may experience one of two outcomes: patients may die, with a chance of infecting others during the resulting funeral (F) before being removed from the model (R), or they may recover, at which point they are similarly removed. The system of ordinary differential equations describing this model is below.

$$\frac{dS}{dt} = -(\frac{\beta_I SI + \beta_H SH + \beta_F SF}{N}) \tag{4.1}$$

$$\frac{dE}{dt} = (\frac{\beta_I SI + \beta_H SH + \beta_F SF}{N}) - \alpha E \tag{4.2}$$

$$\frac{dI}{dt} = \alpha E - [\gamma_H \theta_1 + \gamma I(1-\theta_1)(1-\delta_1) + \gamma_D(1-\theta_1)\delta_1]I \tag{4.3}$$

$$\frac{dH}{dt} = \gamma_H \theta_1 I - [\gamma_{DH}\delta_2 + \gamma_{IH}(1-\delta_2)]H \tag{4.4}$$

$$\frac{dF}{dt} = \gamma_D(1-\theta_1)\delta_1 I + \gamma_{DH}\delta_2 H - \gamma_F F \tag{4.5}$$

$$\frac{dF}{dt} = \gamma_I(1-\theta_1)(1-\delta_1)I + \gamma_{IH}(1-\delta_2)H + \gamma_F F \tag{4.6}$$

Table 4.1: Model Parameters and Fitted Values for a Model of an Ebola Epidemic in Liberia and Sierra Leone, 2014.

Parameter	Liberia fitted values	Sierra Leone fitted values
Contact rate, community (βI)	0.160	0.128
Contact rate, hospital (βH)	0.062	0.080
Contact rate, funeral (βF)	0.489	0.111
Incubation period ($\frac{1}{\alpha}$)	12 days	10 days
Time until hospitalization ($\frac{1}{\gamma_H}$)	3.24 days	4.12 days
Time from hospitalization to death ($\frac{1}{\gamma_{DH}}$)	10.07 days	6.26 days
Duration of traditional funeral ($\frac{1}{\gamma_F}$)	2.01 days	4.50 days
Duration of infection ($\frac{1}{\gamma_I}$)	15.00 days	20.00 days
Time from infection to death ($\frac{1}{\gamma_D}$)	13.31 days	10.38 days
Time from hospitalization to recovery ($\frac{1}{\gamma_{IH}}$)	15.88 days	15.88 days
Fraction of infected hospitalized (θ_1)	0.197	0.197
Case fatality rate, unhospitalized (δ_1)	0.500	0.750
Case fatality rate, hospitalized (δ_2)	0.500	0.750

Model Fitting and Validation

A deterministic version of the model was fit and validated to the current outbreak data using least-squares optimization, with seed values from the Uganda outbreak described in

Legrand et al [18]. The last 15 days of reported cases were given one-quarter of the weight in the model to preferentially fit the most recent data. Based on anecdotal reports from the field (unpublished), candidate optimized fits were accepted such that roughly one-quarter of infections each came from contacts with hospitalized patients or funereal transmission, with the balance being from person-to-person spread within the community. The optimizer was further constrained to plausible parameter values, such as an upper bound of 20 days for infection duration, and 0 to 1 for probabilities or proportions. This model was fit only for Sierra Leone and Liberia, as the outbreak in Guinea has a unique epidemic curve which necessitates an alternative model design beyond the scope of this paper. The fitted parameters for this model, as well as their descriptions, may be found in Table 4.1.

This validated model provides a mathematical description of the epidemic up to the present. In order to forecast into the future, a stochastic version of the model was implemented using Gillespies algorithm with a tau-leaping approximation, which treats individuals as discrete units and converts the deterministic rates in the calibration model into probabilities, allowing random chance to come into play [19, 20]. Using the parameters from the calibration model, as well as the number of individuals in each compartment at the present date, 250 simulations of this model were run until December 31, 2014, giving a fan of potential epidemic trajectories that accounts for uncertainty in the forecast due to chance [21]. All models were implemented in Python 2.7, and the stochastic simulations used the StochPy library [22].

Modeled Interventions

Based on interventions that are technically, but not necessarily socially, feasible in the foreseeable future, we model five scenarios to examine their likely impact on the development of the epidemic. First, we model improved contact tracing by increasing the proportion of infected cases that are diagnosed and hospitalized from the baseline scenario of 51% in Liberia and 58% in Sierra Leone to 80%, 90% and 100%, and a concordant decrease in the time it takes for an infected individual to be hospitalized by 25%. This scenario could also be considered to represent improved access to healthcare, or improved public support for the hospitalization of sick individuals. Second, we explore the impact of simultaneously (1) decreasing the contact rate for hospitalized cases (β_H) to represent the increased use of PPE as supplies and awareness of the outbreak increase as well as (2) eliminating the possibility of post-mortem infection from hospitalized patients due to inappropriate funereal practices. Third, we model both a simultaneous (1) decrease in β_H (lack of post-mortem infection from hospitalized cases) and (2) increase in the proportion of hospitalized cases. This models the effect of a joint, intensified campaign to identify and isolate patients (the conventional means of containing an Ebola outbreak) with the necessary supplies and infrastructure to treat these patients using appropriate infection control practices. Finally, we model a pharmaceutical intervention that increases the survival rate of hospitalized patients by 25%, 50% and 75%, with a moderately high level of contact tracing (80%).

4.5 Results

Model fit and prediction

The deterministic model fit well for both Liberia and Sierra Leone, with the predicted curve of cumulative cases following the reported number of cases in both countries. The end-of-year forecast shows a range of uncertainty for each country of several thousand cases between the most optimistic and pessimistic scenarios. However, the number of cumulative cases is forecast to continue rising extremely rapidly, with the bulk of the epidemic yet to come. This suggests an extremely poor outlook for the course of the epidemic without intensive interventions.

In the baseline end-of-year forecasts for both Sierra Leone and Liberia, person-to-person transmission within the community made up the bulk of transmission events, with a median (IQR: interquartile range) of 117,877 (115,100 120,585) cases arising from the community in the Liberia forecast and $30,611$ (29,667 31,857) in the forecast for Sierra Leone. Both had fewer hospital transmissions $21,533$ (21,025 21,534) in Liberia and $5,474$ (5,306 5,710) in Sierra Leone, than transmissions arising from funerals $35,993$ (35,163 36,789) in Liberia and $9,768$ (9,470 10,137) in Sierra Leone. For brevity, only the results of the Liberia model are reported below, with the results from Sierra Leone in the electronic supplement. The epidemic trajectories for all modeled interventions may also be found in the Appendix.

Basic Reproduction Number

The basic reproduction number (R_0) for the baseline scenario was calculated in the same manner as in Legrand et al [18]. Briefly, R_0 is broken into three components, representing the respective contributions of community, hospital and funereal transmissions, as well as an overall R_0 reflecting the epidemic potential for the disease. In the baseline scenario, we estimate an overall R_0 of 2.22, made up of an R_0 of 1.35 from the community, 0.35 from hospitals and 0.53 from funerals for Liberia. Sierra Leones R_0 was estimated to be 1.78, made up of an R_0 of 1.11 from the community, 0.24 from hospitals and 0.43 from funerals. These estimates are similar to estimates for the current outbreak reported elsewhere. For brevity, only the overall R_0 estimates will be reported here. The breakdown of R_0 by source can be found in the Appendix.

Intensified Contact Tracing and Infection Control

The forecasted distribution of cases under intensified contact tracing is shown in Figure 4.4. There is a shift from community transmission toward hospital transmission, though at extremely high levels of contact tracing and hospitalization, the impact of the intervention on

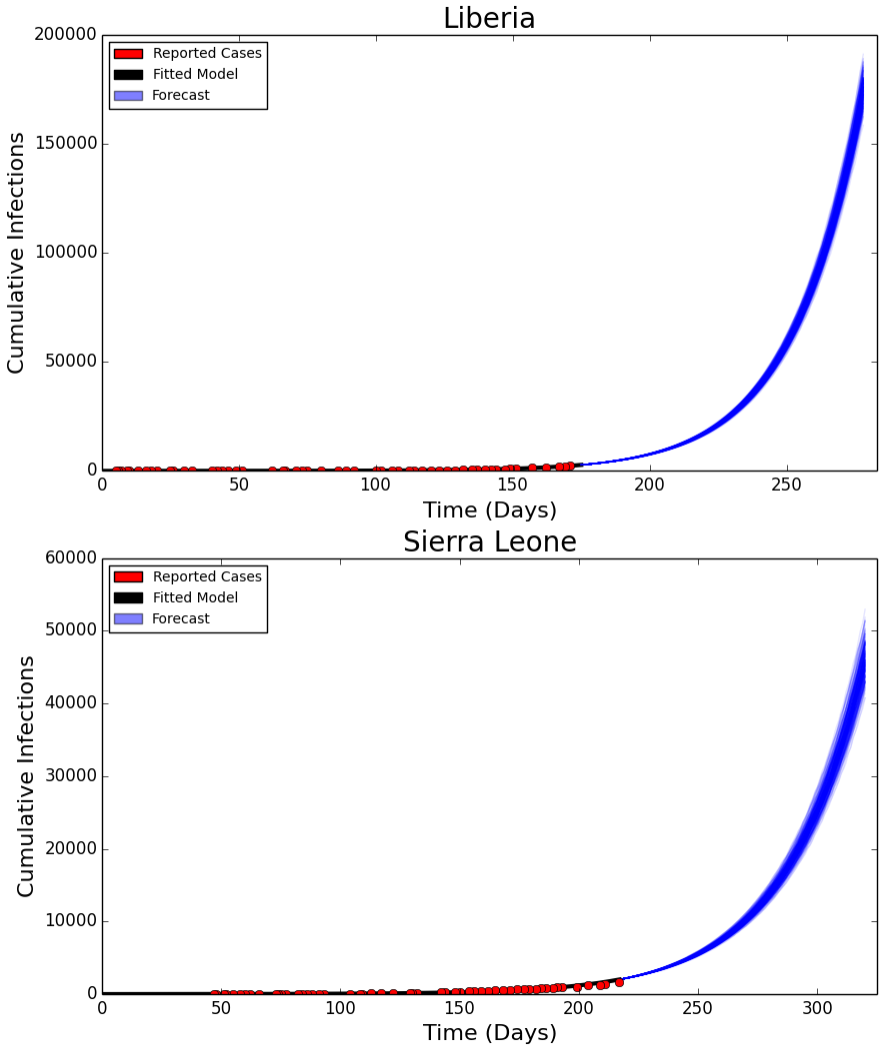

Figure 4.2: **Fitted Compartmental Model for Ebola Epidemic in Liberia and Sierra Leone, 2014, with 250 Iterations of a Stochastic Forecast to December 31, 2014.** Red dots depict the reported number of cumulative cases of Ebola in each country, with the black line indicating the deterministic model fit. Each blue line indicates one of two hundred and fifty stochastic simulated forecasts of the epidemic, with areas of denser color indicating larger numbers of forecasts.

the course of the outbreak also results in fewer hospitalized cases. There is a less pronounced but still substantial downward shift in funeral cases, and a decrease in total cases in Liberia.

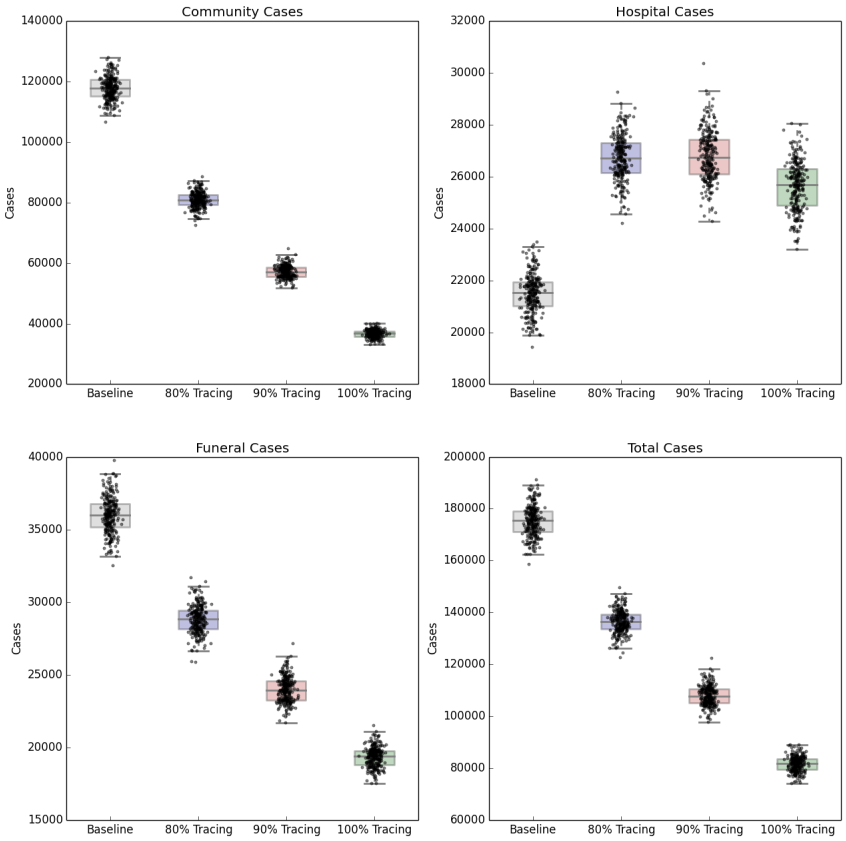

Figure 4.3: **Distribution of Forecast Cases of Community, Hospital, Funeral and Total Cases for Ebola Epidemic, Liberia, 2014, at Baseline, 80%, 90% and 100% of Patients Traced and Hospitalized.** Box plots depict the median, interquartile range and 1.5 times the interquartile range for each scenario. Each individual simulated forecast is shown as a single dot, jittered so as to depict the complete distribution of the data.

The improved infection control scenario decreased β_H to represent decreased risk of hospital transmission due to increased PPE, increased number of healthcare workers or greater awareness of the epidemic resulting in greater care while treating patients with undiagnosed febrile illness. Additionally, it eliminated the potential for post-mortem transmission during the funereal process. This combination of interventions resulted in a marked decrease in the overall number of cases, and a reduction of R_0 to 2.13, 2.05 and 1.96 for 25%, 50% and 75% reductions in the hospital transmission contact rates and improvements in the disposal of the remains of Ebola victims. This decrease is not sufficient to shift the cumulative case

curve off its steep upward trajectory, but only lessens its magnitude. The 80%, 90% and 100% of patients traced and hospitalized scenarios resulted in an overall reduction of R_0 to 2.11, 2.01 and 1.89 respectively.

