Infectious Disease Modeling: Enhancing Epidemic Preparedness and Response

HARVARD UNIVERSITY

Graduate School of Arts and Sciences

DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the
Committee on Higher Degrees in Population Health Sciences,
have examined a dissertation entitled

"Infectious Disease Modeling: Enhancing Epidemic Preparedness and Response"

presented by

Rebecca Kahn

candidate for the degree of Doctor of Philosophy
and hereby certify that it is worthy of acceptance.

Dr. Marc Lipsitch, D.Phil., Committee Chair, Harvard T.H. Chan School of Public Health
Dr. Caroline Buckee, D.Phil., Harvard T. H. Chan School of Public Health
Dr. Rui Wang, Ph.D., Harvard Medical School, Harvard T. H. Chan School of Public Health

In lieu of all Dissertation Advisory Committee members' signatures,
I, Tyler J. VanderWeele, Ph.D., appointed by the Ph.D. in Population Health Sciences,
confirm that the Dissertation Advisory Committee has examined the above dissertation,
presented by Rebecca Kahn and hereby certify that it is worthy of acceptance as of 1 October 2020.

Date:1 October 2020

Infectious Disease Modeling:
Enhancing Epidemic Preparedness and Response

A dissertation presented

by

Rebecca Kahn

to

The Department of Epidemiology

Harvard T.H. Chan School of Public Health

&

The Department of Population Health Sciences

Harvard Graduate School of Arts and Sciences

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

In the subject of

Population Health Sciences (Field of Study: Epidemiology)

Harvard University

Cambridge, MA

October 2020

Infectious Disease Modeling: Enhancing Epidemic Preparedness and Response

Abstract

Recent outbreaks of Ebola, Zika, and COVID-19, among others, have shown how infectious diseases can decimate economies and destroy lives. Infectious disease models are important tools for preparing for, preventing, and responding to such epidemics. Here, we use infectious disease modeling to analyze past outbreaks, prepare for future outbreaks, and respond to ongoing outbreaks, with the goal of informing public health response.

We first analyze past Ebola and cholera outbreaks and build a simulation model to understand the role the incubation period, the time between exposure and symptom onset, has on epidemic trajectory. We find that diseases with longer incubation periods, such as Ebola, where infected individuals can travel further before becoming infectious, result in more long-distance sparking events and less predictable disease trajectories, as compared to the more predictable wave-like spread of diseases with shorter incubation periods, such as cholera. Second, we assess if augmenting classical randomized controlled trials of vaccines with pathogen sequence and contact tracing data can permit these trials to estimate vaccine efficacy against infectiousness, or the reduction in onward transmission from a vaccinated person who is infected compared to an unvaccinated infected person. Through simulations of a transmission model and a vaccine trial, we find that these data sources enhance identifiability of this key measure of vaccine efficacy. Finally, we simulate studies of SARS-CoV-2 seroprotection. We find that in studies assessing whether seropositivity confers protection against future infection, time varying epidemic dynamics can cause confounding; it is therefore necessary to adjust for geographic location and time of enrollment in order to reduce bias. These methods and findings demonstrate how infectious disease modeling can be used to enhance epidemic preparedness and response.

Table of Contents

Acknowledgements

I first want to thank my committee members, Marc Lipsitch, Caroline Buckee, and Rui Wang, for their mentorship and guidance over the past four years on these chapters and other projects. I have learned so much from each of them and am grateful for everything they have taught me. I am looking forward to continuing to work with all of them.

I also want to thank Marc and Caroline for being truly inspirational PIs who have not only taught me how to do good science but also the importance of translating that science into action to improve public health. The two of them, along with the other PIs, students, and postdocs, make CCDD an incredible place to work, and I feel so fortunate to be a part of such a collaborative and fun group.

I am so grateful to have had amazing co-authors on each of these chapters and would like to thank all of them for their collaboration, mentorship, and support.

- Chapter 2: Corey Peak, Juan Fernández-Gracia, Alexandra Hill, Amara Jambai, Louisa Ganda, Marcia Castro, and Caroline Buckee. (https://www.pnas.org/content/117/9/5067.short?rss=1)
- Chapter 3: Rui Wang, Sarah Leavitt, Bill Hanage, and Marc Lipsitch (https://www.medrxiv.org/content/10.1101/2020.09.14.20193789v1)
- Chapter 4: Lee Kennedy-Shaffer, Yonatan Grad, James Robins, and Marc Lipsitch. (https://academic.oup.com/aje/advance-article/doi/10.1093/aje/kwaa188/5900104)

Additionally, I want to thank Eric DiGiovanni for his help over the years. I always knew I could turn to him to answer my questions, and I am so grateful to him for cohosting my defense.

Special thanks to Inga Holmdahl and Kelsey Vercammen. I could not have gotten through the masters or PhD without their friendship, support, texting, and Thai food. I am so lucky to have them both in my life.

I want to thank my family, especially my grandmothers, Carolyn Gold and Jane Kahn, and parents, Susan and Bobby Kahn, for being my biggest cheerleaders and providing unconditional and unending support throughout my entire life.

Finally, I want to thank Rahul Nayak for being the best partner and friend I could have ever asked for and for knowing way more about all of these papers than I'm sure he ever wanted to know. I cannot thank him enough for his love and support.

List of Tables and Figures

Chapter 1. Introduction

Recent outbreaks of Ebola,[1,2] Zika,[3] and COVID-19,[4] among others, have shown how infectious diseases can decimate economies and destroy lives. Infectious disease models are important tools for preparing for, preventing, and responding to such epidemics. Models allow us to test assumptions,[5] identify key sources of uncertainty, examine interventions and advocate for policies or programs, explore the impact of different parameters, or plan the design and analysis of trials in advance.[6]

Outbreak science is an emerging field that seeks to integrate mathematical modeling of infectious diseases more systematically into public health decision making.[7] In the following chapters, we use infectious disease modeling to analyze past outbreaks, prepare for future outbreaks, and respond to ongoing outbreaks, with the goal of informing public health response and enhancing trial designs.

In the second chapter, we analyze the trajectories of recent back-to-back outbreaks of Ebola and cholera in Sierra Leone. This analysis motivates a question regarding the role of the incubation period, the time between exposure and symptom onset, in predictability of outbreak spread. To answer this question, we develop a simulation model to compare metrics of outbreak trajectory across a range of incubation periods.

In the third chapter, we aim to assess if vaccine trials can be designed to measure the vaccine candidate's impact on infectiousness. Even if the vaccine does not prevent everyone from getting infected, does it have an impact on further transmission? Estimating this key measure, however, requires knowledge of who infected whom. We therefore simulate sampling

of genome sequences and contact tracing data to attempt to reconstruct transmission networks to inform estimation of vaccine efficacy against infectiousness.

In the fourth chapter, we identify biases that can arise in studies that assess whether or not prior infection with SARS-CoV-2 confers protection against future infection. Using simulation models, we demonstrate when these biases occur and identify ways to ameliorate them. Accurate estimates of SARS-CoV-2 seroprotection will be critical for understanding the dynamics of this pandemic and implementing measures to control it.

Looking across a range of pathogens, including Ebola, cholera, and SARS-CoV-2, we show how infectious disease modeling can shed insight on previous outbreaks that can be useful for informing response efforts for future outbreaks. This work also underscores the importance of using simulations to aid in the design and analysis of vaccine and seroprotection studies to minimize bias and enhance the information obtained from these trials conducted in urgent settings. Through these chapters, we aim to show that infectious disease modeling has the potential to change the course of epidemics and save lives.

Chapter 2. Incubation periods impact the spatial predictability of outbreaks: analysis of cholera and Ebola outbreaks in Sierra Leone

2.1 ABSTRACT

Forecasting the spatiotemporal spread of infectious diseases during an outbreak is an important component of epidemic response. However, it remains challenging both methodologically and with respect to data requirements as disease spread is influenced by numerous factors, including the pathogen's underlying transmission parameters and epidemiological dynamics, social networks and population connectivity, and environmental conditions. Here, using data from Sierra Leone we analyze the spatiotemporal dynamics of recent cholera and Ebola outbreaks and compare and contrast the spread of these two pathogens in the same population. We develop a simulation model of the spatial spread of an epidemic in order to examine the impact of a pathogen's incubation period on the dynamics of spread and the predictability of outbreaks. We find that differences in the incubation period alone can determine the limits of predictability for diseases with different natural history, both empirically and in our simulations. Our results show that diseases with longer incubation periods, such as Ebola, where infected individuals can travel further before becoming infectious, result in more long-distance sparking events and less predictable disease trajectories, as compared to the more predictable wave-like spread of diseases with shorter incubation periods, such as cholera.

2.2 INTRODUCTION

Epidemics of emerging infectious diseases such as Ebola and Zika underscore the need to improve global capacity for surveillance and response.[3,8,9] Forecasting the spatiotemporal spread of infectious diseases during an outbreak can enable responders to stay ahead of an epidemic.

However, it remains challenging both methodologically and with respect to data requirements[10,11] as disease spread is influenced by multiple factors, including: the pathogen's underlying transmission parameters and epidemiological dynamics; social networks and population connectivity; and environmental conditions.[12–15] Previous forecasting efforts have had varying levels of success in predicting the total number of cases and spatiotemporal spread of outbreaks like Ebola, and few have actually been used in real time in the midst of an epidemic.[13] Efforts to understand the likely performance of forecasts have shown that heterogeneity in contact structure and number of secondary infections can pose challenges, but reasonable predictions can be made in some cases, depending on disease-specific parameters.[12] However, the epidemiological attributes that determine predictability remain uncertain in real-world settings.[16–18]

The time from when individuals are infected to when they become infectious (the latent period) and to when they become symptomatic (the incubation period), and the relationship between the two, have been shown to play a large role in the epidemic potential of diseases.[15,19,20] In particular, transmission that occurs during the incubation period before an individual develops symptoms can contribute to rapid disease spread. When the latent period is shorter than the incubation period for an infectious individual, pre-symptomatic transmission can be a strong driver of the total number of secondary infections by an infectious individual in a completely susceptible population (i.e. R_0).[19,20] Indeed, the basis of contact tracing protocols during an outbreak reflect the need to identify and contain individuals during the incubation period, and the relative effectiveness of interventions such as symptom monitoring or quarantine significantly depends on the relationship between infectiousness and symptoms.[20] Additional related metrics, the generation interval (i.e. the time between infection of an infector-infectee

pair) and the serial interval (i.e. the time between symptom onset of an infector-infectee pair), as well as their variances, can further impact the growth rate and total number of infections during an epidemic.[21] The incubation period is also likely to play a particularly important role in determining the spatial spread of an epidemic because one's typical travel may continue prior to symptom onset, whereas travel behavior may change or stop altogether during illness,[22] particularly when symptoms are severe or immobilizing; even if symptoms are mild, if one knows they are infected, behavior may also change, impacting transmission.

Back-to-back epidemics of cholera (2012-2013) and Ebola (2014-2015) in Sierra Leone present a unique opportunity to compare the spatial dynamics of two epidemics in the same population caused by pathogens with notable similarities in both the drivers of outbreaks and the interventions used to curtail them, including oral rehydration.[23,24] Both are transmitted through contact with contaminated diarrhea or vomitus (plus other bodily fluids for Ebola), and the reproductive number (R_0) for both diseases is thought to be between 1 and 3.[25,26] Both diseases can cause immobilizing gastrointestinal symptoms of diarrhea and vomiting and, untreated, their case fatality rates can exceed 50%.[27,28] Cultural factors and rituals, such as traditional funeral practices, are known to influence the spread of both cholera[29] and Ebola,[30] while water, sanitation, and hygiene (WASH) programs are often used to slow the spread of each.[31] Both epidemics occurred against a backdrop of an immunologically naïve population. Although it seems likely that travel patterns and the density and distribution of people were broadly similar over the time period in question, regular movements may have been more impacted during the Ebola epidemic than during the cholera epidemic due travel restrictions, particularly during the multi-day lockdowns.[32] One critical difference between the dynamics of these diseases, however,

is the incubation period, which is estimated at a median of 8-12 days between infection and onset of symptoms for Ebola[8] and only 1-2 days for cholera.[33]

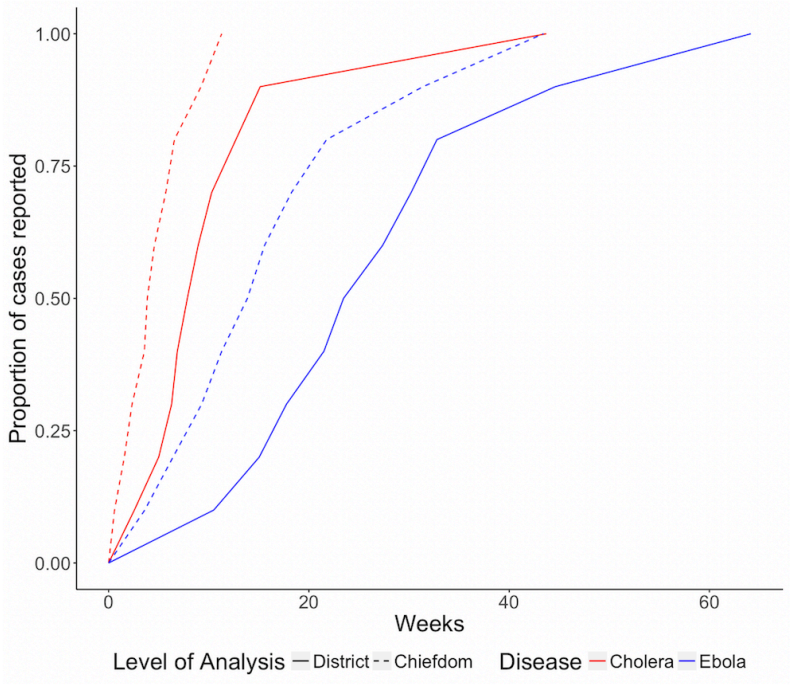
We hypothesize that the disease incubation period may be a particularly influential driver of different patterns of disease spread through space and time. We analyze the spatiotemporal dynamics of a cholera outbreak and an Ebola outbreak in Sierra Leone, both of which occurred over a similar time period. We develop a simulation model of the spatial spread of an epidemic and examine the impact of the incubation period on the dynamics of spread and the predictability of outbreaks. We find that differences in the incubation period alone can determine the limits of predictability for these diseases with different natural history, both empirically and in our simulations. Our results show that diseases with longer incubation periods, such as Ebola, where infected individuals can travel further before becoming infectious, result in more long-distance sparking events and less predictable disease trajectories, as compared to the more predictable wave-like spread of diseases with shorter incubation periods, such as cholera.