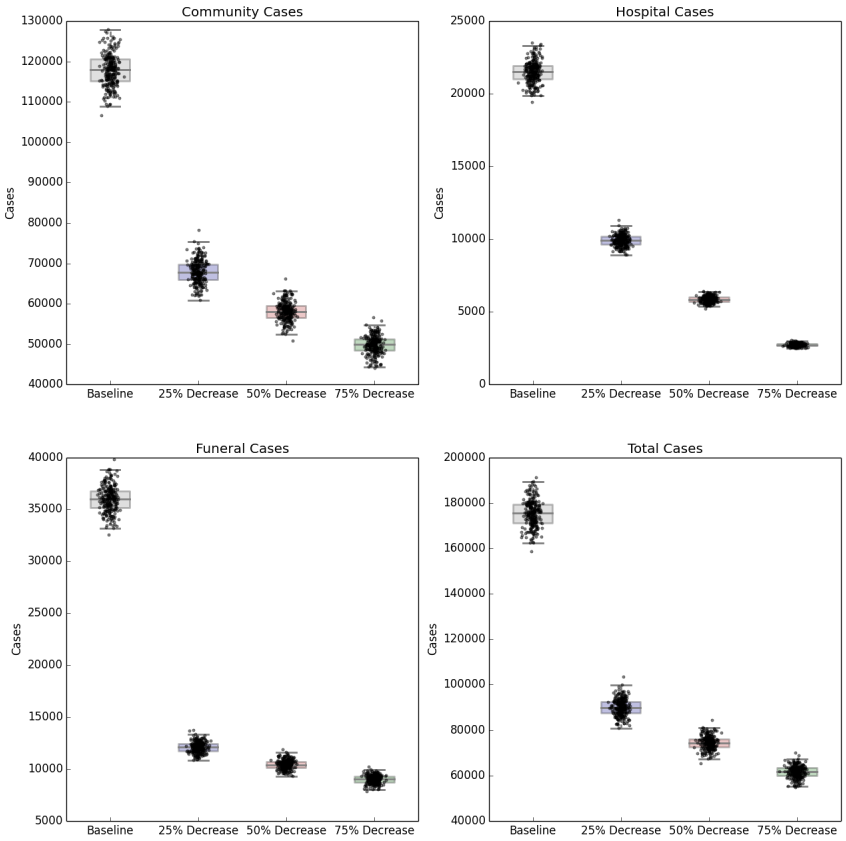

Figure 4.4: **Distribution of Forecast Cases of Community, Hospital, Funeral and Total Cases for Ebola Epidemic, Liberia, 2014, at Baseline, 25%, 50% and 75% Reductions in Hospital Transmission Contact Rates (β_H).** Box plots depict the median, interquartile range and 1.5 times the interquartile range for each scenario. Each individual simulated forecast is shown as a single dot, jittered so as to depict the complete distribution of the data.

Major reductions in all sources of cases were seen, with the most dramatic drop in the relative number of cases arising from the reduction of within-hospital transmissions. However, even with substantially reduced transmission and a decrease in the burden of mortality from the outbreak, improved infection control was also insufficient to push the epidemic off its steep

upward trajectory.

Figure 4.5 shows the median decrease in cases as compared to baseline for simulations combining increased contact tracing and a reduction in the risk of hospital transmission from those who are isolated and treated. The most optimistic of these scenarios, with complete contact tracing and a 75% reduction in hospital transmission results in more than $165,000$ fewer total cases over the course of the forecasted period, as compared to the baseline scenario (Figure 4.6). The overall R_0 in this scenario is reduced to 1.72. This represents a ten-fold reduction in the number of cases, and is a major improvement to the epidemic trajectory. However even under this scenario, the epidemic is slowed and mitigated, rather than fully stopped, with transmission still occurring after the end of the year.

Increased Availability of Pharmaceutical Interventions

The introduction of a pharmaceutical intervention that dramatically improves the survival rate of hospitalized patients also leads to a less severe outbreak, shown in Figure 4.7. Compared to contact tracing alone, there is a small reduction in the number of hospitalized cases (as the scenario implies no change in infection control practices), but a stronger decrease in the number of community, funeral and overall cases depending on the efficacy of the hypothetical pharmaceutical. An efficacy that reduces the case fatality rate of hospitalized patients by 25%, 50% or 75% results in a corresponding reduction of R_0 to 2.03, 1.94 and 1.85 respectively. As with the other forecasts above, this intervention also fails to halt the progress of the epidemic, though it does considerably reduce the burden of disease.

4.6 Discussion

The control of Ebola outbreaks in the past has been a straightforward, albeit difficult application of infection control and quarantine policies. In principle, these types of interventions should be applicable to this outbreak as well. However, it remains unclear whether they can be implemented at the unprecedented scale of the current outbreak. This study attempts to address whether or not aggressive interventions could arrest, or at least mitigate, the epidemic.

Our findings suggest that, for at least in the near term, some form of coordinated intervention is imperative. The forecasts for both Liberia and Sierra Leone in the absence of any major effort to contain the epidemic paint a bleak picture of its future progress, which suggests that we are in the opening phase of the epidemic, rather than near its peak. These findings are in line with predictions from other models which, despite using different methods and different data sources, have all estimated similar basic reproductive numbers, and forecast that the epidemic is currently beyond the point where it can be easily controlled [3, 23, 7].

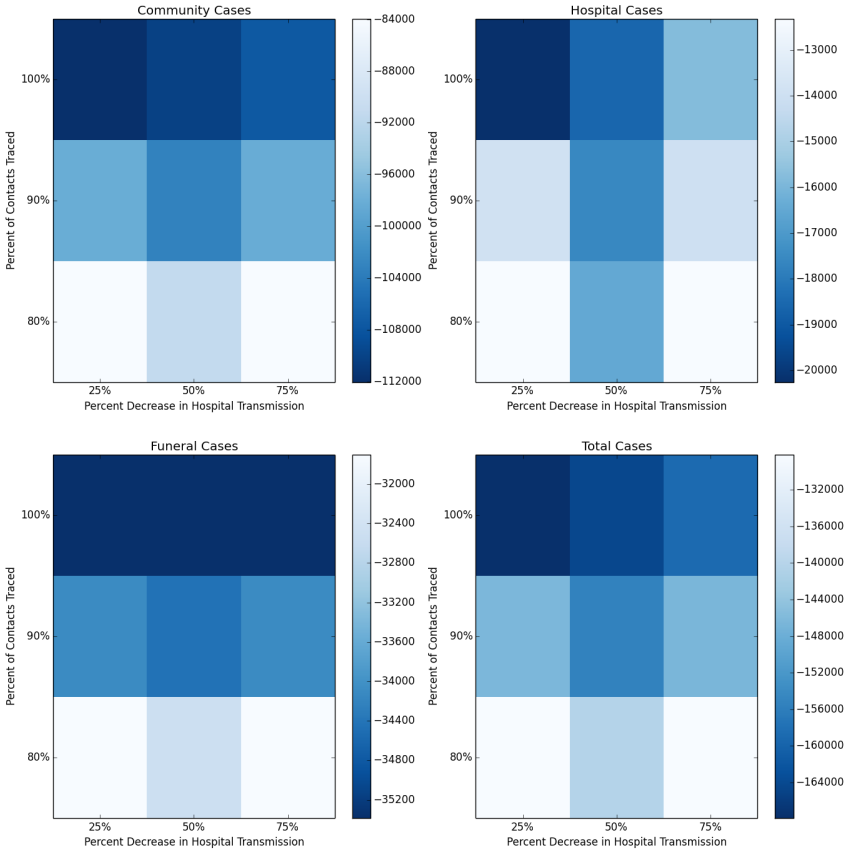

Figure 4.5: **Distribution of Forecasted Cases of Community, Hospital, Funeral and Total Cases for Ebola Epidemic, Liberia, 2014, at Baseline, 25%, 50% and 75% Reductions in Hospital Transmission Contact Rates (β_H) with 80%, 90% and 100% of Patients Traced and Hospitalized.** Each box represents the median result of 250 forecasted epidemics, each with a % of contacts traced and a % decrease in hospital transmission. Areas of deeper blue indicate progressively greater reductions of the median number of cases.

Of the modeled interventions applied to the epidemic, the most effective by far is a combined strategy of intensifying contact tracing to remove infected individuals from the general population and placing them in a setting that can provide both isolation and dedicated care. This intervention requires that clinics have the necessary supplies, training and personnel to follow infection control practices. Although both of these interventions in isolation also have an impact on the epidemic, they are much more effective in parallel. In particular, the

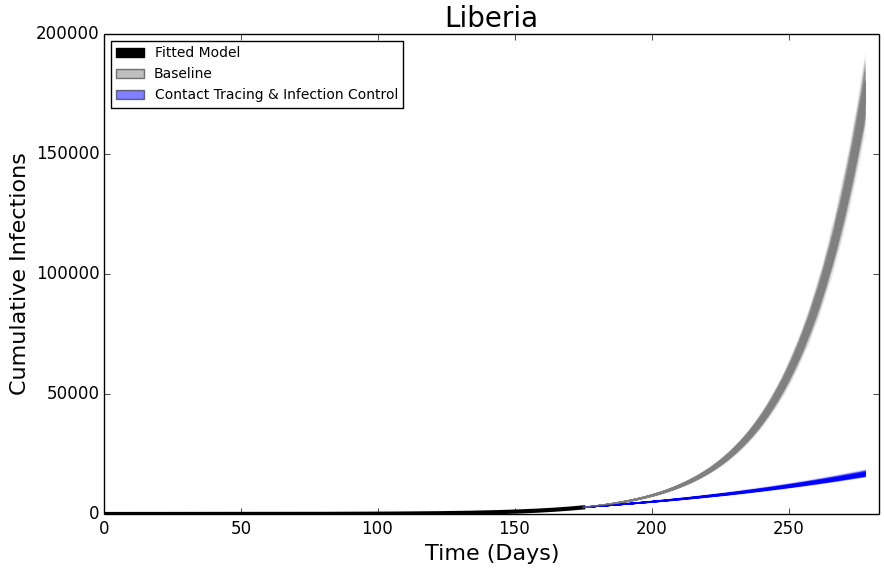

Figure 4.6: **Forecasted Cumulative Cases for Ebola Epidemic, Liberia, 2014 with 75% Reduction in Hospital Transmission Contact Rates (β_H) with 100% of Patients Traced and Hospitalized.** The solid black line represents the deterministic model fit of the epidemic to present, with each grey line representing a single simulated forecast with no interventions in place, and each blue line representing a single simulated forecast of the epidemic with 100% of contacts traced, a 75% reduction in hospital transmission (β_H) and no post-mortem infections from hospitalized patients. Areas of darker color indicate more forecasts with that result.

slight increase in cases in Sierra Leone at the lower end of the modeled contact tracing range, when unaccompanied by a concordant increase in infection control, highlights the necessity of these two interventions being implemented side-by-side.

The hypothetical mass application of a novel pharmaceutical like the one administered to two American aid workers had a much smaller impact on the course of the epidemic itself. While certainly lessening the burden of mortality of those infected (in the most optimistic scenario modeled, reducing the case-fatality rate from 50% to 12.5%), the downstream effects of such an intervention are relatively minor, as there is no suggestion that any candidate treatments have a substantial impact on transmission. This will impact not only the evaluation of those treatments, which should thus focus primarily on their patient-level efficacy, but also in communicating to the media and the public that, despite the use of these drugs, the benefits that will be seen from them are in decreases to mortality due to infection, not in a halt to the epidemic itself.

Despite the considerable impact the proposed interventions have on the burden of disease, none of them are forecast, at least in the short term, to halt the epidemic entirely. It is

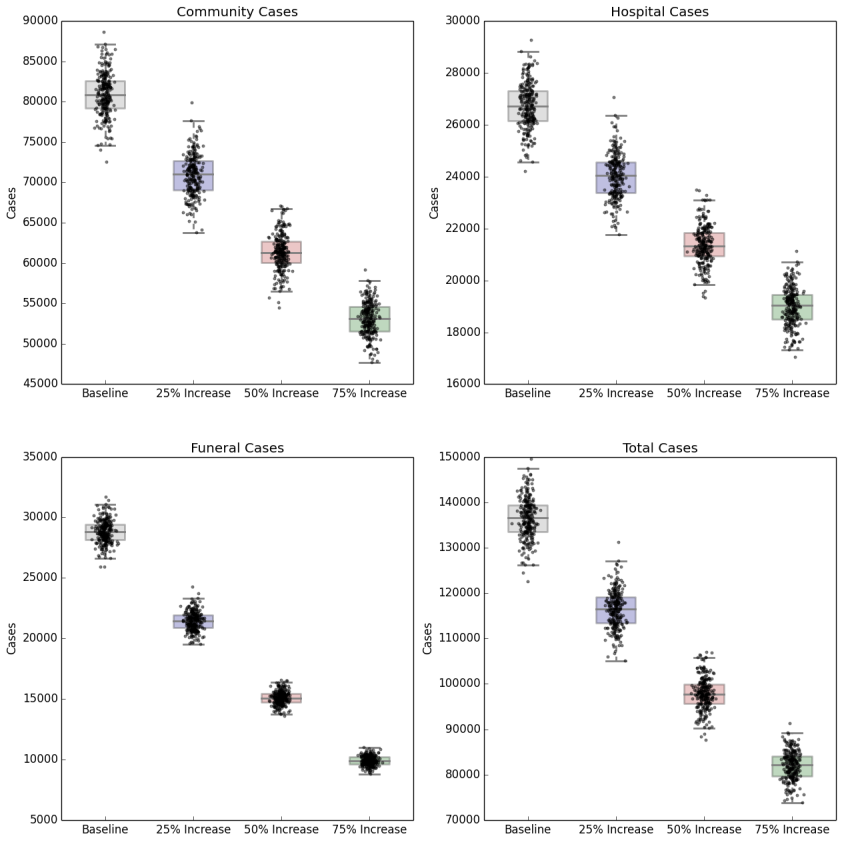

Figure 4.7: **Distribution of Forecast Cases of Community, Hospital, Funeral and Total Cases for Ebola Epidemic, Liberia, 2014, at Baseline, 25%, 50% and 75% Reductions in Case Fatality Rate Due to a Hypothetical Pharmaceutical Intervention.** Box plots depict the median, interquartile range and 1.5 times the interquartile range for each scenario. Each individual simulated forecast is shown as a single dot, jittered so as to depict the complete distribution of the data.

possible these interventions will have a longer-term impact on the epidemic, however we have avoided projecting out further than the end of the year due to the inherent uncertainty in an emerging epidemic. This in turn suggests another communication challenge for public health planners. While all of the proposed interventions are worth pursuing, and will have an impact on the epidemic and public health, the attention of the international community must be sustained in the long term in order to ensure the necessary supplies and expertise remain present in the affected areas. Additionally, in light of public resistance to the limited

types of these interventions already in place, public awareness and acceptance of intensified interventions must be built, as there will not be an immediate cessation to the epidemic, or even necessarily a clear sign that the situation is improving.

This study is not without limitations. As with all mathematical models, the results of the study depend on the assumptions about the natural history of Ebola, its epidemiology, and the values of the parameters used as well as the quality of the data used to fit the model. Particularly, this model is validated against data that records cases on their time of reporting, rather than their time of onset, so the model time series may be shifted by several days. Because of the relatively small number of historical Ebola outbreaks, the unusual size of this outbreak, and the difficulty collecting data during an emerging epidemic, the accuracy of this model and its parameters is difficult to ascertain. It does however represent our best understanding of the epidemic using available data, and the model has proven capable of predicting the ongoing development of the epidemic, as well as having been used to model previous Ebola outbreak. The uncertainty inherent in model prediction has been addressed with the short forecasting window, and with the use of stochastic simulation to aid in quantifying uncertainty inherent within the model system.

The ongoing Ebola epidemic in West Africa demands international action, and the results of this study support that many of the interventions currently being implemented or considered will have a positive impact on reducing the burden of the epidemic. However, these results also suggest that the epidemic has progressed beyond the point wherein it will be readily and swiftly addressed by conventional public health strategies. The halting of this outbreak will require patient, ongoing efforts in the affected areas and the swift control of any further outbreaks in neighboring countries.

4.6.1 Technical note

Based on comments from readers of the manuscript, we believe it is valuable to explore the way multiple competing pathways for an individual within the Legrand model, and by extension the model used in this paper, are handled. In a generic SEIR model, the system of differential equations is given by:

$$\frac{dS}{dt} = -\beta_I SI \tag{4.7}$$

$$\frac{dE}{dt} = \beta_I SI - \gamma E \tag{4.8}$$

$$\frac{dI}{dt} = \gamma E - \delta I \tag{4.9}$$

$$\frac{dR}{dt} = \delta I \tag{4.10}$$

This can be thought of as a deterministic (or stochastic, when simulated using a stochastic simulation algorithm) version of a Markov process, moving people from E to I and I to R with the corresponding probabilities on each step. In this interpretation, the distribution for residence time in each state is exponential, with means of γ^{-1} and δ^{-1} respectively.

Because the model used in this manuscript has multiple potential transitions for two of the compartments (I and H), it uses a somewhat more complex transition scheme that is not as intuitive. Rather than the transition between two states being a single process governed by a single residence time parameter, the people transitioning out of a compartment are split into the fractions following each potential "path" out of that compartment, each with its own corresponding residence time. This is shown below in Figure 4.8 with both processes broken out, and Figure 4.9 with the processes shown together.

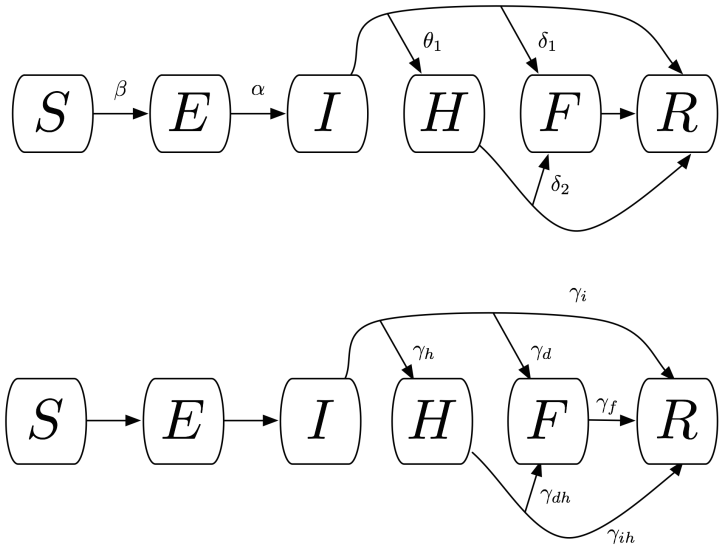

Figure 4.8: The fractions of people distributed across different paths through the compartments (top) and their respective transition rates (bottom).

In effect, while technically within the same compartment, patients who will experience different outcomes are treated as being on distinct, independent pathways. This is a conceptually different process than individuals all having the possibility of each outcome, and transitioning between them based entirely on waiting times. As such, the mean residence times estimated by the model are not weighted means, but the means for each particular path, where the "outcome" of that path is already known. The composition of a system of equations that does provide weighted means is relatively straightforward. However, the authors have elected to use the same system as Legrand et al. to maintain consistency with much of the existing

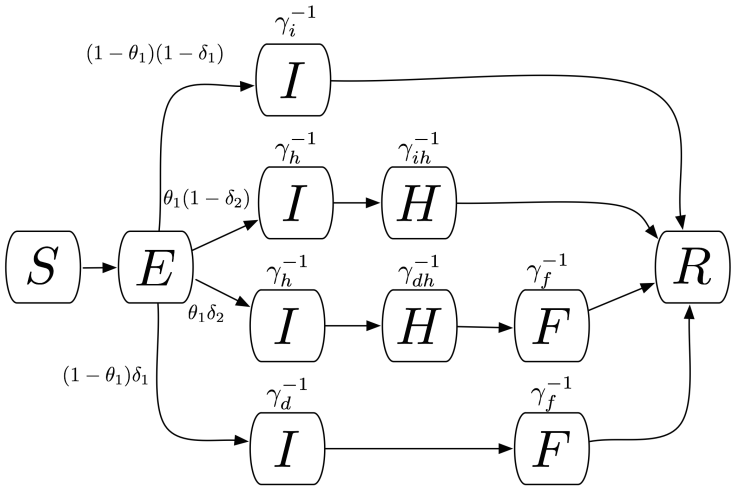

Figure 4.9: Expanded view of the model used in this manuscript, with each path labeled by its overall probability and the transition rates represented by the mean residence times associated with each compartment in each path.

Ebola literature.

Acknowledgements

The authors would like to thank Alessandro Vespignani and Nina Fefferman for their comments, and Katie Dunphy, Jesse Jeter, P. Alexander Telionis, James Schlitt, Jessie Gunter and Meredith Wilson for their assistance and support. The authors would also like to acknowledge the participants in the weekly briefings organized by DTRA, BARDA and NIH, including Dave Myer, Aiguo Wu, Mike Phillips, Ron Merris, Jerry Glashow, Dylan George, Irene Eckstrand, Kathy Alexander and Deena Disraelly for their comments and feedback.

4.6.2 Funding statement

This work was funded by NIH MIDAS Grant 5U01GM070694-11 and DTRA Grant HDTRA1-11-1-0016 and DTRA CNIMS Contract HDTRA1-11-D-0016-0001. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The authors have declared that no competing interests exist.

Subsequent literature

The forecasts in this work were generated in September; the manuscript was published three weeks later in *PLoS Currents Outbreaks*. At the time of the study, the Ebola outbreak in Sierra Leone and Liberia was still in the initial growth phase, which was approximately exponential. By mid-October, there were indications that the outbreak was slowing down. In Liberia in particular, the actual observed number of cases was consistently below the forecasted projections. By late October, the same trend was seen in Sierra Leone, although to a lesser degree.

At first it was not clear if the downturn was real, or if the surveillance and reporting capacity of the affected countries was saturated. Very limited information was available on the setup of the in-country surveillance systems, and even less was known about the performance of those systems. To an outside observer, the case counts would behave the same in either scenario; both would result in either a decline or near complete cessation of new cases reported. However, as several more weeks passed it became clear that the epidemic really was slowing.

At the time of this writing, the weekly incidence of cases in Liberia has fallen to less than 100 and has maintained that decline from early September, when peak weekly incidence was around 350 cases. In Sierra Leone however, progress has been slower. The epidemic curve appears to still be on the climb as weekly incidence continues to grow. The country has had between 300-500 cases per week every week since mid-September. Guinea has yet another pattern, characterized by comparatively few but steady case counts. Weekly incidence is generally less than 150, but intensity waxes and wanes periodically. Analysis at the district level indicates that hot spots of active mini-outbreaks are continually flaring up and then dying down, only to reappear elsewhere in the country.

Given current epidemic conditions, the forecasts generated for publication overestimated the future number of cases. There are several contributing factors. The interventions modeled could only capture two to three combinations of interventions at the most. During an active outbreak, it would be unthinkable to limit the number of interventions applied. In an epidemic response situation, every intervention available is deployed in combination, making the downstream effects on incidence potentially much more significant. Second, the impact of social distancing and adaptive human behavior was not modeled, because it is very difficult to quantify and represent in the models even under the best of circumstances. There is some evidence that in Liberia, the country with the most dramatic initial growth and then subsequent control, changes in human behavior were primarily responsible for the epidemics downturn[24].

Approximately six weeks after this model was developed, the World Health Organization Ebola Response team published an analysis of the patient databases from the affected countries in the *New England Journal of Medicine*. The manuscript contained the true values for many epidemiological parameters that we had estimated in our model, among quite a lot

else. In the absence of the necessary data, we had used an optimization routine, the Nelder-Mead simplex algorithm implemented using the fmin function in the Python package scipy, to navigate the parameter space. The optimizer minimized the least squares between cumulative case counts and modeled cumulative counts. A discussion of the technical drawbacks of this approach can be found in the limitations section of this thesis.

Comparison of the parameters we derived with data from the patient databases revealed our values to be quite suitable. Other modeling efforts relied on parameter values in previously published literature, which underestimated the incubation period and duration of infection by nearly half, as seen in Table 4.2. This comparison suggests that when data from the field is not available, computer-guided parameterization can produce sufficiently close value estimates to allow for early modeling.

Table 4.2: Comparison of Ebola model parameters to WHO reported values

Source	Incubation	Onset to recovery	Onset to death	Onset to hospitalization	R_0
Liberia WHO	11.7	15.4	7.9	4.9	1.8
Liberia VBI	12.0	15.0	13.3	3.2	2.2
Sierra Leone WHO	10.8	17.2	8.6	4.6	2.2
Sierra Leone VBI	10.0	20.0	10.4	4.1	1.8
Gomes	7.0	10.0	9.6	5.0	1.8
Althaus	5.3	5.6			1.6-2.5
Towers	5, 7, 10	5, 7, 10			1.3-2.1
Meltzer	6	6			

4.6.3 Social context of modeling as outbreak response

Despite the accuracy of the early forecasts and their role in identifying and confirming a change in the dynamics of the epidemic, doubts surfaced. Our group and many others involved in similar work received questions about why our models failed to produce accurate long term forecasts, and what their utility could be given their short time horizon. One high profile piece was published as a news item in *Nature* by author Declan Butler. Our group, together with colleagues from around the country, wrote a rebuttal to Butler, and to others who were unclear on the role of modeling. Our piece was published as Correspondence in *Nature*, and is reproduced here in full. A shortened version appeared in print. The extensive author list can be found online. A piece on the similar topic of why employ modeling in outbreak scenarios, was published shortly thereafter by lead author Eric Lofgren in *Proceedings of the National Academy of Sciences*.

> In "Models overestimate Ebola cases", Declan Butler asserts that models of the Ebola epidemic have "failed to accurately project the outbreak's course".

This misrepresents the role of epidemic modeling in outbreak response.

Models will never provide perfect forecasts, and do not claim to. They provide quantitative insight into possible futures of the outbreak under certain conditions. Unlike weather forecasts, where humans cannot change the course of events, epidemic forecasts stimulate adaptive behaviors that change the outbreaks course. This is in part why early forecasts sometimes overestimate the future number of cases. By analogy, stock market forecasting influences investor behavior and in doing so sets the markets on a new path.

The very models Dr. Butler describes as 'failed' helped inspire and inform the strong international response that may have led to the slowing of the epidemic seen today. We consider this a success. Later models assessed the potential impact of various public health interventions and policy decisions. As those interventions were implemented and as behaviors changed, case counts that diverged from the modeled baseline were early indicators that the outbreak response was having an impact.

Without these insights, there is very little to guide decision makers in the midst of an outbreak other than intuition. Sociobiological processes like epidemics are affected by countless unobserved variables, and uncertainty is a given. But without any rigorous attempt to synthesize available information, policy makers may only rely on their own internal model of how the epidemic will develop. These "gut instinct" models suffer from the same problems, but without any of the transparency or clarity of more formal models. Instead of portraying Ebola models as having 'missed the mark', we encourage a closer inspection of the importance of models beyond providing forecasts.

4.7 Bibliography

[1] World Health Organization. Ebola virus disease in Guinea, March 2014. URL `http://www.afro.who.int/en/clusters-a-programmes/dpc/epidemic-a-pandemic-alert-and-response/outbreak-news/4063-ebola-hemorrhagic-fever-in-guinea.html`.

[2] Christian L Althaus. Estimating the Reproduction Number of Ebola Virus (EBOV) During the 2014 Outbreak in West Africa. *PLoS Currents Outbreaks*, 2014.

[3] Marcelo F. C. Gomes, Ana Pastore y Piontti, Luca Rossi, Dennis Chao, Ira Longini, M. Elizabeth Halloran, and Alessandro Vespignani. Assessing the International Spreading Risk Associated with the 2014 West African Ebola Outbreak. *PLoS Currents Outbreaks*, 2014. ISSN 2157-3999. doi: 10.1371/currents.outbreaks.

cd818f63d40e24aef769dda7df9e0da5. URL http://currents.plos.org/outbreaks/?p=40803.

[4] David Fisman, Edwin Khoo, and Ashleigh Tuite. Early Epidemic Dynamics of the West African 2014 Ebola Outbreak: Estimates Derived with a Simple Two-Parameter Model. *PLoS Currents Outbreaks*, 2014. ISSN 2157-3999. doi: 10.1371/currents.outbreaks.89c0d3783f36958d96ebbae97348d571. URL http://currents.plos.org/outbreaks/?p=41147.

[5] Sherry Towers, Oscar Patterson-Lomba, and Carlos Castillo-Chavez. Temporal Variations in the Effective Reproduction Number of the 2014 West Africa Ebola Outbreak. *PLoS Currents Outbreaks*, (1), 2014. ISSN 2157-3999. doi: 10.1371/currents.outbreaks.9e4c4294ec8ce1adad283172b16bc908. URL http://currents.plos.org/outbreaks/?p=42655.

[6] Tanja Stadler, Denise Kühnert, and David A Rasmussen. Insights into the Early Epidemic Spread of Ebola in Sierra Leone Provided by Viral Sequence Data. *PLoS Currents Outbreaks*, 2014.

[7] Martin I Meltzer, Charisma Y Atkins, Scott Santibanez, Barbara Knust, Brett W. Petersen, Elizabeth D. Ervin, Stuart T. Nichol, Inger K. Damon, and Michael L. Washington. Estimating the Future Number of Cases in the Ebola Epidemic Liberia and Sierra Leone , 2014 2015. *Morbidity and Mortality Weekly Report*, 63(3):2014–2015, 2014.