2.3 RESULTS

We first summarize the cholera and Ebola epidemics in terms of their dynamics in time and space. More cases were reported during the cholera epidemic (22,691) than during the Ebola epidemic (11,903); however, far fewer cholera cases were fatal (324 vs. 3,956). Both epidemics lasted for similar periods of time, with cholera (January 7, 2012 – May 14, 2013) occurring two years prior to Ebola (May 18, 2014 – September 12, 2015). Data for both outbreaks were reported at the chiefdom level, the third-level administrative units. The times between the onset of an outbreak and when half or all of its cases were reported were longer when outbreaks were aggregated by district (second-level administrative units, comprised of chiefdoms), instead of

chiefdom (**Figure 2.1**), which has implications for the optimal scale for surveillance and response measures. The median time for a chiefdom cholera outbreak to report half its case total was 3.9 weeks, and median outbreak duration was 11.3 weeks. The median time for district outbreaks to report half their cholera cases was 7.9 weeks, and the median outbreak duration was 43.7 weeks. Analysis of Ebola revealed similar trends, with chiefdoms reporting half of their cases at a median of 13.9 weeks and median outbreak duration of 43.3 weeks, and districts reporting half of their cases at a median of 23.5 weeks and median outbreak duration of 64.1 weeks.

Figure 2.1. The proportion of cholera and Ebola cases reported over time differed between district and chiefdom level



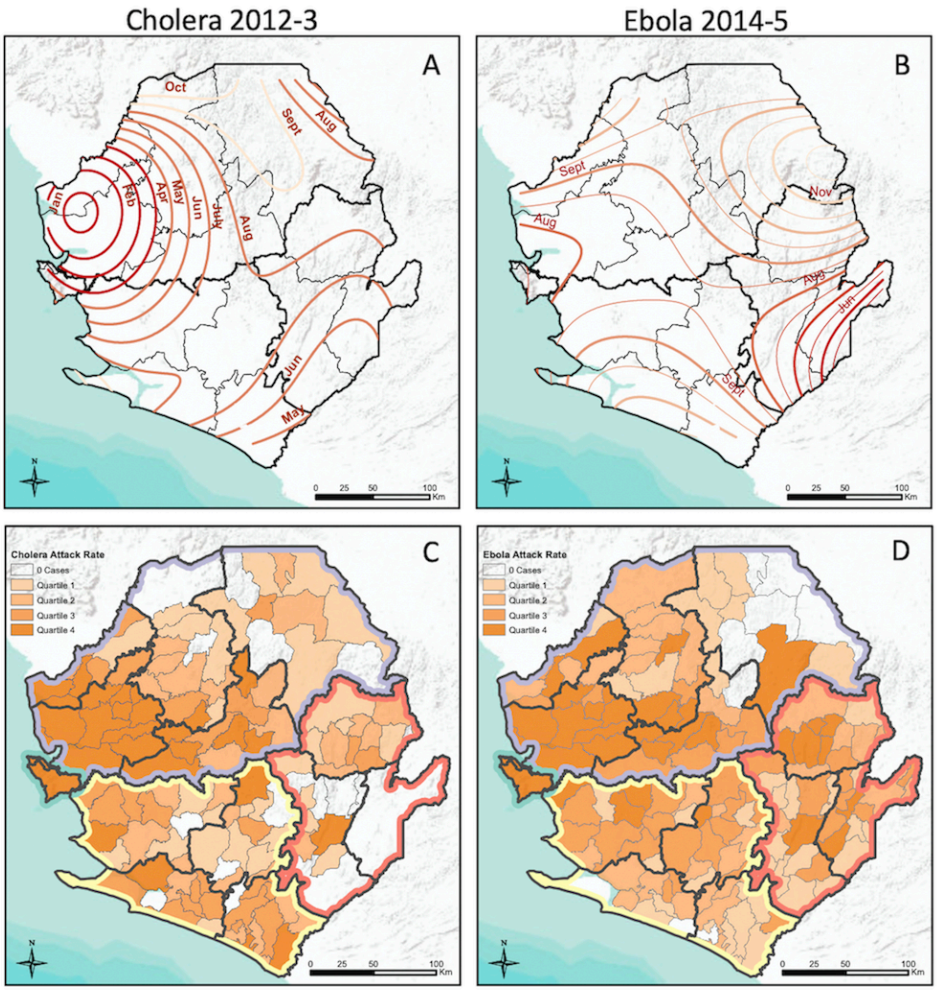
The times between the onset of an outbreak and when half or all of its cases were reported were longer when outbreaks were aggregated by district instead of chiefdom, which has implications for the optimal scale for surveillance and response measures. The median time for a chiefdom cholera outbreak to report half its case total was 3.9 weeks and a median of 7.9 weeks for district cholera outbreaks. For Ebola, chiefdoms reported half of their cases at a median of 13.9 weeks and districts at a median of 23.5 weeks.

Both the cholera and Ebola epidemics were widespread, each reaching more than 75% of the country's chiefdoms. However, their trajectories differed. The spread of cholera from the northwest followed a radial spatial dispersion gradually in all directions for the first six months, while Ebola spread from the southeast for two months before rapid expansion to the northwest which sparked the national epidemic (**Figure 2.2 A-B**; **Supplementary Movie 2.1**). These findings were statistically supported by space-time analysis of each epidemic, which revealed clusters of high case reporting of both diseases in western Sierra Leone and unique clusters of cholera in the south and Ebola in the east (**Supplementary Figure 2.1**). The wave front of chiefdom cholera outbreak onset progressed more slowly and gradually than for Ebola, which exhibited faster and more discontinuous expansion as shown by the larger spacing between monthly contour lines (**Figure 2.2 A-B**). Despite their different trajectories, the geography of the epidemics largely overlapped, with clusters of high cumulative attack rates of cholera and Ebola observed in the north and west regions of Sierra Leone (**Figure 2.2 C-D**) and confirmed through Local Moran's I methods (**Supplementary Figure 2.2**).

As a daily estimate of transmission intensity, we recorded the effective reproductive number (R_t) and its variation over time nationally and by region (**Figure 2.3**). While some areas sustained transmission (i.e., $R_t > 1$) of both cholera and Ebola for many days (e.g., Freetown in the west and Kenema Town in the east), 75% of chiefdoms during the cholera outbreak and 44% of chiefdoms during the Ebola outbreak recorded either zero cases or zero days with $R_t > 1$ (**Supplementary Figure 2.3**). As expected, transmission intensity of both diseases was positively correlated in chiefdoms near each other (**Supplementary Figure 2.4**). Correlation decayed with distance, consistent with local disease spread, and inter-chiefdom distances of over 100km eliminated any evidence of positive correlation of disease presence, chiefdom outbreak

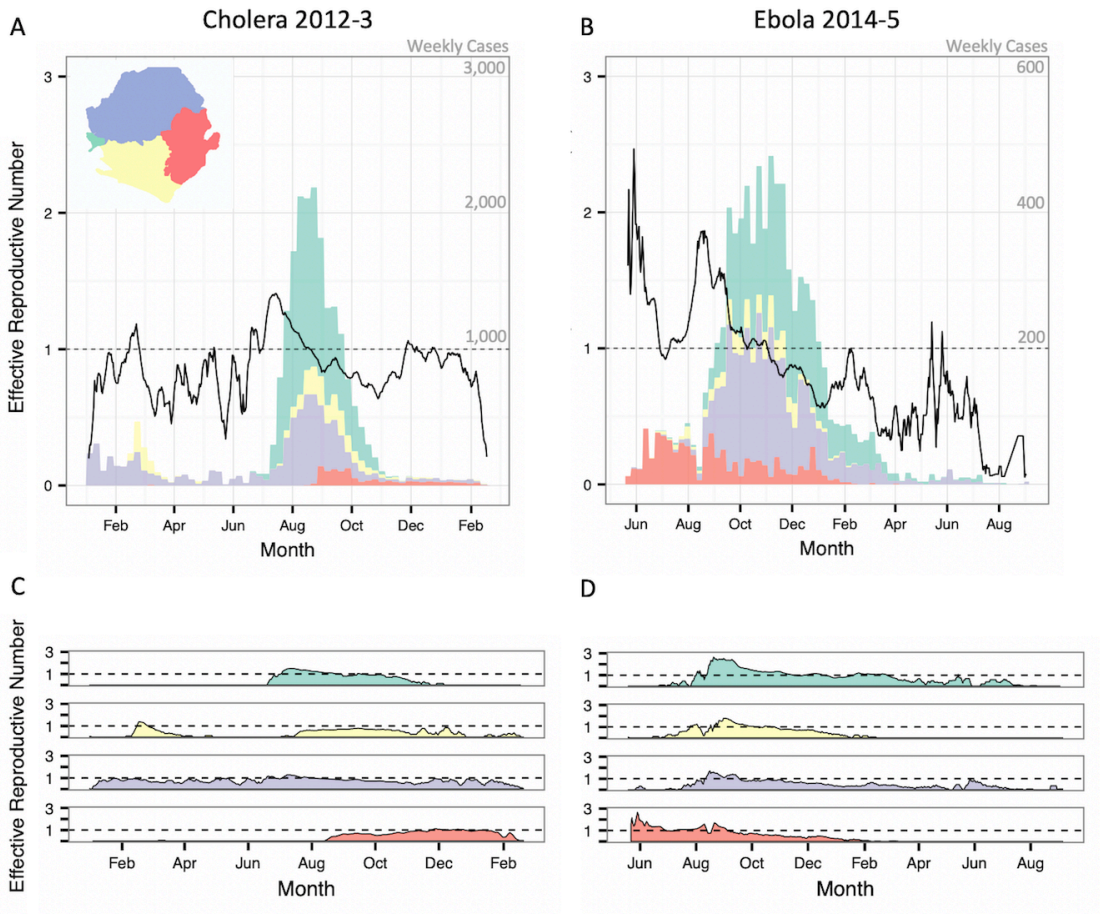
time, case count, and cumulative attack rate (**Supplementary Figure 2.4**). These metrics appear more highly correlated in space for cholera than for Ebola, although the confidence intervals overlap (**Supplementary Figure 2.4**).

Figure 2.2. Spatial trend contours of disease spread and chiefdom attack rates



Spatial trend contours of disease spread and chiefdom attack rates highlight similarities and differences between the two epidemics. Spatial trend contours of cholera (A) and Ebola (B) spread from areas in dark red to light orange; thicker lines (A & B) show monthly increments and thinner lines (B) show 2 week increments. Each line of the same color represents the same timescale. Larger spacing between lines represents faster spread. Thick black lines denote regions and thin black lines denote districts. Chiefdom attack rate quartile for cholera (C) and Ebola (D) vary over space and regions. Colored boundaries denote regions, followed by bold black borders for districts and thin borders for chiefdoms. *Figure made by coauthor Corey Peak.*

Figure 2.3. Weekly case counts and effective reproductive number



Weekly case counts show outbreak trajectory in the four regions of the country. The bars in A and B indicate the weekly case count on independent y-axes of cholera and Ebola, respectively. Black lines show maximum likelihood estimates of R_t of cholera and Ebola epidemics nationally (A and B, respectively) and in each region (C and D, respectively). *Figure made by coauthor Corey Peak.*

Simulations

To examine the role of the incubation period in the spread of disease, we simulated outbreaks characterized by varying incubation periods among agents distributed evenly on a spatial lattice, with movement between populations in the lattice based on a gravity model. These simulations show a systematic relationship between the incubation period and spatiotemporal patterns of disease spread. As expected, simulated epidemic curves of diseases with shorter

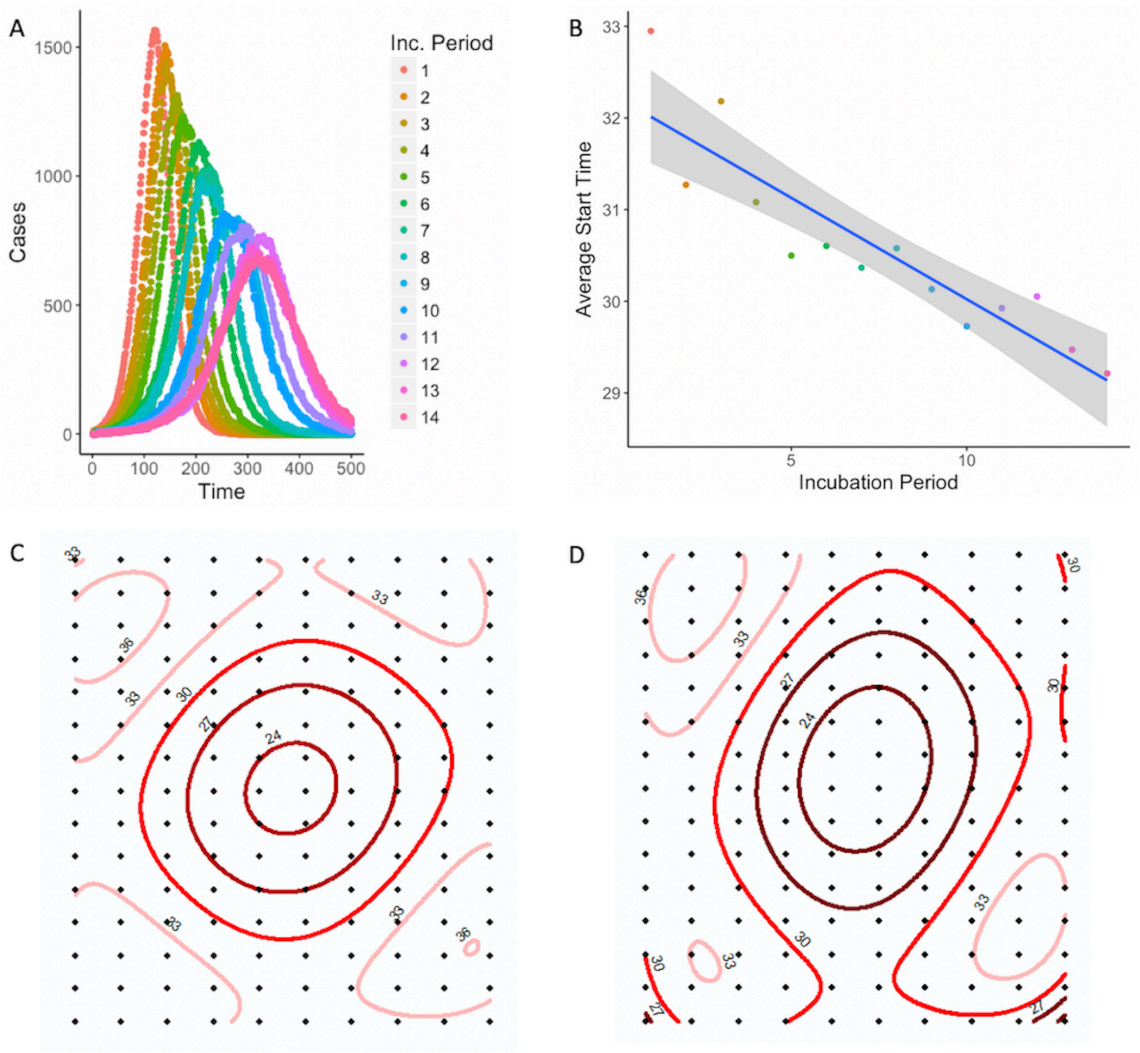
incubation periods were more acute while diseases with longer incubation periods peaked later (**Figure 2.4 A**). Although epidemics tend to last longer for diseases with longer incubation periods, the spread of the disease to more distant locations can progress more quickly, causing a discontinuous and more rapidly spreading wave front (**Figure 2.4 C-D**). In the first 50 days of our simulations, locations further from the origin of the epidemic experienced cases earlier on average in simulations with longer incubation periods compared to those with shorter incubation periods, likely due to long-distance sparking events from infected agents traveling during the incubation period (**Figure 2.4 B**). The dispersion kernel $K^x(d)$, the probability that an agent will end up at a position separated a distance d from the initial position after x days, is more homogeneously spread and has non-vanishing probabilities at greater distances the higher the incubation period (x), explaining the enhancement in sparking events (**Supplementary Figure 2.5**).

Simulations on a lattice with relative population size based on Sierra Leone's chiefdom census data (as opposed to the evenly distributed populations in the original lattice simulations) support the finding that the duration of epidemics is longer on a district (i.e. group of lattice points) rather than chiefdom (i.e. individual lattice point) scale, with duration lengthening with increasing incubation periods (**Figure 2.5 A**).

Consistent with the correlation analysis comparing Sierra Leone's cholera and Ebola outbreaks, time series from simulated outbreaks with shorter incubation periods were more highly correlated than those from simulations with longer incubation periods, with correlation decaying as distance between locations on the lattice increased (**Figure 2.5 B-C**). Higher correlation suggests increased predictability, which the results of the overlap function support (**Figure 2.6**). As the incubation period lengthened, the average predictability during the

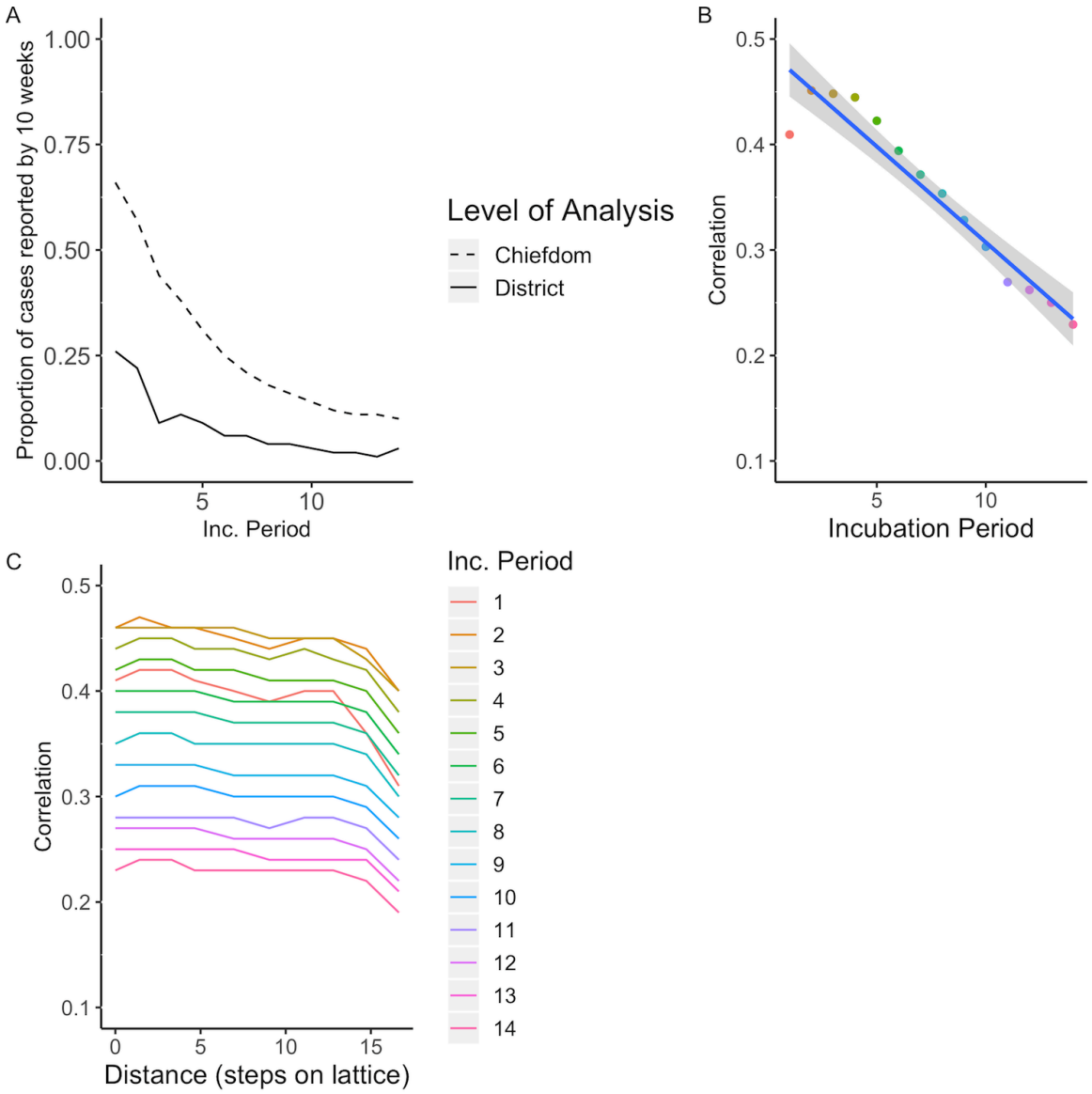
beginning of the outbreak decreased as the epidemics spread via unpredictable sparking patterns. Predictability plateaued as the outbreaks became widespread.

Figure 2.4. Simulated epidemic results



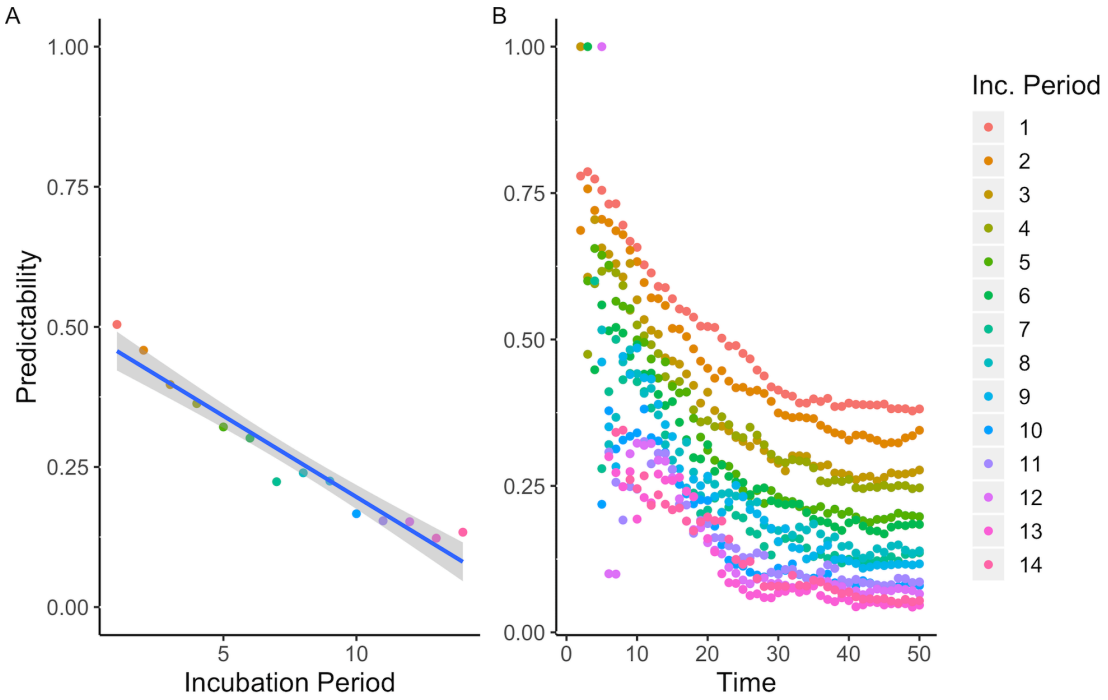
Results of 700 simulations of 14 different incubation periods show the impact of incubation period on disease spread. Epidemics with shorter incubation periods are more acute than epidemics with longer incubation periods (A). The average start time of epidemics at all locations over the first 50 days of outbreak is later for shorter incubation periods than longer (B). Spatial trend contours of first 50 days of simulated outbreaks with shorter incubation period (2 days) (C) and longer incubation period (10 days) (D), spreading from areas in dark red to light red, show that shorter incubation periods result in a more wave front spread and longer incubation periods result in more long-distance sparking events; numbers show average start day relative to start of the outbreak.

Figure 2.5. Impact of the incubation period on outbreak dynamics



The incubation period impacts the timing of outbreaks and as a result, the correlation. As the incubation period increases, the proportion of cases reported by 10 weeks, when a reactive vaccination campaign might begin, decreases in simulated epidemics (A). As the incubation period increases, the average correlation overall (B) and by distance from origin of simulated outbreaks (C) decreases.

Figure 2.6. Incubation period impact on predictability of outbreak spread



The incubation period impacts the predictability of disease spread. As the incubation period increases, the average overlap (predictability) of the first 50 days (A) and over the first 50 days (B) of simulated outbreaks decreases.

2.4 DISCUSSION

Analysis of the cholera and Ebola epidemics revealed commonalities and differences in the way these pathogens spread throughout Sierra Leone, and our simulations suggest the differences in the incubation period reproduce these differences. Spatial diffusion of Ebola occurred more quickly than cholera, as evidenced by the wave front contour lines and further supported by statistical tests considering a subset excluding cholera cases before the brief respite in June (**Supplementary Figure 2.6**). Additionally, cholera metrics were more correlated in space than Ebola metrics. Our model simulations suggest that these findings are potentially due to the counter-intuitive role of the longer incubation period for Ebola as compared to cholera. Travel during the incubation period will be a key driver of geographic disease dispersion and

predictability, especially in a population of individuals who decrease mobility when ill. Consequently, diseases with longer incubation periods will tend to have more long-distance sparking events caused by infected, but healthy, individuals traveling during the incubation period. This will result in faster epidemic dispersion to distant, unpredictable locations. These findings are in line with Marvel et al.'s results, which found epidemic wave fronts are less likely to occur for mobility kernels that decay more slowly;[34] when the incubation period is longer, the effective mobility kernel can span to more distant places, making sparking events more probable given the same number of transmission events.