[8] Joseph a Lewnard, Martial L Ndeffo Mbah, Jorge a Alfaro-Murillo, Frederick L Altice, Luke Bawo, Tolbert G Nyenswah, and Alison P Galvani. Dynamics and control of Ebola virus transmission in Montserrado, Liberia: a mathematical modelling analysis. *The Lancet Infectious Diseases*, 3099(14), October 2014. ISSN 14733099. doi: 10.1016/S1473-3099(14)70995-8. URL http://linkinghub.elsevier.com/retrieve/pii/S1473309914709958.

[9] Isaac I Bogoch, Maria I Creatore, Martin S Cetron, John S Brownstein, Nicki Pesik, Jennifer Miniota, Theresa Tam, Wei Hu, Adriano Nicolucci, Saad Ahmed, James W Yoon, Isha Berry, Simon Hay, Aranka Anema, Andrew J Tatem, Derek MacFadden, Matthew German, and Kamran Khan. Assessment of the potential for international dissemination of Ebola virus via commercial air travel during the 2014 west African outbreak. *The Lancet*, 385(9962):29–35, October 2014. ISSN 01406736. doi: 10.1016/S0140-6736(14)61828-6. URL http://linkinghub.elsevier.com/retrieve/pii/S0140673614618286.

[10] Shi Chen, Brad J White, Michael W Sanderson, David E Amrine, Amiyaal Ilany, and Cristina Lanzas. Highly dynamic animal contact network and implications on disease transmission. *Nature Scientific Reports*, 4:4472, January 2014. ISSN 2045-2322. doi: 10.

1038/srep04472. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3966050\&tool=pmcentrez\&rendertype=abstract.

[11] Sylvain Baize, Delphine Pannetier, Lisa Oestereich, Toni Rieger, Lamine Koivogui, N'faly Magassouba, Barrè Soropogui, Mamadou Saliou Sow, Sakoba Keïta, Hilde De Clerck, Amanda Tiffany, Gemma Dominguez, Mathieu Loua, Alexis Traoré, Moussa Kolié, Emmanuel Roland Malano, Emmanuel Heleze, Anne Bocquin, Stephane Mély, Hervé Raoul, Valérie Caro, Dániel Cadar, Martin Gabriel, Meike Pahlmann, Dennis Tappe, Jonas Schmidt-Chanasit, Benido Impouma, Abdoul Karim Diallo, Pierre Formenty, Michel Van Herp, and Stephan Günther. Emergence of Zaire Ebola Virus Disease in Guinea - Preliminary Report. *The New England Journal of Medicine*, pages 1–8, April 2014. ISSN 1533-4406. doi: 10.1056/NEJMoa1404505. URL http://www.ncbi.nlm.nih.gov/pubmed/24738640.

[12] Sylvie Briand, Eric Bertherat, Paul Cox, Pierre Formenty, Marie-Paule Kieny, Joel K. Myhre, Cathy Roth, Nahoko Shindo, and Christopher Dye. The International Ebola Emergency. *The New England Journal of Medicine*, 371:13, August 2014. ISSN 1533-4406. doi: 10.1056/NEJMp1409903. URL http://www.ncbi.nlm.nih.gov/pubmed/25140858.

[13] World Health Organization. WHO : Ebola Response Roadmap Situation Report (8 October 2014). Technical Report October, 2014.

[14] IDSA Ebola Guidance, 2014. URL http://www.idsociety.org/2014_ebola/.

[15] Nell Greenfieldboyce. A Virtual Outbreak Offers Hints of Ebola's Future, August 2014. URL http://www.npr.org/blogs/health/2014/08/14/340346575/a-virtual-outbreak-offers-hints-of-ebolas-future.

[16] Kai Kupferschmidt. Disease modelers project a rapidly rising toll from Ebola, August 2014. URL http://news.sciencemag.org/health/2014/08/disease-modelers-project-rapidly-rising-toll-ebola.

[17] Denise Grady. U.S. Scientists See Long Fight Against Ebola, September 2014. URL http://www.nytimes.com/2014/09/13/world/africa/us-scientists-see-long-fight-against-ebola.html.

[18] J Legrand, R F Grais, P Y Boelle, a J Valleron, and A Flahault. Understanding the dynamics of Ebola epidemics. *Epidemiology and Infection*, 135(4): 610–21, May 2007. ISSN 0950-2688. doi: 10.1017/S0950268806007217. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2870608\&tool=pmcentrez\&rendertype=abstract.

[19] DT Gillespie. Exact stochastic simulation of coupled chemical reactions. *Journal of Physical Chemistry*, 81:2340–2361, 1977.

[20] DT Gillespie. Approximate accelerated stochastic simulation of chemically reacting systems. *The Journal of Chemical Physics*, 115(4):1716–1733, 2001.

[21] Eric T Lofgren. Visualizing results from infection transmission models: a case against confidence intervals. *Epidemiology*, 23:738–741, 2012.

[22] Timo R Maarleveld, Brett G Olivier, and Frank J Bruggeman. StochPy: a comprehensive, user-friendly tool for simulating stochastic biological processes. *PloS One*, 8(11):e79345, January 2013. ISSN 1932-6203. doi: 10.1371/journal.pone.0079345. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3832486\&tool=pmcentrez\&rendertype=abstract`.

[23] WHO Ebola Response Team. Ebola Virus Disease in West Africa - The First 9 Months of the Epidemic and Forward Projections. *The New England Journal of Medicine*, September 2014. ISSN 1533-4406. doi: 10.1056/NEJMoa1411100. URL `http://www.ncbi.nlm.nih.gov/pubmed/25244186`.

[24] Abhishek Pandey, Katherine E Atkins, Jan Medlock, Jeffrey P Townsend, James E Childs, and G Tolbert. Strategies for containing Ebola in West Africa. *Science*, 16 (October), 2014.

Chapter 5

Open epidemiology for outbreak response

5.1 Background

The first half of this decade has brought many challenges in emerging infectious disease surveillance and control. Outbreaks of influenza H7N9, chikungunya, Middle East Respiratory Syndrome Coronavirus, and Ebola virus disease, among others, have underscored the need for robust international public health. Detecting and understanding these events in a timely manner is more important now than ever before in the history of public health. However, the same interconnectedness that turns a local health event into a global concern, makes information exchange even faster. These challenges have also brought opportunities. The same technological developments that make local epidemics global threats also allow an unprecedented exchange of information. Data can be published to the internet even as it is being collected. Scientists, public health professionals, clinicians, and local stakeholders can converse regardless of geographic distance and language barriers.

In the years since the advent of digital epidemiology, methods for conducting this work have emerged. These methods take advantage of the numerous sharing platforms that allow researchers to share and discuss their work online in real time. Data are typically aggregated from online public sources like Census records and outbreak situation reports. These data are typically in HTML or PDF format, so they must made machine readable either by hand or using software like Tabula. For data compilations that require some degree of inference like assembling line listings, researchers often compare data sets to sort out confusion and minimize uncertainty. It is helpful if researchers then share these machine-readable sets on services like Github or Figshare, to minimize duplicative efforts. Preliminary results are then shared through social networking sites like Twitter, or through personal blog posts. Active sharing of early results is a departure from traditional research procedure, where findings

are generally finalized and published before they are released.

There are several common concerns that arise when discussing open data. Some people worry that publishing the data will lead to criticism of the public health department, which could be demoralizing or distracting to those trying to do their jobs. While distracting from the outbreak response is certainly best avoided, withholding data is not protection against that outcome. Journalists, third party stakeholders, and local community members are all just as likely to criticize the outbreak managers as data consumers. And within the bounds of respectful professionalism, that outside assessment is usually in the best interest of the outbreak outcome as a whole. Accountability is rarely a mistake.

Another common concern is that preparing the data for publication is time-consuming and will draw responders away from their other responsibilities. This concern is understandable - engaging in these activities take time. However, the data are surely already being prepared for regular analysis internally to guide the response efforts. Second, there are innumerable groups and individuals who would be willing to aid with the data preparation effort, often for free. The time-sink component is an obstacle that can be handled. And finally, a shift in thinking must occur away from data provision as a non-critical activity to one of the central functions of outbreak responders. Although data publication does not directly impact the outbreak in the way that e.g. contact tracing does, on a larger scale it almost certainly has significant downstream effects.

Similarly, some are worried that because the source data are messy, publishing it will misinform onlookers, leading them to draw incorrect conclusions. A similar argument proposes that data might be used by modelers to build models that draw the wrong conclusions, or project inaccurate forecasts. On the other hand, publishing no data at all is more likely to lead to that outcome, as people scramble to piece together whatever scraps of insight they can uncover. It is widely known that outbreak data is messy; end-users understand that, and methodologies exist to account for that uncertainty. Furthermore, the more eyes that are on the data, the more opportunities there are to identify and correct biases or errors. Crowdsourcing the data quality improvement is much more efficient than leaving it to people who have other critical responsibilities, and probably less of the needed skill set.

These concerns highlight the gaps the open epidemiology community need to address in order to bring open data into the mainstream. Some efforts have been made towards this end during the course of this dissertation work. The following example is a case study of the accuracy of Ebola forecasts using publicly available data, compared to private patient database data.

5.2 Case study: Ebola

Beginning in late summer, Sierra Leone and Liberia published near-daily situation reports on their website. The 'sitreps' contained case counts and deaths by county, and sometimes

additional information like the number of contacts under followup, and laboratory reporting capacity. For the duration of the outbreak I converted those situation reports from their original PDF format to machine-readable format. I posted them on the internet regularly so that other researchers involved in the Ebola response could better access the data. As of this writing, the repository remains the only place on the internet that the digitized data is available. The repository received two to three thousand views a day at its peak, and has maintained hundreds of views per day even during relative lulls.

These sitreps were the only information about the outbreak available to people outside the Ministries of Health of the affected countries and the WHO. It was the primary source of data used to inform the response operations for every governmental and nongovernmental organization involved in the outbreak. The digitized copies allowed users to work with the data directly without having to first convert it by hand. Although this manual conversion is time intensive, sharing it meant that it only had to be done once, so the work was not duplicated across every stake holding organization.

The response I received from users of the data I posted underscores the crucial importance of open data in public health events. Disaster response is not limited to the local organizations with official responsibility to respond. Many external groups and stakeholders also deploy. Without open data, they have nothing to guide or inform their actions. And even for the local clinics and public health departments who do have direct access to the source data, analytical capacity is usually in short supply, if not non-existent. Providing open data makes the response more tractable as a whole, and is a critical part of designing effective interventions.

As the case of the Ebola sitreps demonstrated, publishing the data online is a useful and necessary, but not sufficient, step. A more useful choice than providing it in PDF is to also provide a digital copy. The data are not stored at their source in PDF format, so there is no reason to believe providing a machine-readable version would be additional work for the groups who own it. Furthermore, as the outbreak progressed, old source data on the Ministries of Health website was removed, so in many cases my repository is the only record of that data that exists publicly.

However, the open data is not without limitations. Sitrep data reflects only the report date of new cases; there is no information available on when those cases fell ill. In some cases, the onset date for newly reported cases may have been many days or weeks ago. A comparison of the time series shows that as expected, the time series from the sitrep data lags behind the patient database, seen in Figure 5.1. Sitrep data represents report date, or the time when cases were reported to the Ministries of Health. The patient database records onset date of the disease, which is a more accurate representation of the natural history of the outbreak.

In order to better understand the limitations of making disease forecasts using open source data, I compared our forecast performace of our previously-described Ebola model in Sierra Leone using patient database data, and open sitrep data. Model structure and optimization routines are described in Chapter 4. Parameters describing the clinical course of the disease,

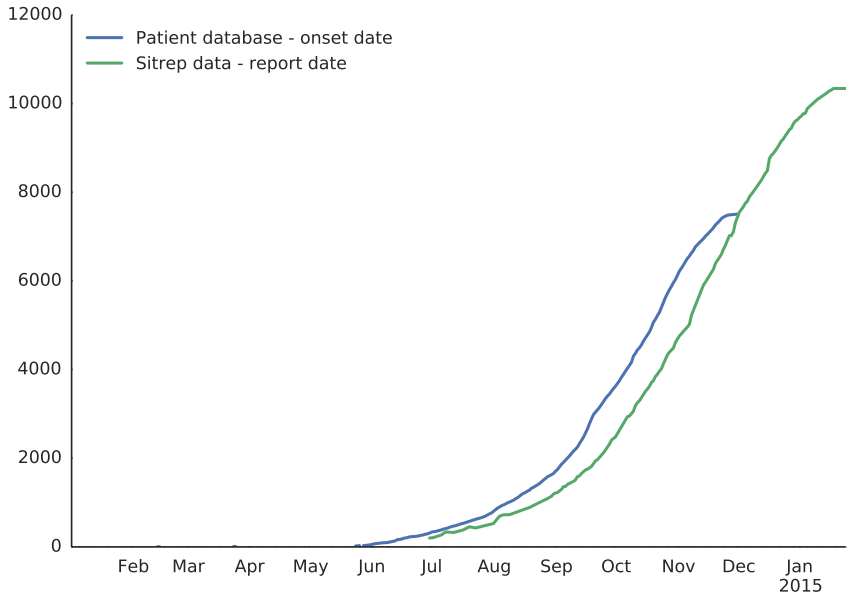

Figure 5.1: **Comparison of cumulative incidence in Sierra Leone.** Cumulative incidence of publicly available sitrep data compared to cumulative incidence of onset dates from the patient database.

like incubation period, were held constant between the models; the only parameter modified was the optimization routine that identifies the β parameters for each of the three infectious compartments. Six forecasts were generated, using successively more data. The first forecast used data up to July 18, shortly after the outbreak was first recognized in Sierra Leone. Successive forecasts used 21 days more data each. These forecasts were compared to the 'ground truth' time series of onset date incidence from the patient database.

Results in Figures 5.2, 5.3 and 5.4 show that sitrep data actually does a better job generating accurate forecasts in the earliest days of the outbreak; patient database forecasts drastically underestimated the future number of cases. Although the exact mechanism for surveillance and reporting in the affected countries is unknown, it's possible that the sitrep data stream is more inclusive when counting possible Ebola cases, and therefore early sitrep tallies overestimated the number of Ebola cases in the country. This overestimate may have inadvertently bolstered the accuracy of the forecasts by indicating exponential growth before that growth was actually established in the country. The corresponding result is that although sitrep data did a good job producing forecasts in the short term, long term forecasts overestimated the expected future number of cases.