Similar results were also obtained when infectious agents did not decrease mobility when ill, suggesting that travel during the incubation period has more influence on correlation and predictability than travel during the infectious period. While many other factors will influence wave speed, continuity, and epidemic synchrony, our simulations showed that small changes in the incubation period can powerfully influence epidemic dynamics. For example, environmental persistence of Vibrio cholerae in a local water source can potentially lead to a longer serial interval for local transmission[35] and a fatter right tail in offspring distribution via super-spreading. Following the dynamics of cholera and Ebola, our models assumed that the incubation and latent periods were equal; however, pre-symptomatic infectiousness may be an important factor increasing spatial heterogeneity of onward transmission especially in the context of decreasing mobility when ill. For a disease with pre-symptomatic infectiousness, we would expect to continue to see a positive correlation between long-range sparking events and the incubation period as well as a greater likelihood of intermediate-range sparking events as infectiousness increasingly precedes symptom onset and travelers transmit en route.

The incubation period has already been recognized as an important component for understanding epidemics and control,[19] with the conventional knowledge that long incubation periods allow more time for responders to scale-up interventions against the overall epidemic and are therefore advantageous for disease control efforts. Here we demonstrated a counter-intuitive mechanism whereby a longer incubation period may in fact hinder a response by decreasing the predictability of outbreaks and increasing their geographic scope as well as of the needs of surveillance and response. We use simulations to reproduce the double-edged sword of the influence of the disease incubation period on reactive interventions.

Reactive vaccination strategies exist for both cholera and Ebola outbreaks, and a better understanding of spatiotemporal spread can facilitate locally-preemptive vaccination to target locations at high risk of introduction.[36–38] Reactive vaccination campaigns must consider both the expected duration of an outbreak at a given spatial scale and the predictability of its spread. We found that both epidemics lasted longer at the district level than chiefdom level, likely due to the larger spatial scale of the districts. For cholera, we showed that chiefdom outbreaks tended to report half their cases within approximately 4 weeks, suggesting reactive vaccination of a chiefdom triggered by detection of a case may not be early enough to avert an outbreak and instead intervening at a wider scale, such as districts, might provide more favorable timing for intervention targeting. We posit for future study that regional-ring vaccination strategies may be better suited to diseases with short incubation periods, while contact-ring vaccination strategies may be better suited to diseases with longer incubation periods due to their regional unpredictability and the longer intervals between generations in infection.

There are limitations to our work with regards to data as well as methods. Few cholera cases were confirmed during the epidemic and therefore we depend on the clinical definition as

well as the cases that were detected and recorded by the surveillance system. Ebola surveillance data are similarly prone to differences in reporting rates, but the use of only confirmed cases yielded similar results to those reported above using both confirmed and suspected cases. Our estimates for the effective reproductive number depend on, and absorb the limitations of, case data, serial interval estimates, and the chiefdom connectivity matrix. Specifically, we assume all cases in our dataset acquired infection from others in the dataset, thereby excluding missing cases and asymptomatic transmitters. However, this method has been shown to be robust to cases missing at random and we furthermore expect the role of asymptomatic transmission to be limited for both diseases due to the strong correlation between pathogen load, symptoms, and infectiousness.[39,40]

Further, we assume no changes to the serial interval for either cholera or Ebola during the course of the epidemics. For cholera specifically, waterborne transmission could potentially lead to a heavy right-tail in serial intervals or change the distribution as pathogen accumulates or clears from a drinking source. Household data in Bangladesh, where the role of water contamination is expected to be large, suggest few serial intervals beyond 7 days.[41] The geographic spread of cholera in Sierra Leone from the northwest and south towards the center of the country was not consistent with the direction of key waterways in the country, which primarily run from the eastern highlands to the western shores, suggesting population density and human-to-human contact likely played a larger role than water sources in this outbreak.

Finally, our simulation model provides a proof-of-concept test of the hypothesis of the impact of the incubation period on disease spread and makes several simplifying assumptions. These assumptions could be relaxed in future work, including the complete overlap of symptoms and infectiousness and constant or structured diffusion of agents, for example without increased

probability of returning "home," which could decrease the overall distance exposed agents travel and therefore lower the probability of longer range sparking events. One could use other models for the mobility of the agents, such as the one by Song et al.[42] which includes probabilities for returning to already visited places, as well as for exploration of locations not previously visited. In general, complex travel patterns are difficult to measure in real populations and are highly context-specific, interacting in critical ways with the epidemiological drivers of epidemics examined here.

The threat of cholera and Ebola re-emergence in Sierra Leone remains a concern.[43] We have shown that differences in incubation period alone are a powerful driver of geographic dispersion and merit further study. Although this study only examines one epidemic from each disease, the size of these epidemics, combined with simulation results from our model, can lend information towards a better understanding of each disease and our ability to predict disease spread. This work can inform development of international preparedness and response strategies and ensure timely and effective interventions.

2.5 METHODS

Data

Cholera cases were reported to the Sierra Leone Ministry of Health and Sanitation by treatment facilities throughout Sierra Leone between January 1, 2012 and May 15, 2013. Following standard WHO definitions,[44] a suspected cholera case was defined as acute onset of watery diarrhea or severe dehydration in a person aged five years or older in a region without a known cholera outbreak; once the Government of Sierra Leone declared an outbreak of cholera on February 27, 2012, any case of acute watery diarrhea could henceforth be included as a

suspected cholera case. Data were compiled and anonymized by the WHO for analysis, with case reports temporally resolved by day and spatially resolved by chiefdom. For Ebola, we used a published dataset of 8,358 confirmed and 3,545 suspected Ebola cases reported to the Sierra Leone Ministry of Health and Sanitation from May 2014 to September 2015.[45] Our analysis included both suspected and confirmed cases of Ebola according to standard WHO definitions.[46] Population estimates for 2012 and 2014 were imputed by chiefdom using a linear fit between chiefdom population estimates from the 2004 and 2015 Population and Housing Censuses.[47] Data and code are available on Github.[48]

Sierra Leone has four administrative regions, which are divided into fourteen districts. Freetown, the capital and largest city, is comprised of two districts; the remaining twelve districts are subdivided into 149 chiefdoms, with a median of 11.5 chiefdoms per district. Chiefdom, as the finest administrative unit available for cases of both cholera and Ebola, was considered the unit of observation and the unit of analysis (with the exception of cases in Freetown which were solely reported at district level), as it is the likely scale of intervention campaigns like vaccination. To understand what would have been observed at a coarser spatial scale that is more common for surveillance, we additionally aggregated cases by district.

Spatiotemporal analysis

We defined the first outbreak week for each chiefdom as the week of the first reported case in that chiefdom. We visualized outbreak spread using a contour map of outbreak wave front direction and speed.[45] Contours of spatial spread were generated using ArcMap 10.3.1 Spatial Analyst extension by applying a fourth degree polynomial trend interpolation of chiefdom onset dates and generating contour lines of this surface in 2–4 week increments. With

this method, more closely-spaced contour lines indicate slower propagation, similar to the slope of a topographic map of geographic elevation.

To identify space-time clusters, using the SaTScan software package,[49] we ran a retrospective discrete Poisson-based Scan Statistic over the entirety of the outbreaks for which data were available, namely 16 months of cholera data and 17 months of Ebola data. Disease case reports were assumed to be Poisson-distributed given chiefdom population size. The unit of time aggregation for the analysis was specified as the median serial interval for each disease (5 days for cholera[50,51] and 13.3 days for Ebola[8]).

We calculated spline correlograms for four chiefdom outbreak metrics to measure spatial correlation of date of first case, case count, attack rate, and disease presence (yes/no). The maximum centroid-to-centroid distance was set to 150 km, approximately the radius of Sierra Leone. We used the spline.correlog function of the R package ncf for each disease and all chiefdom pairs.[52]

We estimated the daily effective reproductive number (R_t), the average number of onward infections generated by cases with onset on day t, using methods described by Wallinga and Teunis and extended to metapopulations by White et al.[53,54] This maximum likelihood method estimates the probability that an observed case was the infector for each subsequent case by leveraging information on the daily case count, the serial interval distribution, and a weights matrix that quantifies relative contact frequency within and between chiefdoms. The serial interval for cholera was assumed to follow a gamma distribution (rate = 0.1, shape = 0.5) with a median of five days, as has been used previously after consideration of both fast, person-to-person, and slow, environmental, transmission routes.[50,51] The serial interval for Ebola was assumed to follow a gamma distribution (rate = 0.17, shape = 2.59) with a median of 13.3 days

derived from the estimates by the WHO Ebola Response Team.[8] The contact frequency between two given chiefdoms was assumed to decrease with squared distance between the chiefdom centroids. Additional weights matrices with different functional forms for distance decay yielded qualitatively similar measurements of R_t.

Model

We simulated an agent-based model with 45,000 agents distributed equally in 150 locations, evenly spaced on a 15 x 10 lattice. Infected agents progressed through a traditional Susceptible-Exposed-Infectious-Recovered (SEIR) compartmental transmission framework. We assumed the incubation period (i.e. the time from exposure to symptom onset) overlapped completely with the latent period (i.e. the time from exposure to onset of infectiousness). Similarly, the duration of illnesses (5 days) aligned with the duration of infectiousness. The serial interval, which comprises both the incubation period and duration of infectiousness, can strongly influence epidemic dynamics. However, to isolate the impact of pre-symptomatic travel on spatiotemporal patterns of disease spread, in our simulations, we held the duration of infectiousness constant. The attack rate also remained constant throughout the epidemics, with an R_0 of 1.5; simulations with larger R_0s (e.g. 3) returned similar results. Movement of agents between two locations was simulated through a daily travel connectivity matrix **A** based on a gravity model, whereby connectivity was proportional to the population sizes of each location and the inverse squared distance between them.[55] Different parametrizations of the gravity model, as well as simulations with relative population size based on Sierra Leone's chiefdom census data,[56] yielded similar results. Note that the effective dispersion kernel after x days $K^x(d)$ mentioned in the simulation results is different from the daily mobility matrix **A**. In our simulations the mobility matrix is fixed independently of the disease, but the dispersion kernel

that is relevant for each disease depends on the incubation period. The element A_{ij} of the mobility matrix **A** describes the probability that an agent will travel from location i to location j in one day. These elements depend on the populations of those locations and the distance between them as in a gravity model for mobility. The dispersion kernel $K^x(d)$ measures the probability of finding an agent at a distance d from the place where she was x days before. Therefore the dispersion kernel is a direct consequence of the daily mobility matrix. It will tell us, for infected agents, the probability of being at a distance d from where they became infected after x days. As travel is stopped once they become infectious, the relevant dispersion kernel for each disease will be the one for which x equals the incubation period.

Susceptible, exposed, and recovered individuals had a daily probability of movement. To simulate the impact of a reduction in mobility during illness, agents in the model had their movement reduced as far as zero throughout the course of their period of infectiousness (and, equivalently, illness). Holding all other parameters constant, we conducted 700 simulations of epidemics for incubation periods ranging from 1 to 14 days. We seeded the epidemic at the same location near the center of the lattice for all simulations.

Synchrony was assessed with the R package ncf functions mSynch and Correlog.Nc,[52] which both estimate the correlation between the time series in each of the 150 locations across the 500 days of the simulations, with the latter incorporating distance.[52]. To assess the impact of the incubation period on the initial speed of spread, we calculated the average start time across all locations in the first 50 days of the outbreaks as well as at increasing distances from the location on the lattice where the outbreaks began.

To estimate the predictability of outbreak spread in space and time, we adapted an overlap function used to measure predictability of a SARS outbreak.[15] In each simulation, a

vector $\pi_j(t)$ represents the proportion of all infected individuals at time (t) who are at location (j). In a system with high predictability, $\pi_j(t)$ will be similar across simulations. The overlap between simulations I and II can be estimated by: $\Theta(t) = \sum_j \sqrt{\pi_j^I(t) * \pi_j^{II}(t)}$. $\Theta(t)$ ranges from 0 to 1 with a higher value indicating more overlap and thus more predictability. We estimated predictability at each time point by calculating the average of the overlap functions for each pair of simulations for each incubation period. We calculated the average overlap across time points to provide a summary metric for predictability of each incubation period.

Data Availability

Code and data are available on Github:[48] https://github.com/rek160/Sierra-Leone-Cholera-Ebola.

Chapter 3. Leveraging pathogen sequence and contact tracing data to enhance vaccine trials in emerging epidemics

3.1 ABSTRACT

Advance planning of the design and analysis of vaccine trials conducted during infectious disease outbreaks increases our ability to rapidly define the efficacy and potential impact of a vaccine and inform public health response. Vaccine efficacy against infectiousness (VE_I) is an important measure for understanding the full impact of a vaccine, yet it is currently not identifiable in many vaccine trial designs because it requires knowledge of the vaccination status of infectors. Recent advances in pathogen genomics have improved our ability to accurately reconstruct transmission networks. We aim to assess if augmenting classical randomized controlled trial designs with pathogen sequence and contact tracing data can permit these trials to estimate VE_I.

We develop a transmission model with a vaccine trial in an outbreak setting, incorporate pathogen sequence evolution data and sampling as well as contact tracing data, and assign probabilities to likely infectors. We then propose and evaluate the performance of an estimator of VE_I. We find that under perfect knowledge of infector-infectee pairs, we are able to accurately estimate VE_I. Use of sequence data results in imperfect reconstruction of the transmission networks, biasing estimates of VE_I towards the null, with approaches using deep sequence data performing better than approaches using consensus sequence data. Inclusion of contact tracing data reduces the bias.

Pathogen genomics enhance identifiability of VE_I from individually randomized controlled trials, but imperfect transmission network reconstruction biases the estimates towards the null and limits our ability to detect VE_I. Given the consistent direction of the bias, estimates

obtained from trials using these methods will provide lower bounds on the true VE_I. A combination of sequence and epidemiologic data results in the most accurate estimates, underscoring the importance of contact tracing in reconstructing transmission networks.

3.2 INTRODUCTION

Vaccine trials conducted during epidemics of emerging infectious diseases provide an important opportunity to test the safety and efficacy of vaccine candidates. Increasing our ability to quickly and accurately understand the impact of a vaccine candidate in the urgent setting of an outbreak is critical for enhancing public health response. The use of the ring vaccination strategy in the *Ebola ça Suffit* trial during the 2013-2016 West African Ebola outbreak highlighted the importance of developing innovative designs for trials conducted during an ongoing outbreak.[38] It also underscored the need to think through trial design and analysis strategies in advance in order to expedite the rollout of a vaccine trial once an outbreak starts and to identify the best methods for obtaining high quality efficacy estimates in outbreak settings.[57]

Multiple components of vaccine efficacy can be estimated from a vaccine trial.[58] Individually randomized controlled trials (iRCTs) estimate vaccine efficacy against susceptibility to infection (VE_S), the direct effect of the vaccine on vaccinated individuals.[58] If reducing susceptibility to infection is the only effect of the vaccine, then this measure, combined with information on contact network structure and pathogen transmission dynamics, can be used to estimate the total effect of a vaccination program, a combination of the direct and indirect (i.e. herd immunity) effects. Vaccine efficacy against infectiousness (VE_I), the reduction in onward transmission from a vaccinated person who is infected compared to an unvaccinated infected person, is another important measure for understanding the impact of a vaccine.[58] Even if a vaccine does not protect everyone who is vaccinated from getting infected, its impact on

infectiousness for those who are vaccinated but nevertheless become infected plays a critical role in both outbreak dynamics and also cost-effectiveness of a vaccine program. The significance of understanding interventions' effects on future transmission is exemplified by the efforts of HIV treatment-as-prevention programs to reduce patients' viral loads to undetectable levels in order to prevent onward transmission.[59,60]

In order to estimate VE_I, the vaccination status of infectors must be known. VE_I is therefore potentially measurable in household studies[61,62] and partner transmission studies, such as HIV vaccine trials[63] because in these settings, infector-infectee pairs can be identified (by assuming that household members or partners are the infectors), and thus the vaccination status of infectors is known. However, VE_I is not currently identifiable in population-level vaccine trials, such as those often conducted during an infectious disease outbreak, because the transmission network, and consequently the vaccination status of infectors, are typically unknown.

Recent advances in pathogen genomics have improved our ability to accurately reconstruct transmission networks.[13,64–69] The West African Ebola epidemic and the ongoing COVID-19 pandemic have demonstrated our growing capacity to use sequence data in outbreak settings,[53,70–74] and recent work has highlighted the potential for deep sequence data to add resolution to transmission networks.[75–77] We aim to assess if augmenting classical randomized controlled trial designs with pathogen sequence data, as well as contact tracing data, would permit these trials to estimate VE_I by reconstructing transmission networks and identifying the trial status of infectors.

3.3 METHODS

We define θ as the risk ratio for becoming infected if one receives vaccine vs. control, or $1 - VE_S$, and Φ as the relative infectiousness of a vaccinated person who is infected compared to a control who is infected, or $1 - VE_I$. At the conclusion of a vaccine trial, the ratio of the proportion of people infected by vaccinated individuals to the proportion of people infected by controls is a product of both the vaccine's effect on susceptibility to infection and its effect on infectiousness among those who are infected. With knowledge of who infected whom, using the ratio of infector vaccination status, we can therefore calculate VE_I:

$$\theta\Phi = (1 - VE_S)(1 - VE_I) = \frac{\#\text{ infected by vacc}/\#\text{ vacc}}{\#\text{ infected by control}/\#\text{ control}}$$

$$VE_I = 1 - \frac{\#\text{ infected by vacc}/\#\text{ vacc}}{\#\text{ infected by control}/\#\text{ control}} / (1 - VE_S)$$

We simulate a compartmental network model of an outbreak, together with a vaccine trial, the details of which have been previously described.[78] Individuals are grouped into communities, with many connections between individuals in the same community and fewer between individuals in different communities. Introduction of infection into the network occurs at a time-varying rate, and the disease natural history in the communities follows a stochastic susceptible, exposed, infectious, recovered (SEIR) model, with Ebola-like parameters (**Table 3.1**). Each individual has a daily probability of infection from their infectious contacts in the network. Individuals are enrolled into an iRCT, with 50% randomized to vaccine and 50% to control. The vaccine's efficacy against susceptibility to infection is "leaky", with 60% efficacy ($VE_S = 0.60$), meaning upon each exposure, the vaccine reduces a vaccinated individual's chance of infection by 60%. The vaccine's efficacy against infectiousness is 30%, meaning infectiousness among infected vaccinated individuals is 30% lower than among infected

unvaccinated individuals ($VE_I = 0.30$). **Table 3.2** shows the number of infections expected for each type of infector-infectee pair from the trial simulations.

Table 3.1. Parameters

Parameter	Value in baseline model	Values in supplement
R_0[25]	1.5	
Incubation period[79]	9.7 days	
Infectious period[79]	5 days	
VE_S	0.6	0.8
VE_I	0.3	0.7
Number of communities	2	
Size of community	5,000	
Probability of connection within community	0.02	
Probability of connection between communities	0.001	
Importations from main population over trial period [78]	20	
Trial length (days)	300	
Genome length[64]	18,958	
Mutation rate (per genome per generation)	0.012	0.003
Bottleneck (size of pathogen inoculum at transmission)	10	2
Cluster threshold[80]	0.2	0.1

To estimate VE_I, we first make the unrealistic assumption of complete knowledge of the transmission network, with perfect ascertainment of who infected whom and their infection and recovery times. We then relax the assumption of perfect knowledge of who infected whom. Using the R package *seedy*,[81] we incorporate pathogen evolution and sampling of both consensus and deep sequence data into the simulations, specifying parameters such as genome length, mutation rate, and bottleneck size (Table 1). As the choice of parameters, particularly mutation

rate and bottleneck size, greatly impacts our ability to reconstruct transmission networks,[75] we vary parameters across simulations to assess their impact on our ability to estimate VE_I.

Table 3.2. Expected number of infections

Infector (column) *Infectee (row)*	Vaccinee	Control	Any participant	Ratio of infectees
Vaccinee	$apq\theta^2\Phi$	$apq\theta$	$apq\theta(1+\theta\Phi)$	θ
Control	$apq\theta\Phi$	apq	$apq(1+\Phi)$	
Any participant	$apq\theta\Phi(1+\theta)$	$apq(1+\theta)$	$apq(1+\theta)\ (1+\theta\Phi)$	
Ratio of infectors	$\theta\Phi$			

A proportion *p* is randomized to vaccine and to control. In the absence of vaccination, a proportion *a* would become infected, and a proportion *2q* of all exposures to infection of participants would come from other trial participants (with *1-2q* external exposures). $\theta = 1 - VE_S$, or the risk ratio for becoming infected if one receives vaccine vs. control, and $\Phi = 1 - VE_I$, or the relative infectiousness of a vaccinated person who becomes infected to a control who becomes infected.

For each infectee we then assign a probability to each potential source of infection, based on comparisons of the sequence data for the candidate source(s) and each index case using four different approaches. In the first two approaches we use consensus sequence data. First, we assign probabilities to potential infectors based on the inverse of the genetic distance between the infectee and potential infectors. Second, we use a geometric-Poisson approximation of SNP distance to assign probabilities to potential infectors; this approach assumes genetically similar sequences are more likely to be infector-infectee pairs, while also accounting for mutation rate and times of infection.[26] Third, we weight potential infectors by the number of rare variants (i.e. minority variants not seen in the consensus sequence that are rare in the population) they share with each infectee, which may be identified through deep sequence data and has previously been shown to provide additional resolution.[75] Fourth, we combine the second and third approaches, using the consensus sequence data in the event that no shared minority variants for an infectee are identified through deep sequence data.[75] For all four approaches, we then weight the

probabilities identified through the sequence data by the probability of infection given the time of symptom onset of the infectee and potential infector(s) based on the serial interval distribution (i.e. the time between when an infector becomes symptomatic and their infectee becomes symptomatic).

Using each of these approaches, we then estimate the ratio of the number of cases infected by a vaccinated person to the number of cases infected by a control. We do this in three ways for each approach (see supplemental text 1 for more details). First, we weight each identified potential infector by the probability assigned to them and sum the probabilities by vaccination status. Second, we split the probabilities for each infectee into clusters based on the largest gap in probabilities between potential infectors.[80] If the gap is larger than the specified threshold, we use the normalized probabilities from the infector(s) in the top cluster; otherwise we exclude that infectee from the analysis. Third, we use only the vaccination status of the most likely infector(s) for each infectee. Using the estimated ratio of the trial status of the potential infectors and the estimate of VE_S from the trial, we then estimate VE_I, using the equation above. To incorporate the data from the network obtained through contact tracing efforts during an epidemic, we also conduct all of the approaches described above in a data set restricted to only potential infectors who are contacts of the infectees (i.e. connections in the network model).

We propose the following procedure for estimating the standard errors of the estimates under the approaches that perform best. For a given simulation and estimate of VE_I, we obtain a bootstrap estimate of the standard error as follows. We first sample with replacement from the infected individuals. We then construct a bootstrapped data set using each infected individual from the sample and all of their potential infectors identified by the approach. We estimate VE_I from the bootstrapped data set and then repeat these steps 100 times. The standard deviation of

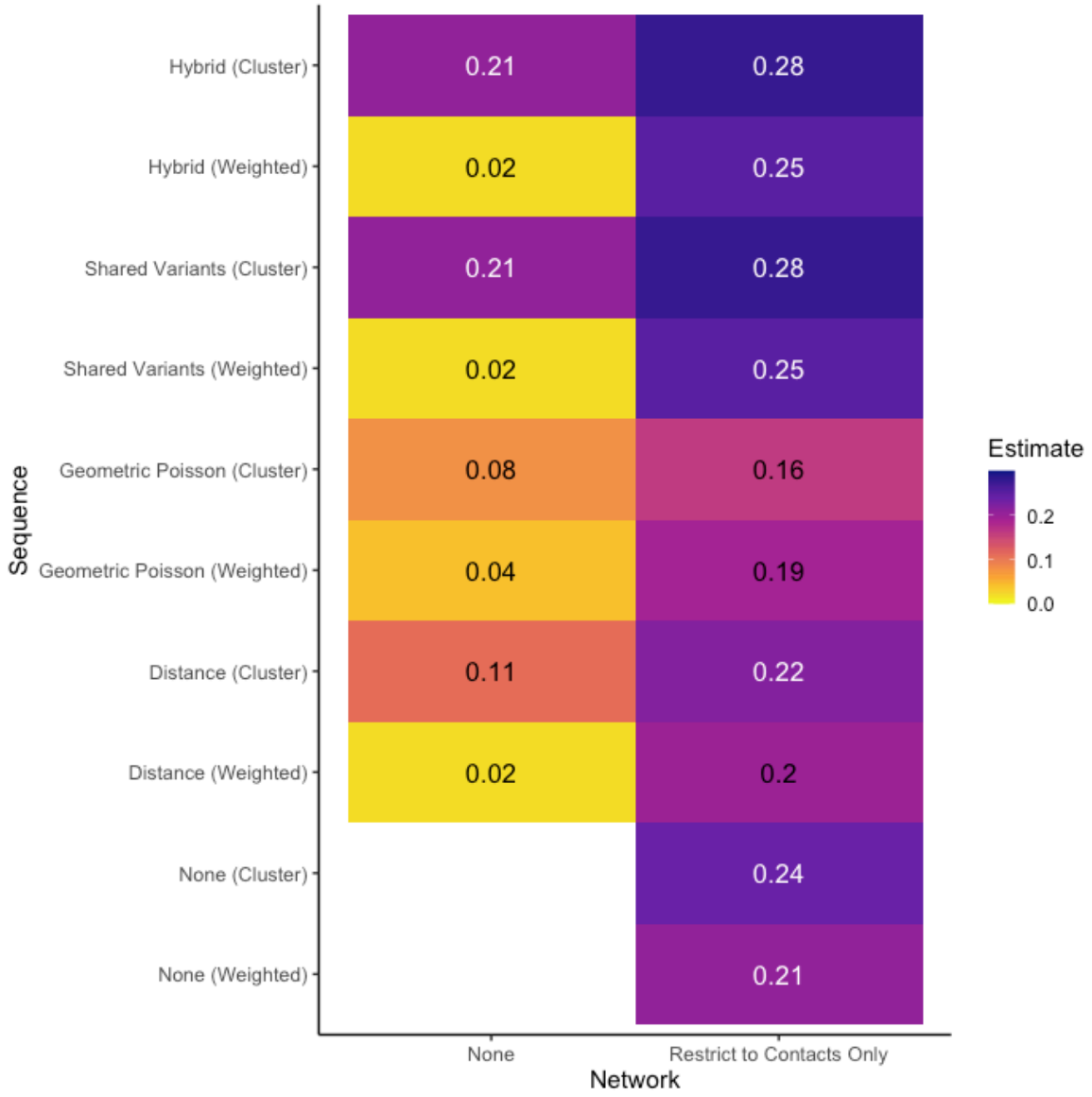
the 100 bootstrapped estimates is the standard error of the VE_I estimate. This approach could be used with real data observed in a real trial and resembles the bootstrapping clusters approach (i.e., clusters are treated as units for resampling) for clustered data.[83]

3.4 RESULTS

As expected, under perfect knowledge of the transmission network, VE_I is estimated correctly (median of 500 simulations: estimate = 0.29, standard error = 0.19), while imperfect reconstruction of the transmission networks using sequence data results in bias towards the null away from the true VE_I of 0.30 (**Figure 3.1**). This imperfect reconstruction is due to the identification of multiple potential infectors for each infectee. For example, another infectee infected by the infector of an index case may share the same number of rare variants as the index case and thus be identified in the top cluster of potential infectors. Of the methods using only sequence data, the shared variant approach using deep sequence data and the hybrid approach return results closest to the true value of VE_I, while the approaches using consensus sequence data alone return estimates closer to the null. The approaches using clustering result in more accurate estimates of VE_I (**Figure 3.1**) than the methods weighting all possible infectors, or methods using only the most likely infector(s) (**Supplementary Figure 3.1**).