As the outbreak progressed and exponential growth ended, the two models performed comparably. Both did excellent through the initial growth phase. After exponential growth

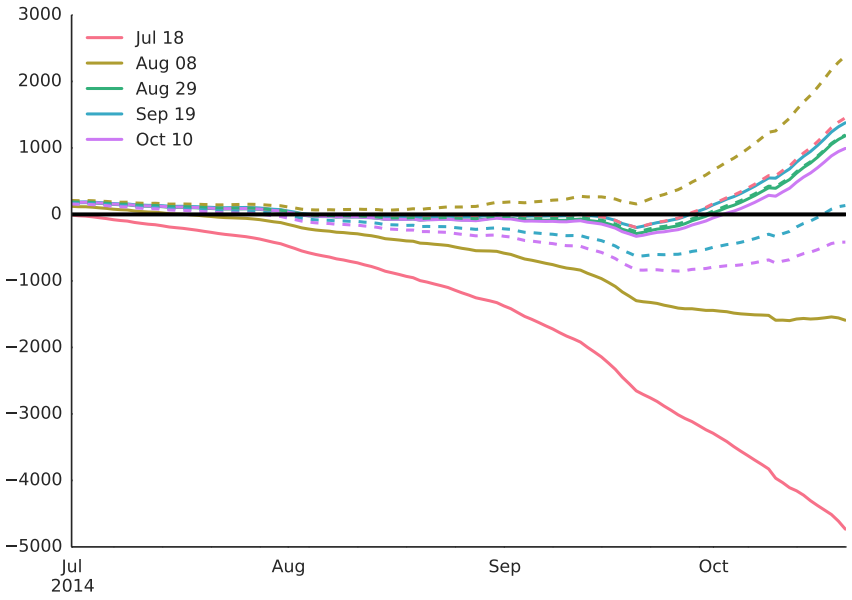

Figure 5.2: **Difference of forecasted cases and actual cases in Sierra Leone.** Solid lines represent forecasts derived from the patient database model; dashed lines indicate forecasts from the sitrep model.

ended, open data models under-predicted later case counts, and patient database models over-predicted. The relative success of each model at identifying the approximate number of future cases supports the value of models as tools to help inform decision makers about the expected course of the outbreak. Furthermore, the performance of the open data model in particular is testament to the importance and utility of open data to support real-time outbreak epidemiology.

5.3 Open data resources

It is not enough to consume data as an open epidemiologist. To fully develop the ecosystem, researchers need to contribute data, tools and resources of their own. As a complement to the work described in this dissertation, I developed an open source Python package called epipy that contains tools for open epidemiologists to use for manipulating epidemiology data. A description of epipy, and in particular a tool I developed to reconstruct and visualize transmission chains of emerging zoonoses is included below. Epipy is available for download for free at https://www.cmrivers.github.io/epipy.

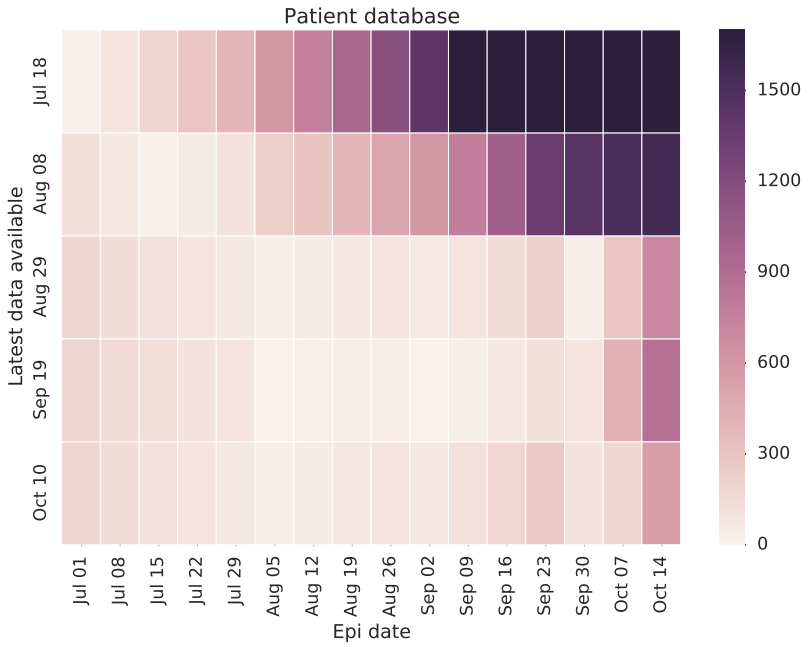

Figure 5.3: **Absolute difference of forecasted and actual values from models parameterized using patient database data.**

Abstract

We present case tree plots and checkerboard plots for visualizing contagions. The visualizations are best suited for diseases like SARS, MERS-CoV and H7N9 for which there are a limited (less than 200) number of cases, with data available on human to human transmission. They a) allow for easy estimation of epidemiological parameters like basic reproduction number b) indicate the frequency of introductory events, e.g. spillovers in the case of zoonoses c) represent patterns of case attributes like patient sex both by generation and over time.

Introduction

Zoonoses represent an estimated 58% of all human infectious diseases, and 73% of emerging infectious diseases [1]. Careful tracking of zoonotic disease is a major focus of global public health protection strategy. Recent examples of zoonotic outbreaks include Severe Acute Respiratory Syndrome, H1N1, and Middle East Respiratory Syndrome, which have caused thousands of deaths combined [2, 3, 4]. Early identification of new outbreaks is critical to successful containment of these diseases.

The current toolkit for visualizing data from these emerging diseases is limited. One popular option is the epidemic curve, which is a histogram of new cases over time. Epidemic curves

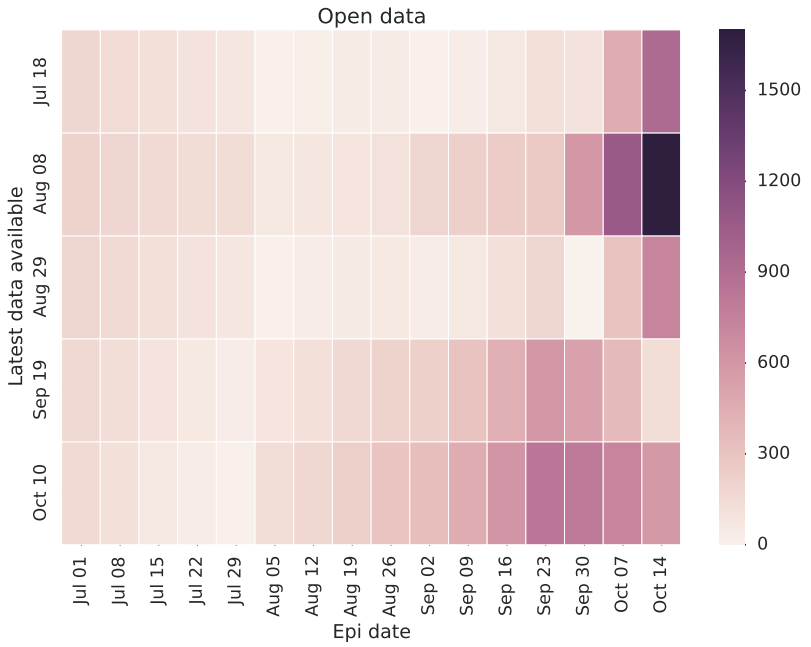

Figure 5.4: **Absolute difference of forecasted and actual values from model parameterized using publicly available sitrep data.**

are limited in that they do not indicate how cases are related to one another, nor can they represent the presence of an animal source. Network diagrams are a useful though less popular option. These diagrams can depict individual human clusters, but often do not have a time component, and cannot represent constellations of unconnected clusters. Furthermore, network diagrams typically require complete information about the structure of the transmission tree. Here we introduce case tree plots and checkerboard plots to address those weaknesses and more clearly represent zoonotic outbreaks.

Description

We present two new visualizations, case tree plots and checkerboard plots, for visualizing emerging zoonoses. Code for the plots are available in the open source python package epipy, which is available on github. Epipy relies heavily on the networkx [5] and pandas [6] packages. In addition to the visualizations introduced here, epipy includes a number of functions for common epidemiology calculations, like odds ratio and relative risk. A function that generates realistic example data is also provided. All plots, data and tables in this manuscript were generated using epipy.

Case ID	Onset date	Cluster ID
1	2013-01-20	FamilyA
2	2013-01-29	FamilyA
3	2013-02-10	HighSchool
4	2013-02-12	
5	2013-02-08	Family A
6	2013-02-14	HighSchool
7	2013-02-22	High School

Table 5.1: An example line list for case tree plot construction

Case tree plots

Case tree plots depict the emergence and growth of clusters of disease over time. Each case is represented by a colored node. Nodes that share an epidemiological link are connected by an edge. The meaning of the color of the node varies based on the node attribute chosen by the plot creator; in many cases, color simply signifies membership to a human to human cluster. However, it could also represent health status (e.g. alive, dead), the sex of the patient, or any other categorical attribute.

Node placement along the x-axis corresponds with the date of illness onset for the case. When the onset date is not known, diagnosis date may be used instead. The y-axis value represents the case generation. Nodes at generation zero are human cases acquired from an animal source. If that infected human passes the disease to two other humans, those two subsequent cases are plotted at generation one. Cases that do not belong to a cluster are not represented on the plot.

To generate a case tree plot, users provide a line list with, at minimum: unique case identifiers, the date of illness onset (or the date the illness was reported, if onset date is not available), and cluster membership, as seen in table 5.1. Any additional relevant variables like patient age and sex may also be included.

Users must also provide the mean and standard deviation of the generation time between cases. Because generation time is not always known in the early days of the outbreak, the incubation period may be a reasonable proxy. The line listing need not specify the chain of transmission; the plot generator will estimate the chain of transmission based on the onset dates. Cases labeled as belonging to the same cluster that have an onset date within one standard deviation of the mean generation time are assumed to be linked.

Checkerboard plots

We have also developed a second visualization, the checkerboard plot, to complement case tree plots. Checkerboard plots, seen in figure 5.6, show how cases in human to human clusters

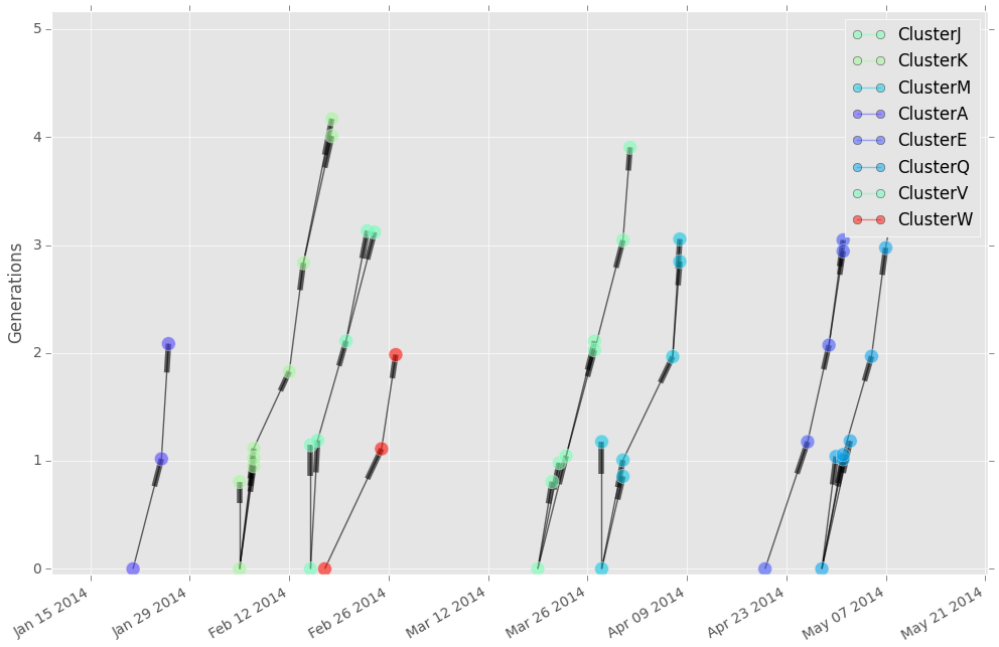

Figure 5.5: **Case tree plot using example outbreak data.**

have arisen over time. They can be used in conjunction with case tree plots, or in situations where representing a hypothetical network structure is inappropriate. Checkerboard plots have the added benefit of showing the unique identifying number for the first and last cases in each cluster, to more easily connect the visualization to the line list.

Each colored block represents a case in the cluster; the first and last cases in each cluster are labeled with their respective case identifiers, so they may more easily be found on the line list. Like case tree plots, the placement of each colored block along the x-axis corresponds with the case's date of illness onset or diagnosis. Cases in the same cluster with onset dates close to each other may overlap. The plot can be generated using the same line listing as described for case tree plots.

Also packaged with epipy are functions to analyze the outbreak structure as understood in case tree plots. Users may generate summary statistics and histograms of case attributes, by generation. For example, figure 5.7 shows the distribution of patient sex by generation for the outbreak depicted in figure 5.5. Users may also calculate the reproduction number for each node in the graph, and produce summary statistics and histograms for the reproduction number of the outbreak as a whole (excluding cases that are not part of any cluster).

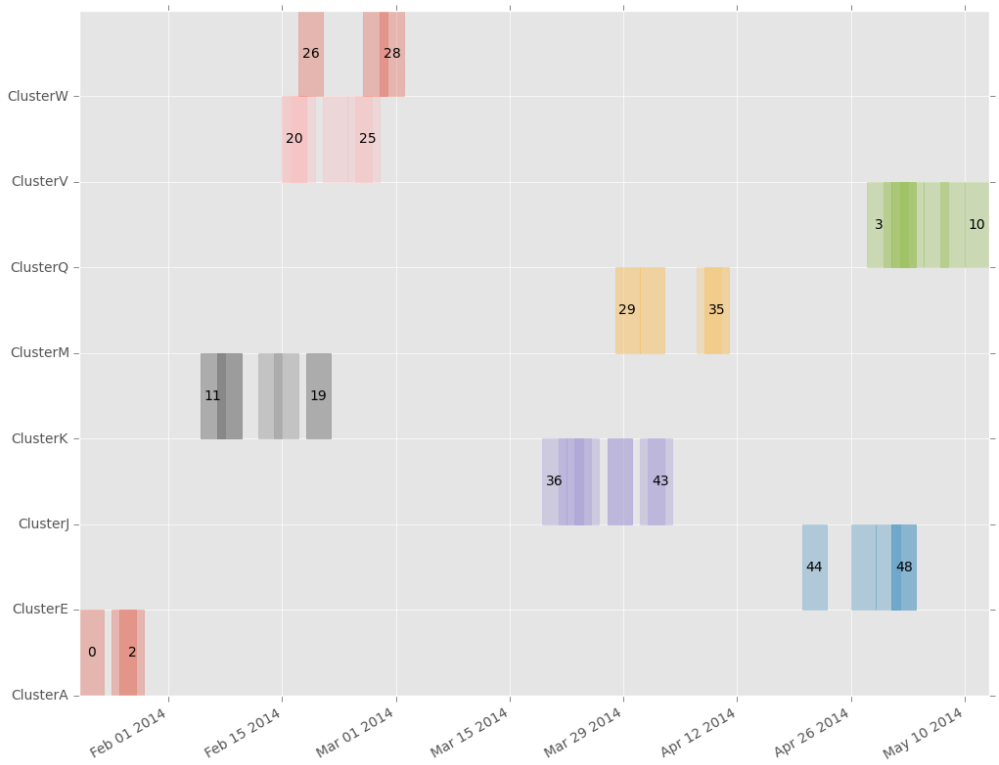

Figure 5.6: **Checkerboard plot using example outbreak data.**

Discussion

Case tree and checkerboard plots can assist epidemiologists with visualizing and analyzing zoonotic diseases with human to human potential. The plots provide valuable insight into the dynamics of an outbreak not available using current visualization tools.