In reality, sequence data are unlikely to be used in isolation, and adding epidemiologic data from the contact network decreases the bias. Using the infector(s) identified from the hybrid and shared variant approaches among potential infectors restricted to contacts results in median estimates of 0.28, close to the true value of 0.30 (**Figure 3.1** & **Supplementary Figure 3.1**).

Figure 3.1. Median VE_I Estimates



The median VE_I estimates from 500 simulations with the baseline parameters, with a true VE_I of 0.3. "None" refers to simulations that use only sequence data, without incorporating any epidemiologic information from the network. "Restrict to contacts only" restricts the analysis of sequence data to potential infectors who are contacts from the network.

The ability to accurately reconstruct transmission networks was previously found to be influenced by parameters such as the bottleneck size and the mutation rate.[75] Varying these and other parameters in our simulations show similar results to the baseline scenario

(**Supplementary Figures 3.2-3.5**), with the hybrid and shared variant approaches performing worse with a lower mutation rate (**Supplementary Figure 3.2**) and a lower bottleneck size (**Supplementary Figure 3.3**), as expected because less shared variant information is available in these settings.[75]

3.5 DISCUSSION

In the case of an outbreak of an emerging infectious disease, the ability to rapidly define the efficacy and potential impact of a vaccine is crucial for improving public health and informing policy decisions. An important component of vaccine efficacy which is often overlooked is its ability not only to guard against acquisition of infection by vaccinated individuals, but also to prevent onward transmission from those who are vaccinated that nevertheless become infected. VE_I is important for fully understanding and modeling the impact of a vaccine and both sequence and contact tracing data have the potential to allow us to estimate VE_I in large individually-randomized controlled trials conducted during an epidemic. Previously, this estimate was only attainable from household and partner studies.[61–63] Advance planning and understanding of the data requirements necessary are critical for obtaining efficacy estimates during the uncertain and urgent setting of an outbreak.

We find that while sequence and contact tracing data have the potential for enabling estimation of VE_I, misclassification of the trial status of infectors due to imperfect reconstruction of the transmission network leads to bias towards the null of VE_I estimates and overall limits our ability to detect an effect of the vaccine on infectiousness. Given the consistent direction of the bias, if an estimate is obtained in a trial using the methods described here, it is expected to be an underestimate of the true VE_I. The approaches using the top cluster of most likely infector(s) identified from the deep sequence shared variants and hybrid data perform the best of all of the

methods using sequence data alone and remain the most accurate method when contact tracing data are incorporated. If deep sequencing data are not available, relying on contact tracing data becomes even more important. The substantial improvement in the estimates when restricting to contacts further underscores the importance of contact tracing for reconstructing transmission networks.

Previous work has pointed to the potential of shared variants identified in deep sequence data, to inform transmission.[75–77] The intuition of this approach is that the pathogen population within an infected host is not composed of identical genomes, but contains some polymorphisms (depending on the population size and the mutation rate). If the transmission bottleneck is sufficiently large, more than one of these genotypes may be transmitted, and the finding that individuals share the resulting polymorphism is then a likely indication of transmission. The methods described here will therefore have variable efficacy for different pathogens. For example, influenza has a high mutation rate,[64] so there is likely sufficient phylogenetic signal and within host variation to support reconstruction of the transmission network and estimation of VE_I. Initial genomics analyses of SARS-CoV-2 found a low mutation rate;[84] recently, however, there is evidence of minority variants detectable by deep sequencing,[85] suggesting deep sequencing approaches have the potential to be used in ongoing vaccine trials to estimate VE_I.

Many simplifying assumptions have been made, which could be relaxed in future work. We assume perfect knowledge of infection and recovery times, allowing us to accurately identify the direction of transmission in infector-infectee pairs; in reality, particularly for pathogens with short incubation periods, the direction of transmission may be less clear. We also assume complete and correct sampling of sequence data (which in turn means that everyone in the community is a participant in the trial, as we assume, or at a minimum is followed up in the

trial), full knowledge of the contact network, and complete contact tracing. Approaches such as those in the *TransPhylo* R package could be used to assess where cases are likely missing and the overall proportion of the outbreak that has been sampled.[86] A naïve Bayes approach using additional data on individuals in the trial, such as demographic or geographic covariates, has been shown to improve reconstruction of the transmission network when limited sequence and/or contact tracing data are available.[87] Our methods further absorb the limitations of the *seedy* package, which assumes neutral evolution and does not permit superinfection, although this latter limitation is likely more of a concern for endemic rather than epidemic disease models.

Despite these simplifying assumptions, this work highlights the potential for existing data sources to be used in the midst of an outbreak to estimate a key measure of vaccine efficacy. It further identifies the data sources that will lead to the most accurate estimation and can thus be used for better targeting of the limited resources available for data collection in the midst of an epidemic.

Chapter 4. Potential Biases Arising from Epidemic Dynamics in Observational Seroprotection Studies

4.1 ABSTRACT

The extent and duration of immunity following SARS-CoV-2 infection are critical outstanding questions about the epidemiology of this novel virus, and studies are needed to evaluate the effects of serostatus on reinfection. Understanding the potential sources of bias and methods to alleviate biases in these studies is important for informing their design and analysis. Confounding by individual-level risk factors in observational studies like these is relatively well appreciated. Here, we show how geographic structure and the underlying, natural dynamics of epidemics can also induce noncausal associations. We take the approach of simulating serologic studies in the context of an uncontrolled or a controlled epidemic, under different assumptions about whether prior infection does or does not protect an individual against subsequent infection, and using various designs and analytic approaches to analyze the simulated data. We find that in studies assessing whether seropositivity confers protection against future infection, comparing seropositive individuals to seronegative individuals with similar time-dependent patterns of exposure to infection, by stratifying or matching on geographic location and time of enrollment, is essential to prevent bias.

4.2 INTRODUCTION

The extent and duration of immunity following SARS-CoV-2 infection are critical outstanding questions about the epidemiology of this novel virus.[88] Serologic tests, which detect the presence of antibodies, are becoming more widely available.[89] However, the presence of antibodies, or seroconversion, does not guarantee immunity to reinfection, and experimental data with other coronaviruses raise concerns that antibodies could under some circumstances enhance

future infections.[90] Studies are needed to evaluate the short and long term effects of seropositivity. Understanding the potential sources of bias and methods to alleviate biases in these studies is important for informing their design and analysis.

Serologic studies may be useful for a variety of reasons, including to assess the cumulative incidence of infection within a community, to identify risk factors for transmission, and to determine the extent of clustering of infections within a community.[91,92] While these types of studies are often cross-sectional and use seroconversion as the endpoint, we consider here longitudinal studies where seroconversion is the exposure of interest.

These seroprotection studies may be conducted by starting with a cross-sectional serological survey, where the tested individuals are then followed to identify future infections. To obtain a sufficient cohort of seropositive individuals, enrollment may need to occur on multiple days. The follow-up to identify future infections depends on regular monitoring of symptoms and/or PCR testing for the virus. Consistent case definitions across the study, as well as tracking individual enrollment and seroconversion dates, are key to reduce the risk of misclassification. If cases are defined based on symptom onset, the study outcome will be the association between seropositivity and progression to symptoms. If cases are based on virologic testing, the study outcome will be the association between seropositivity and infection. These endpoints have different public health implications and the choice should depend on the scientific question of interest.[93]

A crude analysis of this longitudinal study would compare time from enrollment to infection between those that are seropositive and those that are seronegative at enrollment. However, because seroprotection studies are observational, as the exposure (i.e., seropositivity) is not assigned at random, potential confounders must be controlled for to obtain unbiased

estimates. Studies of seropositivity and its effect on future infection are particularly prone to confounding because factors that affect someone's risk of infection and therefore their serostatus prior to enrollment (the exposure) are likely similar to factors that affect someone's risk of infection after enrollment (the outcome). For example, individuals in high-risk occupations (e.g., health care workers) are more likely to become seropositive and are more likely to be exposed again once they are seropositive.

Confounding by individual-level risk factors is relatively well appreciated. Less obvious perhaps is that geographic structure[78] or the underlying, natural dynamics of epidemics[94,95] can induce noncausal associations between an exposure and an outcome. For example, even when seropositivity confers no protection against future infection, if the overall size of an epidemic is very different in different communities, individuals in communities with small epidemics will have low prevalence of the exposure (seropositivity) and low incidence of the outcome (infection after enrollment), while individuals in communities with larger epidemics will have higher prevalence of the exposure and higher incidence of the outcome, biasing estimates of the effect of seroprotection. Bias may also occur if individuals are enrolled at different times during an epidemic. If enrollment occurs during an upward trajectory (such as the early exponential phase of an epidemic), individuals enrolled early in the epidemic will be both less likely to be seropositive (exposure) and also less likely to become infected at a given point in time after enrollment (outcome) than those with a later date of enrollment. Moreover, in an epidemic that is controlled (thus with an up-then-down trajectory of incidence) the representation of seropositive individuals will increase with time, but the rate at which these individuals experience the outcome will increase then decrease, creating potential for confounding in either direction.

In this study we take the approach of simulating such studies in the context of an uncontrolled or a controlled epidemic, under different assumptions about whether prior infection does or does not protect an individual against subsequent infection, and using various designs and analytic approaches to analyze the simulated data. By identifying the direction and comparative magnitude of bias of the estimated degree of protection relative to a known true effect of prior infection (known because we have built it into the simulations), we identify means of designing and analyzing such studies that can render them less likely to show bias due to these confounding factors. This framework of simulating studies in the context of an epidemic has been widely used to understand experimental[6] and observational[94,96] studies of risk factors and prevention interventions for infectious disease.

4.3 METHODS

We simulate a stochastic outbreak of a disease in a network of people grouped into communities, with each community's outbreak seeded by introductions over time.[78,79] For each simulation, we generate a network graph, where individuals are grouped into either one community of 10,000 people or 10 communities of 1,000 people each. People are only connected to individuals in their own community, with the probability of such a connection based on an input parameter in the simulation. For "well mixed" communities, every individual is connected to every other individual within their community, while for simulations with "clustered" communities, individuals have a limited number of connections within their community, which creates smaller sub-communities, or "clusters", by chance. In these latter simulations, individuals may have varying numbers of actual connections but all have the same expected number. The network graph of a "well mixed" community is a complete graph, while that of a "clustered" community is a random graph with uniform edge probability. In simulations with 10

communities, all communities are independent of one another, conditional on the introduction of infection from the outside. At each time step in the model, each susceptible individual has a daily probability of infection from each of their infectious contacts of $1 - e^{-\beta}$, where β is the force of infection. Hence $e^{-\beta}$ is the conditional infection-free survival probability over a single day among those at risk at the start of the day. If a subject has n infectious contacts on a given day, the force of infection is nβ and thus the day's conditional probability of infection is $1 - e^{-n\beta}$. Since the number of contacts per individual varies by simulation, β varies by simulation to keep R fixed (see Web Appendix 1). The outbreak is seeded with stochastic introductions into the communities between days one and fifty based on an external force of infection (different from β, see **Supplementary Figure 4.1**), which means in simulations with multiple communities, outbreaks may start at different times in each community, and some communities may avoid infection completely.

The disease natural history follows a Susceptible-Exposed-Infectious-Susceptible' (SEIS') model, where under the null hypothesis (i.e., no immunity) those in the S and S' compartments are equally susceptible, while under the alternative hypothesis, those in S' are less susceptible (in principle, perhaps completely immune, but in keeping with prior evidence about coronaviruses, we assume partially immune).[97,98] In simulations with partial immunity, we make the simplifying assumption that susceptibility is immediately decreased following the infectious period and remains constant over time. Seroconversion is assumed to be detectable at the end of the infectious period. We simulate scenarios with limited control measures in place (R_E=1.5) and scenarios in which control measures that reduce the force of infection per infected individual (β) are implemented at day 120 of the study period, reducing R_E from 2 to 0.8. β is set to yield these values of R_E. **Table 4.1** shows the specific numbers corresponding to these parameters of the

simulations, and Web Appendix 1 describes the generation of the network and outbreak in more detail.

Table 4.1. Parameters

Parameter	**Values**
Number of communities	1, 10
Average community size	1 community simulations: 10,000 10 community simulations: 1,000
Probability of connection with someone within the same community	Well mixed: 1 (everyone is connected to everyone in their community) Clustered: 0.002 probability per edge for 1 community and 0.02 probability per edge for 10 communities
Probability of connection with someone in another community	0
R_E[14]	Controlled: 2.0 → 0.8 Uncontrolled: 1.5
Latent period	5.6 days (gamma distribution with shape = 5, rate = 0.9)
Infectious period	10 days (gamma distribution with shape = 3, rate = 0.3)
Days of simulation	200
Day control begins	Controlled: 120 Uncontrolled: Never
Reduction in β after control	60%
Days of enrollment	Same day: 100 Different days (uncontrolled): 50, 100, 150 Different days (controlled): 100, 150
% of individuals enrolled (unmatched)	50%
Seropositivity protection	0 (null) 50% 95%

R_E = effective reproductive number

For each simulation setting (one or ten communities, well mixed or clustered communities, control measures or not, and seroprotective efficacy), we consider three sampling designs: enrolling individuals on a single day without matching (day 100), enrolling individuals on multiple days (days 50, 100, 150) without matching, and enrolling individuals on multiple days with matching of enrolled seropositive and seronegative individuals. Enrollment on multiple days may occur, for example, if different cross sectional surveys are conducted, and this study enrolls the participants in those surveys. A random sample of individuals are enrolled into the study at these specified time points over the course of the outbreak. We classify individuals as seropositive or seronegative based on their serostatus on day of enrollment into the study, and then we follow them up until they are infected or until the study period ends at day 200. In the unmatched designs, we enroll half of the individuals in each community into the study, with an equal number enrolled on each day of enrollment. In the matched designs, for every seropositive individual enrolled on each day of enrollment, we also enroll one seronegative individual on that day from the same community. This increases the balance between exposure arms but reduces the overall sample size.

For each simulation setting and sampling design, we conduct two analyses. First, we conduct an unstratified analysis in which we calculate the hazard ratio of infection comparing seropositive to seronegative individuals, using a Cox proportional hazards model with time starting from enrollment (i.e., possibly not the same calendar time if individuals enroll on different dates). Second, given the potential for stochasticity to generate heterogeneous outbreaks between communities,[78] we also conduct an analysis stratified by community and day of enrollment to prevent confounding by these variables. In this analysis, a Cox proportional hazards model with time starting from enrollment is fit with a separate baseline hazard function

for each community and day of enrollment combination, but a common hazard ratio due to seropositivity. R code for the simulations and analysis is available on Github,[99] and additional analyses examined are described in **Supplementary Text 4.2**, **Supplementary Figure 4.2**, and **Supplementary Figure 4.3**.

4.4 RESULTS

Figure 4.1 shows the results for 1,000 simulations for each of 36 combinations of parameters (see **Table 4.1**). **Figure 4.1 A–D** summarize results from simulations with limited control measures in place (R_E=1.5). **Figure 4.1 A and C** are under the null, meaning seropositivity provides no protection against reinfection ($\beta^+ = \beta^-$, where β^+ is the force of infection for contact between an infectious individual and a seropositive individual and β^- is the force of infection for contact between an infectious individual and a seronegative individual). In **Figure 4.1B and D**, seropositivity reduces susceptibility by 50% ($\beta^+ = 0.5*\beta^-$) and 95% ($\beta^+ = 0.05*\beta^-$), respectively.

Simulations are in well mixed communities, meaning everyone within a community is connected to each other, except in **Figure 4.1 C** which has random clustering within each community. This clustering leads to correlations between infection status of particular individuals close together in the network and may be understood as creating multiple smaller (albeit overlapping) "communities" within each discrete community.

For simulations with one well mixed community with the same day of enrollment for all individuals (top lines of **Figure 4.1A, B, and D**), a crude analysis returns unbiased results. If enrollment occurs on different days (**Figure 4.1 A, B, and D**, second and third lines), a crude analysis yields an upwardly biased estimate of the hazard ratio, making seropositivity appear

harmful. However, matching on day of enrollment or stratifying the analysis by day of enrollment removes this bias.

Figure 4.1. Hazard ratios

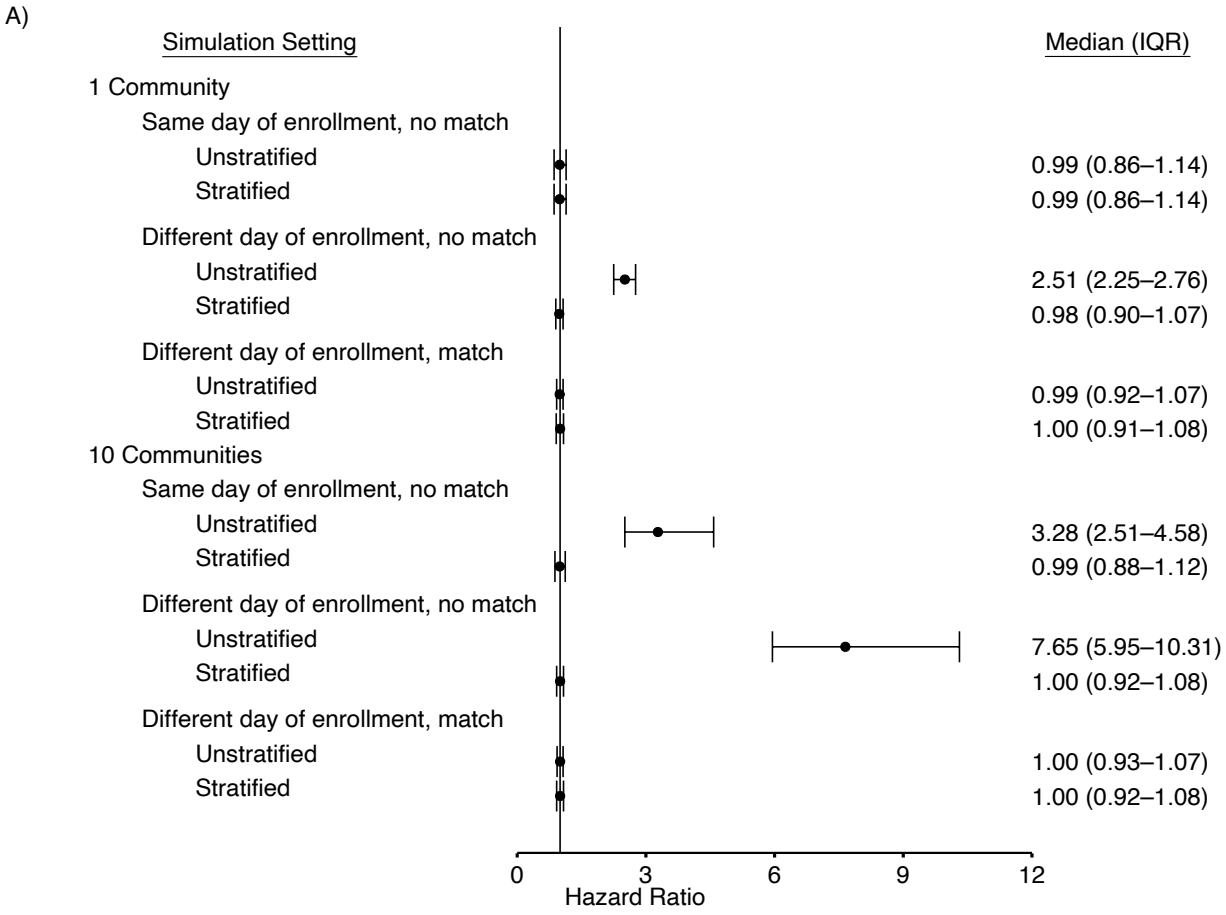


Figure 4.1. Hazard ratios (continued)

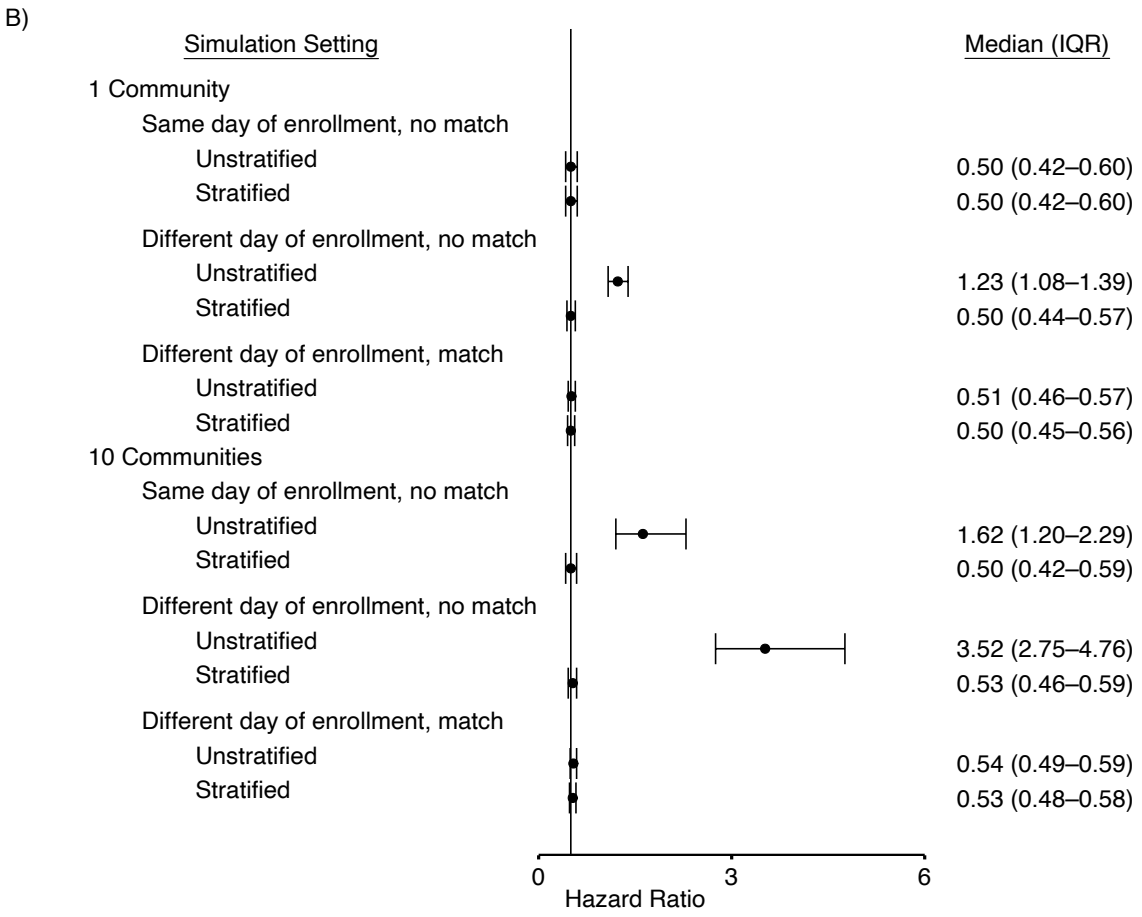


Figure 4.1. Hazard ratios (continued)

C)

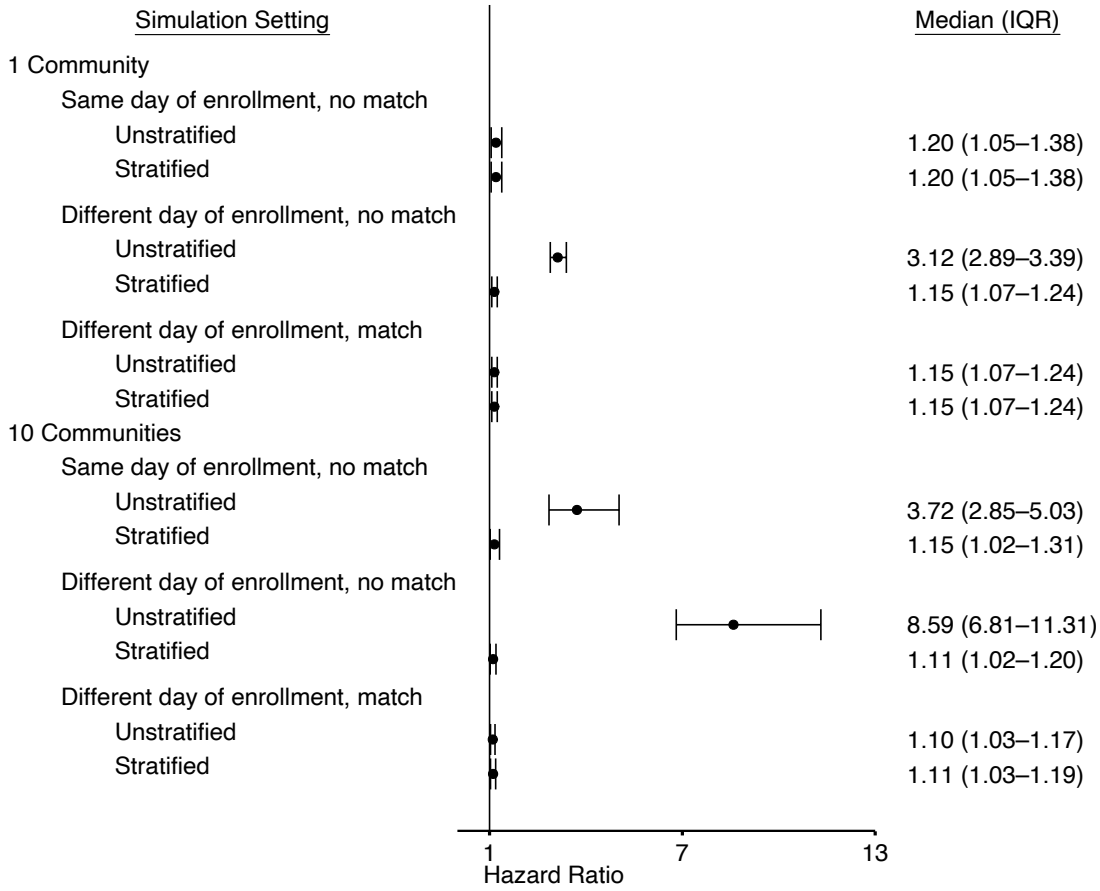


Figure 4.1. Hazard ratios (continued)