The basic reproductive number is easily visualized by evaluating the size and shape of each tree in the case tree plot over time. Functions available in epipy can further assist with basic reproduction number calculations. Clusters of trees that remain short and thin likely have poor human to human transmissibility, whereas trees that become progressively taller and wider as the outbreak progresses may be gaining in transmissibility. Case tree plots also allow for quick identification of trees with an unusally large number of branches. This observation is useful for identifying superspreaders, who can play a critical role in accelerating an outbreak, as was the case in the SARS outbreak [7].

The rate at which new trees emerge is useful for estimating spillover frequency, which in turn is useful for identifying animal hosts. Numerous trees over a short period of time may suggest a domesticated or agricultural animal host rather than a wild animal. As case tree plots become more common, patterns that indicate certain emergence scenarios wil likely be identified, similar to how the shape of epidemic curves can differentiate a point source

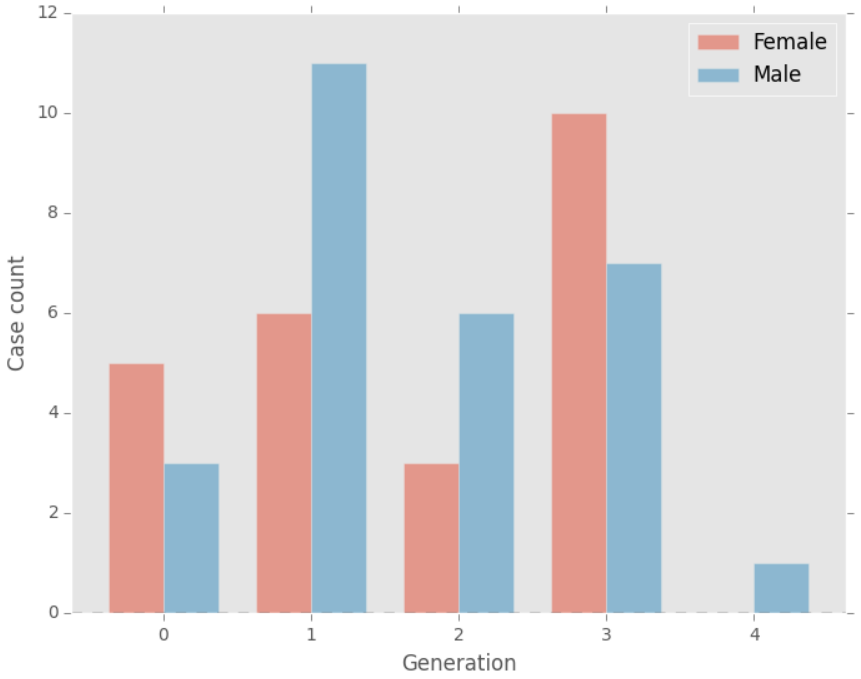

Figure 5.7: **Case counts by generation and patient sex.**

outbreak from a continuous source outbreak.

Finally, case tree plots and accompanying analysis functions describe case attributes by generation and by cluster, which can reveal new insights. For example, a disproportionate number of male index nodes could suggest that exposure to the disease occurs in a male-dominated setting like deer hunting. A high case fatality risk among index nodes compared to non-index nodes suggests that the infection is less virulent when transmitted from person to person than when acquired from an animal source. Revealing this pattern is useful for understanding the epidemiology of the disease, and for identifying effective interventions.

Here we introduce two new plots for representing infectious disease outbreaks. Case tree plots depict spillover events and subsequent transmission of human to human cases of a zoonotic disease over time. Checkerboard plots also represent case clusters over time, but do not attempt to construct transmission trees. These plots visually represent important outbreak dynamics, like the basic reproduction number. They also allow for underutilized analyses like case fatality risk stratified by generation. These plots provide epidemiologists with additional means to visualize and understand zoonotic disease.

5.4 Conclusions

Through open epidemiology, research can be translated into effective interventions in days or weeks, instead of months or years. Opportunities exist to shorten the interval between outbreak detection and control, even in the most remote parts of the world. The best way to get ahead of an outbreak is to facilitate that exchange as much as possible. Researchers, public health professionals, and the public can distribute data, analyses, and recommendations in an instant. They can also self-organize, so that the major public health organizations are not solely responsible for outbreak preparedness and response. In order for this end to be realized, more work is needed to address concerns around data sharing, and digital infrastructure needs to be developed further. The coming years will undoubtedly bring changes in these areas, which I anticipate will benefit global public health immensely.

5.5 Bibliography

[1] Mark E J Woolhouse and Sonya Gowtage-Sequeria. Host range and emerging and reemerging pathogens. *Emerging infectious diseases*, 11(12):1842–7, December 2005. ISSN 1080-6040. doi: 10.3201/eid1112.050997. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3367654\&tool=pmcentrez\&rendertype=abstract`.

[2] World Health Organization. Middle East respiratory syndrome coronavirus (MERS-CoV) - update. Technical report, 2014. URL `http://www.who.int/csr/don/2014\_01\_27mers/en/index.html\#`.

[3] Guillermo Domínguez-Cherit, Stephen E Lapinsky, Alejandro E Macias, Ruxandra Pinto, Lourdes Espinosa-Perez, Alethse de la Torre, Manuel Poblano-Morales, Jose a Baltazar-Torres, Edgar Bautista, Abril Martinez, Marco a Martinez, Eduardo Rivero, Rafael Valdez, Guillermo Ruiz-Palacios, Martín Hernández, Thomas E Stewart, and Robert a Fowler. Critically Ill patients with 2009 influenza A(H1N1) in Mexico. *JAMA : the journal of the American Medical Association*, 302(17):1880–7, December 2009. ISSN 1538-3598. doi: 10.1001/jama.2009.1536. URL `http://www.ncbi.nlm.nih.gov/pubmed/19822626`.

[4] Michael D Christian, Susan M Poutanen, Mona R Loutfy, Matthew P Muller, and Donald E Low. Severe acute respiratory syndrome. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 38(10):1420–7, May 2004. ISSN 1537-6591. doi: 10.1086/420743. URL `http://www.ncbi.nlm.nih.gov/pubmed/15156481`.

[5] Aric Hagberg, Pieter Swart, and Daniel Schult. Exploring network structure, dynamics and function using NetworkX. In *SciPy*, 2008.

[6] Wes McKinney. Data structures for statistical computing in Python. In Stepfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 51–56, 2010.

[7] Zhuang Shen, Fang Ning, Weigong Zhou, Xiong He, Changying Lin, Daniel P Chin, Zonghan Zhu, and Anne Schuchat. Superspreading SARS Events, Beijing, 2003. *Emerging Infectious Diseases*, 10(2), 2004.

Chapter 6

Conclusion

6.1 Summary

Globalization has changed the dynamics of infectious diseases in humans. Highly connected, affordable airplane travel means that an outbreak anywhere in the world is an immediate threat to people far away. The tradeoff is that there is more technology available than ever before to surveil and track those outbreaks. Access to a computer or mobile phone is no longer the luxury that it once was. This dissemination of technology means that more data is created and collected than ever before, and that data is routinely published to the internet. This data deluge is an opportunity for epidemiologists to enhance public health surveillance and response by shortening the window from outbreak emergence to detection, and detection to response and control.

Chapter 1 describes how near real time media reports during the emergence of H7N9 contain enough case information to construct a line listing. We combined the line list we assembled with other publicly available information like Census records and time use surveys to determine the infection risk per exposure hour for shoppers visiting live bird markets. When fused, these data revealed that elderly men are at disproportionate risk for contracting H7N9, suggesting an additional immunological risk factor. This insight was later supported by various other studies which found that dysfunctional host immune response is responsible for the pulmonary damage seen in severe H7N9 cases, and that elderly men are at higher risk for these cytokine storms. This work shows how even simple data fusion efforts can rapidly yield insights of public health importance.

Chapter 2 discusses how we built and parameterized an ordinary differential equation model of Middle East Respiratory Syndrome Coronavirus using publicly available data. Although many of the parameters were missing for that model, by doing parameter sweeps and identifying plausible scenario outcomes we were able to bound the parameter ranges to better understand the simulation space. Results show that spillover event frequency ranges are

compatible with a domestic or agricultural animal. This insight was later supported by other researchers who found that camels are a likely source of human exposure. Our results also show that men ages 19-44 are likely experiencing many infections that go undetected, perhaps because they are mild. Epidemiological surveillance for MERS has been underdeveloped in the affected region, so we await future studies to confirm or refute this finding.

Chapter 3 shows how models of an Ebola virus disease outbreak in West Africa can be built in near real time using publicly available data and inventive parameterization routines. The rapid and iterative nature of the modeling effort enables policymakers and public health professionals to use the models to guide their outbreak response. Our forecasts were used by the Department of Defense to inform their operations in Liberia and later Sierra Leone. Although our long term forecasts overestimated the number of future cases, which was a known and expected outcome, our short term forecasts were consistently in line with the observed data. We also implemented various interventions in the models to better understand what effect various countermeasures could have on the outbreak's future. This effort shows that modelers using agile and iterative modeling techniques can partner with outbreak responders to answer policy questions and provide evidence-based support for operational decision making.

The final chapter discusses the merits, limitations, and criticisms of open data; and describes a Python package I wrote to further develop the open epidemiology infrastructure. Although there are many obstacles to sharing epidemiology data, including privacy and data provenance issues, many of these objections can be overcome with appropriate data sharing agreements and robust data sharing infrastructure. Furthermore, analysis shows that publicly available data can be nearly as reliable as primary data. A comparison of the publicly available Ebola data with the closed, 'gold standard' patient databases shows that models parameterized using open data perform comparably to identical models parameterized using the patient database data.

6.2 Lessons learned

The primary lesson learned is that policy makers have an appetite for the work that modelers and computational epidemiologists conduct. Those policy makers are in charge of making continuous consequential decisions, often with very little guidance or evidence. Any insights epidemiologists can provide them improves their understanding of the situation, and augments understanding of which options would be the most effective. However, in order for the work of modelers to be relevant to decision makers, several goals must be met.

First, the work must be relevant. Although seemingly obvious, academics are often accustomed to working on problems of theoretical importance. Policymakers often have simple questions that need answers. To work with them effectively, academics must be prepared to listen to and meet their needs. Second, the work must be presented in a way that is

meaningful to the audience. Manuscripts or detailed methodological descriptions are of little use to someone who only has a few minutes to understand and synthesize the information. Slideshows that highlight results and recommended courses of action are preferred. Repeat presentations, for example weekly updates, are most effective if presented in a standardized format. That way the audience knows exactly what to expect, and can easily refer back to the information.

Despite these considerations which often fall outside the academic purview, the advantages of partnering with policymakers are numerous. The outcomes of those partnerships have direct relevance to the problem at hand. In the case of Ebola, the insights our findings generated were taken back to the field and used to inform the Ebola response. Government partners are also often important sources of interesting questions or data. They have a different perspective on situations, so their angle can open up new lines of thinking.

Another lesson learned is the importance of making data available to the public. None of the work in this dissertation could have proceeded without open and accessible data. Open data is critical not just for modelers, but for anyone involved in outbreak surveillance and response. Although the need for open data is obvious to people who do not otherwise have access, the primary stewards of the data do not always recognize or agree with its importance. For the data that is openly available, the formatting is such that it is often unusable in its published form. Building awareness around the importance of open data is a crucial effort that all public health professionals should be engaged in. Forging partnerships, pipelines and standards that facilitate the provision of data would go a long way towards enabling the kinds of work that can improve epidemiological response.

A third lesson learned is the importance of fluency with computational tools and thinking in epidemiology. Although most epidemiologists have a thorough understanding of the mathematical and statistical methods needed for their work, they are often not comfortable using the computational tools that would make implementing those methods easier and more reliable. Excel is a popular tool for data analytics, but it is cumbersome, unable to perform certain key operations like data cleaning, and cannot easily be used to build a pipeline for repeat analyses. Knowledge of a programming language like Python or R is critical for improving data handling skills, and that knowledge is something every epidemiologist should gain familiarity with, so epidemiology training should incorporate these skills.

On a broader level, I learned many lessons during the course of this dissertation work in the realm of technical and computational skills, knowledge domains, building productive working relationships, writing for a scientific audience, and more. For example, at the start of this work I had very little familiarity with any tools other than Microsoft Office. My PhD studies allowed me to learn Python, R, unix, and several other tools that are invaluable. My knowledge of epidemiology also flourished, from a sketch of what I had read in textbooks to the intimate knowledge that can only come with analyzing the primary data.

6.3 Limitations

The old modeling adage is that outputs are only as good as the inputs. Data available in the midst of an outbreak, as was the case in the studies described here, are subject to myriad problems and uncertainties. This is especially true in the case of open data. The data are collected under chaotic circumstances, often in resource-poor settings. They are usually poorly documented, and subject to revision - sometimes without comment. They may be incomplete or flat out wrong. Compounding matters, it is often impossible to tell which of these hazards are present with a particular data set. Any time these data are used for model building or analyses, the results produced must be treated with the assumption they are very likely wrong.

A second data limitation separate from the epidemiological data is that there is usually very little information in the scientific literature about emerging zoonoses. Parameters like incubation period or serial interval, transmissibility, or clinical course are often unknown. Sometimes these parameters can be inferred, or optimization routines can be used to narrow down the plausible parameter space, but primary literature is always preferable.

Although there are clear advantages to using open data to model emerging zoonoses in the midst of an outbreak, there are significant drawbacks as well. Awareness of the limitations is critical for properly implementing studies of the kind described in the dissertation, and for clearly communicating expectations and results. Despite this caveat, the other old modeling adage, all models are wrong but some are useful, is just as true here as for models built with gold-standard data. The studies described here provided insight into an emergency situation when there was previously none. Some of our findings, for example the long term Ebola forecasts, were wrong - and we knew they were very likely to be wrong when we produced them, since they did not include any interventions or changes in behavior. But they gave our policymaker partners an upper bound on what to expect, and for that goal they were useful. As long as the results are contextualized and the uncertainty is described, even rough findings can have their place.

6.4 Next steps

Despite the best efforts of public health professionals worldwide, emerging zoonoses are a phenomenon humanity will continue to contend with. Broadly, the next steps are to continue to apply modeling and data analytics to new outbreak situations as they arise. Part of the challenge of emerging zoonoses is that you never know what will be next, and where it will hit. In the interim, next steps include strengthening partnerships with our non-academic colleagues and inventing and refining modeling methods and pipelines.

Each of these projects involved collecting a wealth of data about the disease in question. Because the projects were done while the outbreaks were evolving in order to help inform

response, analyses were streamlined and did not include much data exploration. If a period of relative calm from public health threats should ever occur, the data collected for each of these projects may hold many insights into the nature and dynamics of emerging zoonoses. A planned next step would be to return to those data to learn more so that we know more for the next big outbreak. Of particular interest is early identification of human to human potential by analyzing disease clusters, building off the Python package, epipy, described in Chapter 4.

6.5 Final thoughts

Although modeling of infectious disease is hardly new, emerging zoonoses are seriously underrepresented in the literature. For those models that do exist, they are usually developed well after the outbreak has finished. Although this timeline is sensible from an academic perspective, it is a missed opportunity in terms of improving public health. This dissertation is an effort to improve on those weaknesses. It highlights the value of using publicly available open data to build models of emerging zoonoses to support decision makers in their outbreak response operations.

Appendix - Supplementary material

Supplementary material to *Modeling the impact of interventions on an epidemic of Ebola in Sierra Leone and Liberia*

Scenario	**Community** R_0	**Hospital** R_0	**Funeral** R_0	**Overall** R_0
Baseline	1.35	0.35	0.53	2.22
Contact Tracing (80%)	1.08	0.50	0.53	2.11
Contact Tracing (90%)	0.89	0.58	0.53	2.01
Contact Tracing (100%)	0.69	0.67	0.53	1.89
β_H reduction (25%)	1.35	0.26	0.53	2.14
β_H reduction (50%)	1.35	0.18	0.53	2.04
β_H reduction (75%)	1.35	0.09	0.53	1.96
Contact tracing (100%) and β_H reduction (75%)	0.69	0.50	0.53	1.72
Pharmaceutical (25% efficacy)	1.08	0.51	0.44	2.02
Pharmaceutical (50% efficacy)	1.08	0.53	0.34	1.94
Pharmaceutical (75% efficacy)	1.08	0.54	0.24	1.88

Table A.1: R_0 Estimations for Liberia

Scenario	**Community** R_0	**Hospital** R_0	**Funeral** R_0	**Overall** R_0
Baseline	1.11	0.24	0.43	1.78
Contact Tracing (80%)	0.87	0.47	0.44	1.77
Contact Tracing (90%)	0.78	0.53	0.44	1.75
Contact Tracing (100%)	0.70	0.59	0.44	1.73
β_H reduction (25%)	1.11	0.18	0.43	1.72
β_H reduction (50%)	1.11	0.12	0.43	1.65
β_H reduction (75%)	1.11	0.06	0.43	1.60
Contact tracing (100%) and β_H reduction (75%)	0.70	0.59	0.44	1.73
Pharmaceutical (25% efficacy)	0.87	0.54	0.39	1.80
Pharmaceutical (50% efficacy)	0.87	0.64	0.33	1.83
Pharmaceutical (75% efficacy)	0.87	0.78	0.23	1.88

Table A.2: R_0 Estimations for Sierra Leone

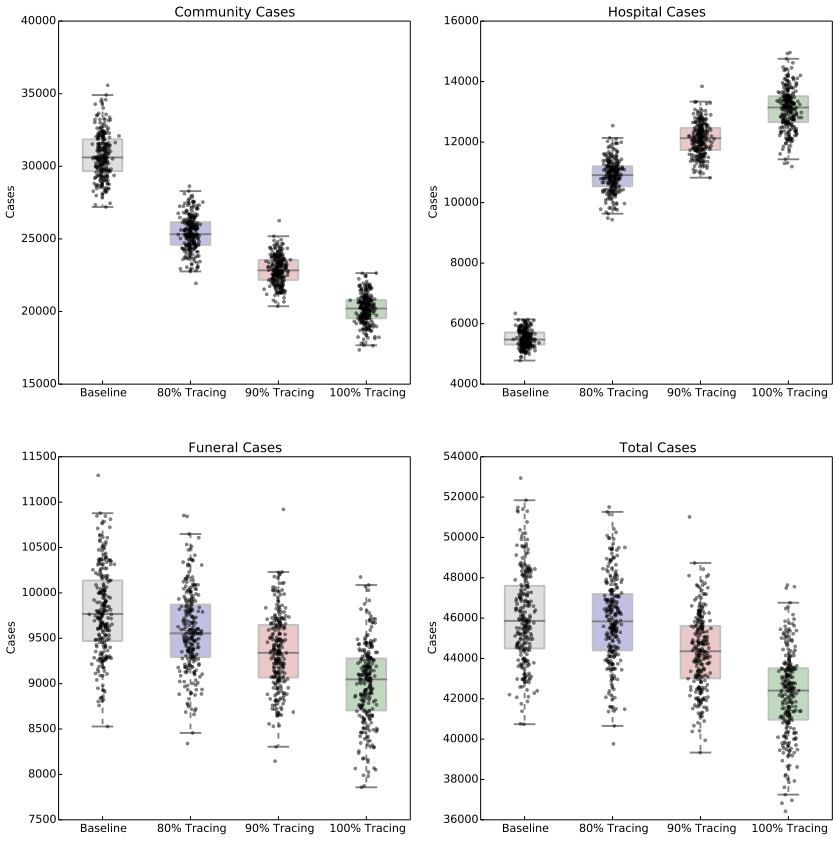

Figure A.1: **Distribution of Forecasted Cases of Community, Hospital, Funeral and Total Cases for Ebola Epidemic, Sierra Leone, 2014, at Baseline, 80%, 90% and 100% of Patients Traced and Hospitalized.** Box plots depict the median, interquartile range and 1.5 times the interquartile range for each scenario. Each individual simulated forecast is shown as a single dot, jittered so as to depict the complete distribution of the data.

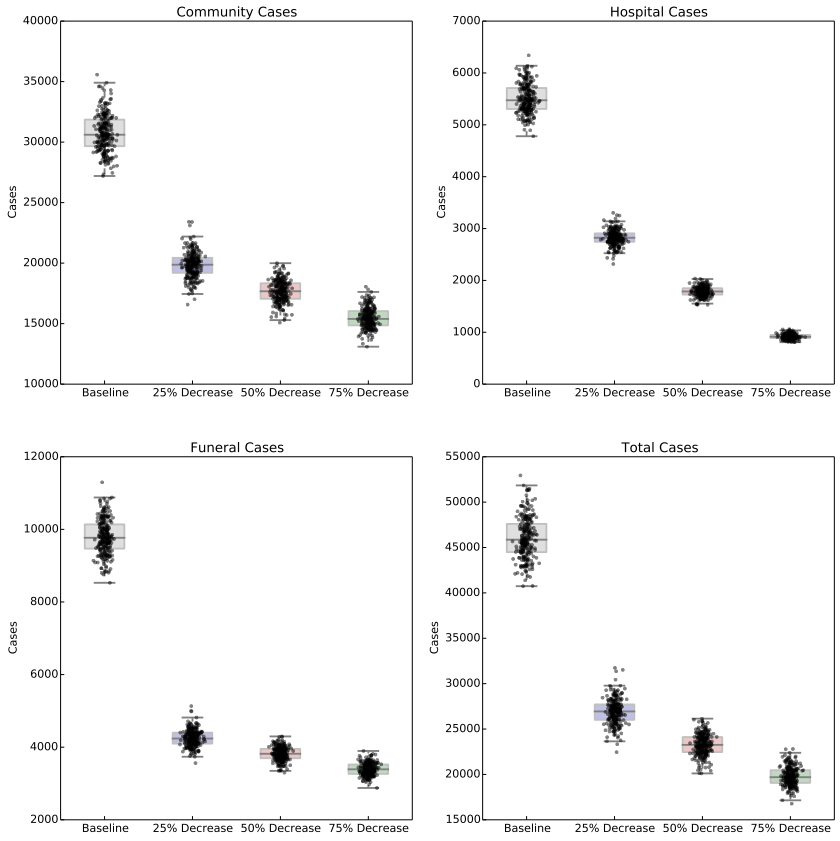

Figure A.2: **Distribution of Forecasted Cases of Community, Hospital, Funeral and Total Cases for Ebola Epidemic, Sierra Leone, 2014, at Baseline, 25%, 50% and 75% Reductions in Hospital Transmission Contact Rates (β_H).** Box plots depict the median, interquartile range and 1.5 times the interquartile range for each scenario. Each individual simulated forecast is shown as a single dot, jittered so as to depict the complete distribution of the data.

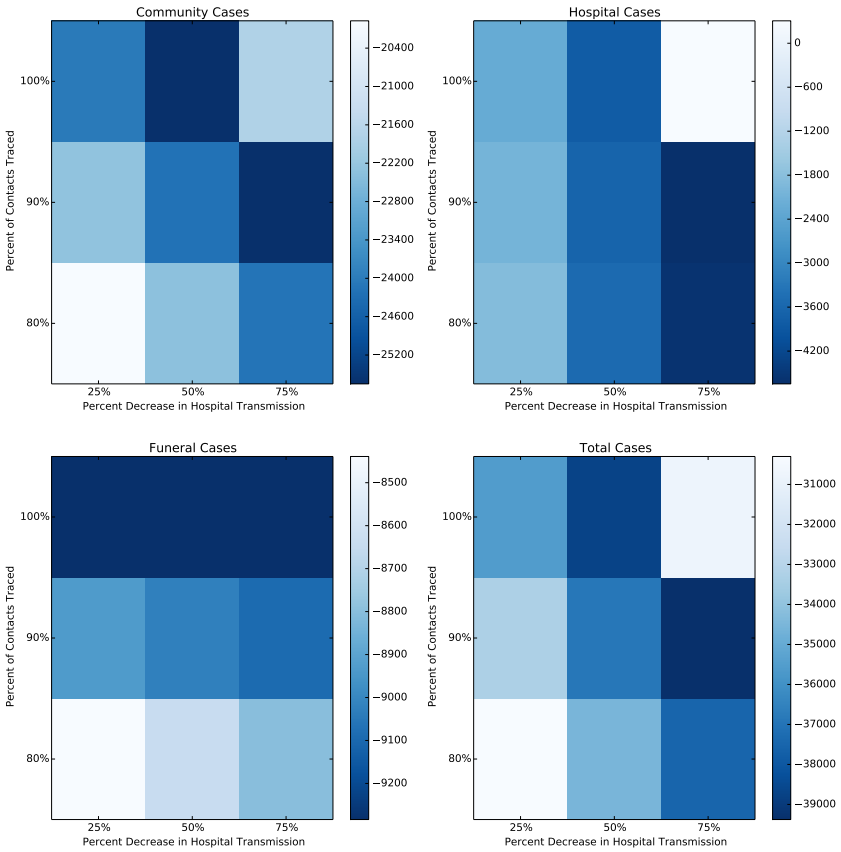

Figure A.3: **Distribution of forecasted cases of Community, Hospital, Funeral and Total Cases for Ebola Epidemic, Sierra Leone, 2014, at Baseline, 25%, 50% and 75% Reductions in Hospital Transmission Contact Rates (β_H) with 80%, 90% and 100% of Patients Traced and Hospitalized.** Each box represents the median result of 250 forecasted epidemics, each with a % of contacts traced and a % decrease in hospital transmission. Areas of deeper blue indicate progressively greater reductions of the median number of cases.

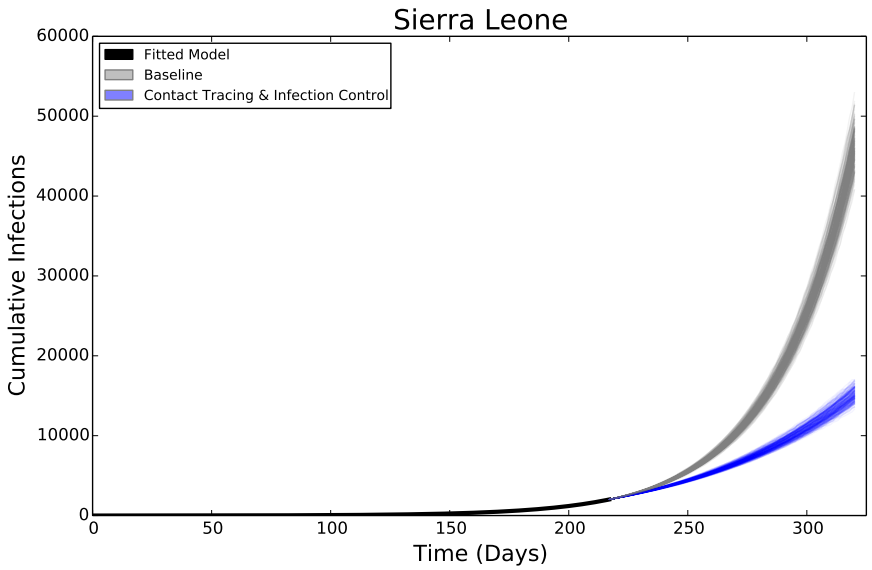

Figure A.4: **Forecasted Cumulative Cases for Ebola Epidemic, Sierra Leone, 2014 with 75% Reduction in Hospital Transmission Contact Rates (β_H) with 100% of Patients Traced and Hospitalized.** The solid black line represents the deterministic model fit of the epidemic to present, with each grey line representing a single simulated forecast with no interventions in place, and each blue line representing a single simulated forecast of the epidemic with 100% of contacts traced, a 75% (reduction in hospital transmission (β_H) and no post-mortem infections from hospitalized patients. Areas of darker color indicate more forecasts with that result.

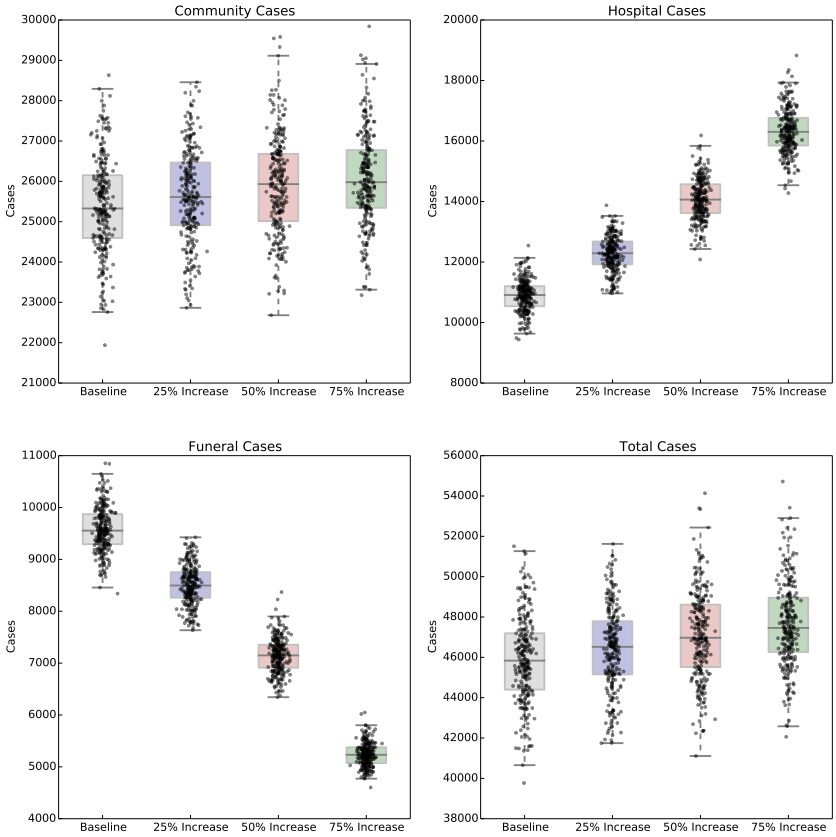

Figure A.5: **Distribution of Forecasted Cases of Community, Hospital, Funeral and Total Cases for Ebola Epidemic, Sierra Leone, 2014, at Baseline, 25%, 50% and 75% Reductions in Case Fatality Rate Due to a Hypothetical Pharmaceutical Intervention.** Box plots depict the median, interquartile range and 1.5 times the interquartile range for each scenario. Each individual simulated forecast is shown as a single dot, jittered so as to depict the complete distribution of the data.

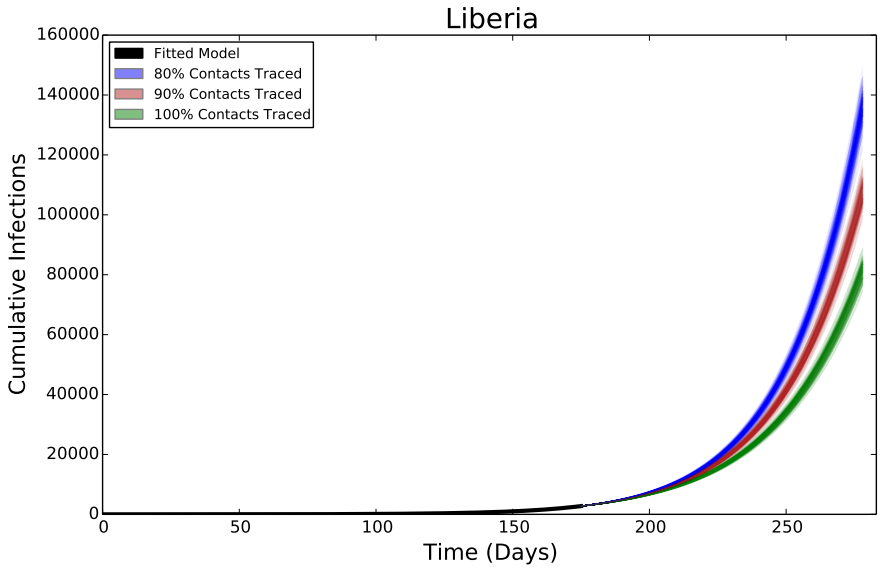

Figure A.6: **Forecasted Cumulative Cases for Ebola Epidemic, Liberia, 2014 with 80%, 90% and 100% of Patients Traced and Hospitalized.** The solid black line represents the deterministic model fit of the epidemic to present, with each colored line representing a single simulated forecast of the epidemic with 80% (blue), 90% (red) or 100% (green) contacts traced. Areas of darker color indicate more forecasts with that result.

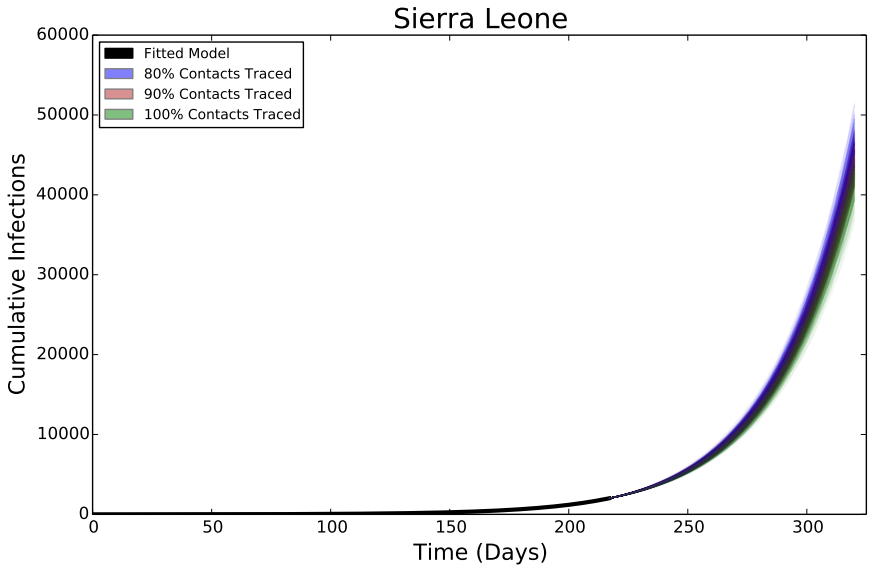

Figure A.7: **Forecasted Cumulative Cases for Ebola Epidemic, Sierra Leone, 2014 with 80%, 90% and 100% of Patients Traced and Hospitalized.** The solid black line represents the deterministic model fit of the epidemic to present, with each colored line representing a single simulated forecast of the epidemic with 80% (blue), 90% (red) or 100% (green) contacts traced. Areas of darker color indicate more forecasts with that result.

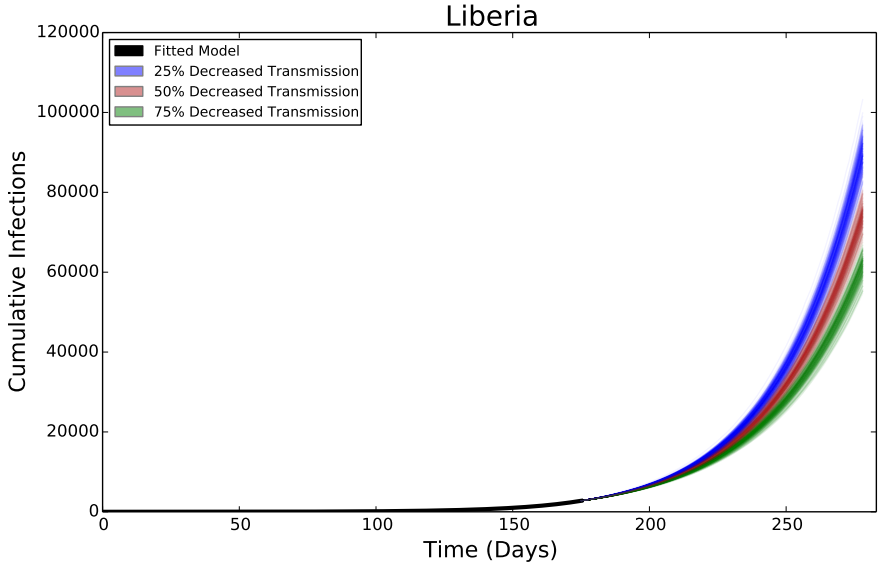

Figure A.8: **Forecasted Cumulative Cases for Ebola Epidemic, Liberia, 2014 at Baseline, 25%, 50% and 75% Reductions in Hospital Transmission Contact Rates (β_H).** The solid black line represents the deterministic model fit of the epidemic to present, with each colored line representing a single simulated forecast of the epidemic with 25% (blue), 50% (red) or 75% (green) reduction in hospital transmission (β_H) and no post-mortem infections from hospitalized patients. Areas of darker color indicate more forecasts with that result.

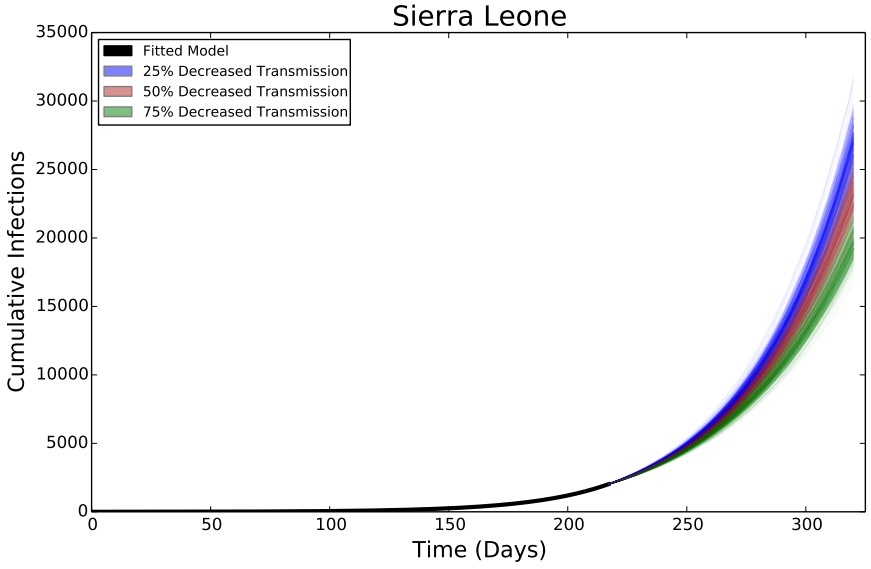

Figure A.9: **Forecasted Cumulative Cases for Ebola Epidemic, Sierra Leone, 2014 at Baseline, 25%, 50% and 75% Reductions in Hospital Transmission Contact Rates (β_H).** The solid black line represents the deterministic model fit of the epidemic to present, with each colored line representing a single simulated forecast of the epidemic with 25% (blue), 50% (red) or 75% (green) reduction in hospital transmission (β_H) and no post-mortem infections from hospitalized patients. Areas of darker color indicate more forecasts with that result.

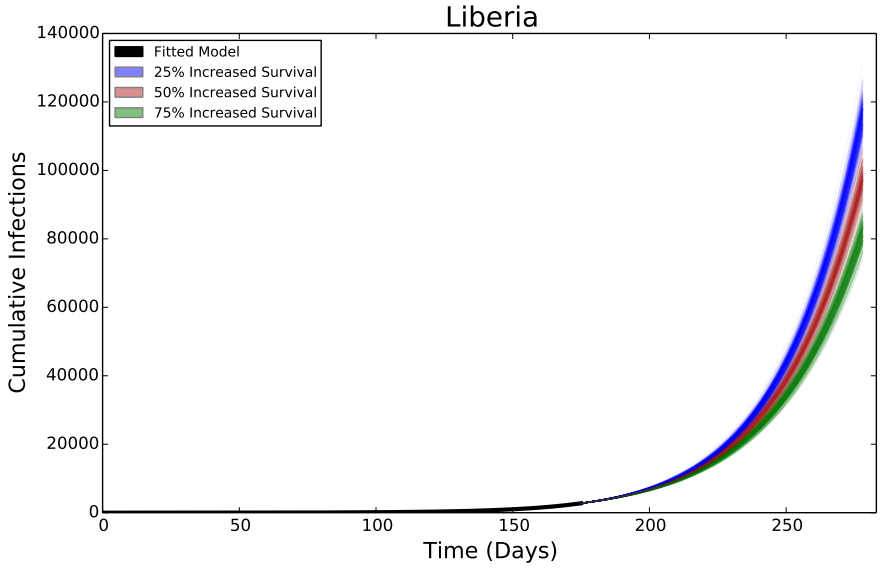

Figure A.10: **Forecasted Cumulative Cases for Ebola Epidemic, Liberia, 2014 at Baseline, 25%, 50% and 75% Reductions in Case Fatality Rate Due to a Hypothetical Pharmaceutical Intervention.** The solid black line represents the deterministic model fit of the epidemic to present, with each colored line representing a single simulated forecast of the epidemic with a hypothetical pharmaceutical intervention that increases hospitalized patient survival by 25% (blue), 50% (red) or 75% (green), along with 80% contact tracing. Areas of darker color indicate more forecasts with that result.

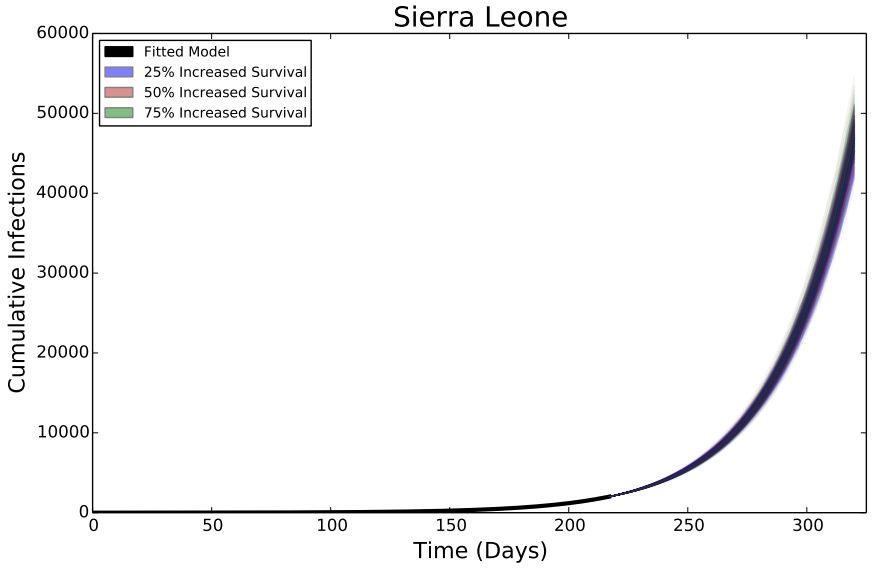

Figure A.11: **Forecasted Cumulative Cases for Ebola Epidemic, Sierra Leone, 2014 at Baseline, 25%, 50% and 75% Reductions in Case Fatality Rate Due to a Hypothetical Pharmaceutical Intervention.** The solid black line represents the deterministic model fit of the epidemic to present, with each colored line representing a single simulated forecast of the epidemic with a hypothetical pharmaceutical intervention that increases hospitalized patient survival by 25% (blue), 50% (red) or 75% (green), along with 80% contact tracing. Areas of darker color indicate more forecasts with that result.