D)

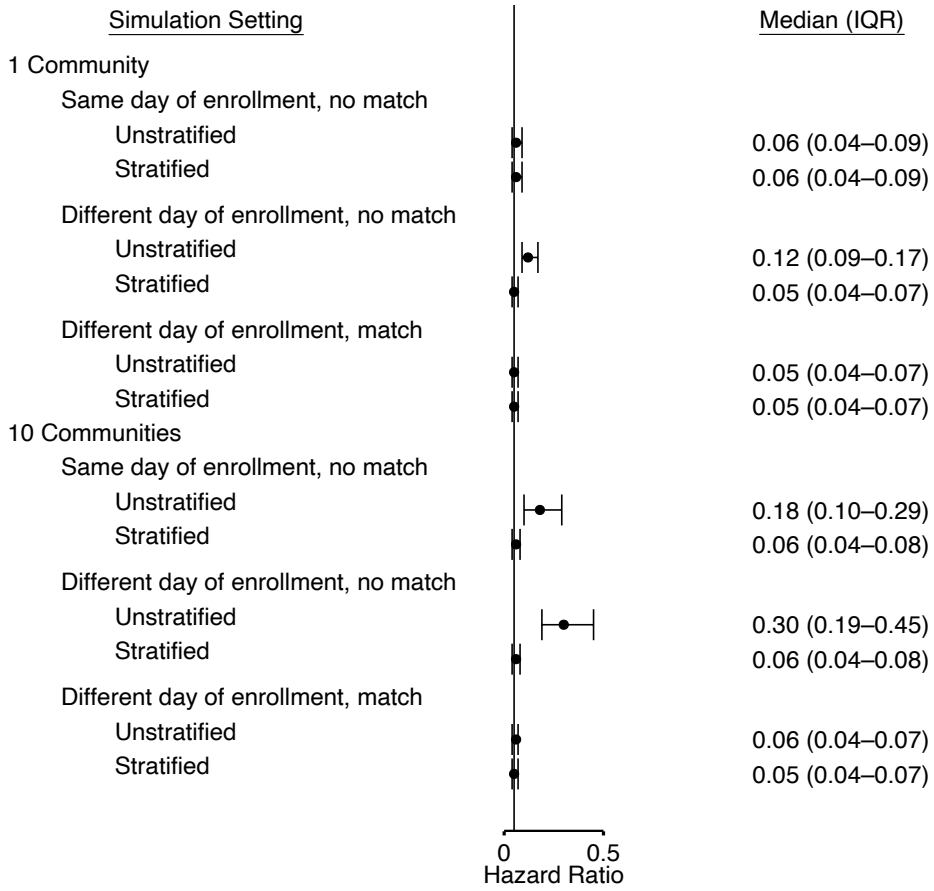


Figure 4.1. Hazard ratios (continued)

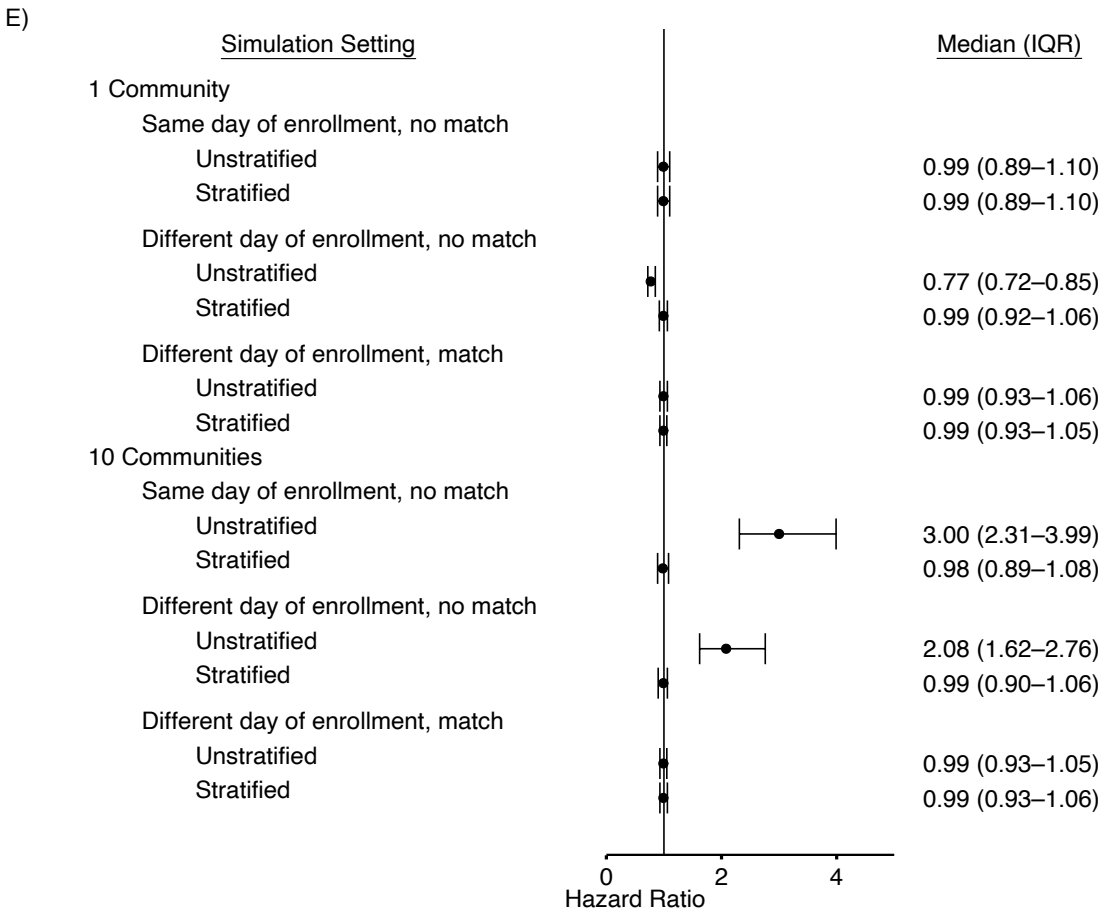
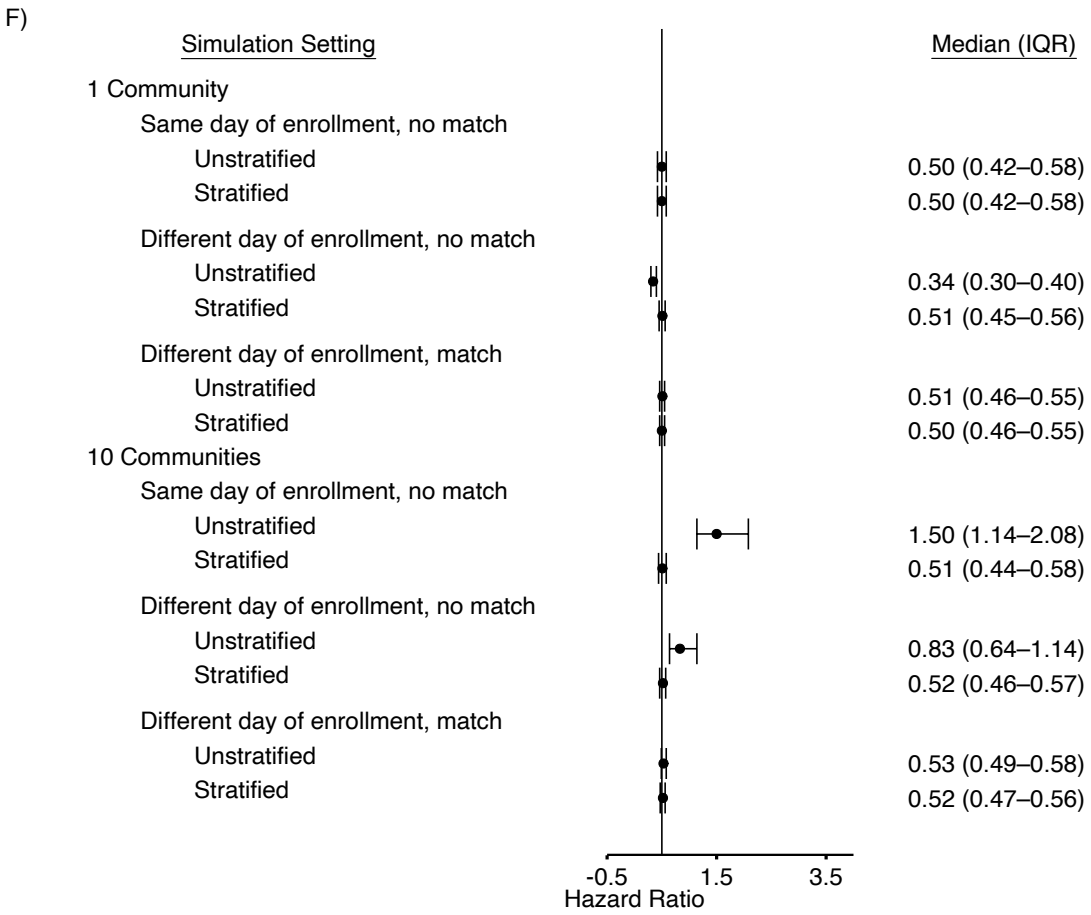


Figure 4.1. Hazard ratios (continued)



The median and IQR of estimated hazard ratios, comparing seropositives to seronegatives, for each set of simulation settings: A) well mixed communities, uncontrolled, null seroprotection; B) well mixed, uncontrolled, 50% seroprotection; C) clustered communities, uncontrolled, null seroprotection; D) well mixed, uncontrolled, 95% seroprotection; E) well mixed, controlled, null seroprotection; F) well mixed, controlled, 50% seroprotection. Note the different x-axis scales. We consider three sampling designs for each simulation setting: enrolling individuals on a single day without matching, enrolling individuals on multiple days without matching, and enrolling individuals on multiple days with matching. In the matched designs, for each seropositive individual enrolled on each enrollment day, a seronegative individual from the same community is also enrolled on that day. We compare analyses stratified by enrollment day and community (black) to unstratified analyses (grey). Simulations with zero events in either the seropositive or seronegative arm were excluded (percent of simulations excluded in each figure: A: 0.85%, B: 1.6%, C: 0.28%, D: 22.1%, E: 4.7%, F: 6.3%). For analyses with a high infection hazard for any enrolled individuals (e.g., Figures 1B, 1D, and 1F with different days of enrollment), the estimated hazard ratio is between the ratio of the force of infection between seropositive and seronegatives (β^+/β^-) and the null HR=1. This occurs because an individual's hazard is not simply the product of their number of contacts and the force of infection. This is not a bias in the conventional sense, but rather a difference between the ratio β^+/β^- and the parameter that is estimated by the Cox model (see Web Appendix 1 for more details).

With multiple communities (and thus multiple, unconnected epidemics, as in the bottom halves of **Figure 4.1A, B, and D**), an unadjusted analysis creates the same upward bias, regardless of whether enrollment is on the same or multiple calendar dates, as the same calendar date does not mean the same phase of the epidemic in each of the communities. Once again, the bias is upward because individuals in communities with larger or more advanced epidemics are exposed to higher hazards and are more likely to be seropositive at baseline (**Figure 4.2 A–D**). As before, the bias can be removed by a matched design or stratified analysis, this time matching or stratifying on both community and day of enrollment. For analyses with a high number of infectious contacts for any enrolled individuals (e.g., **Figure 4.1 B and D** with different days of enrollment), the estimated hazard ratio is between the ratio β^+/β^- and the null HR=1. This occurs because an individual's hazard is not simply the product of their number of contacts and the force of infection. This is not a bias in the conventional sense, but rather a difference between the ratio β^+/β^- and the parameter that is estimated by the Cox model (see details in **Supplementary Text 4.1**). For settings with a lower force of infection or fewer infectious contacts, this difference is imperceptible.

Clustering of contacts within communities (a departure from the assumption of a well mixed epidemic, **Figure 4.1 C**) produces an upward bias even in the matched design and stratified analyses. As noted, this reflects that the different parts of the network have different local prevalence at any given time, resulting in a milder form of the same heterogeneity-induced bias seen when there are many discrete communities. Because these clusters of high and low prevalence areas overlap and arise during the study, there is no a priori way to adjust for them.

Figure 4.2. Daily hazards

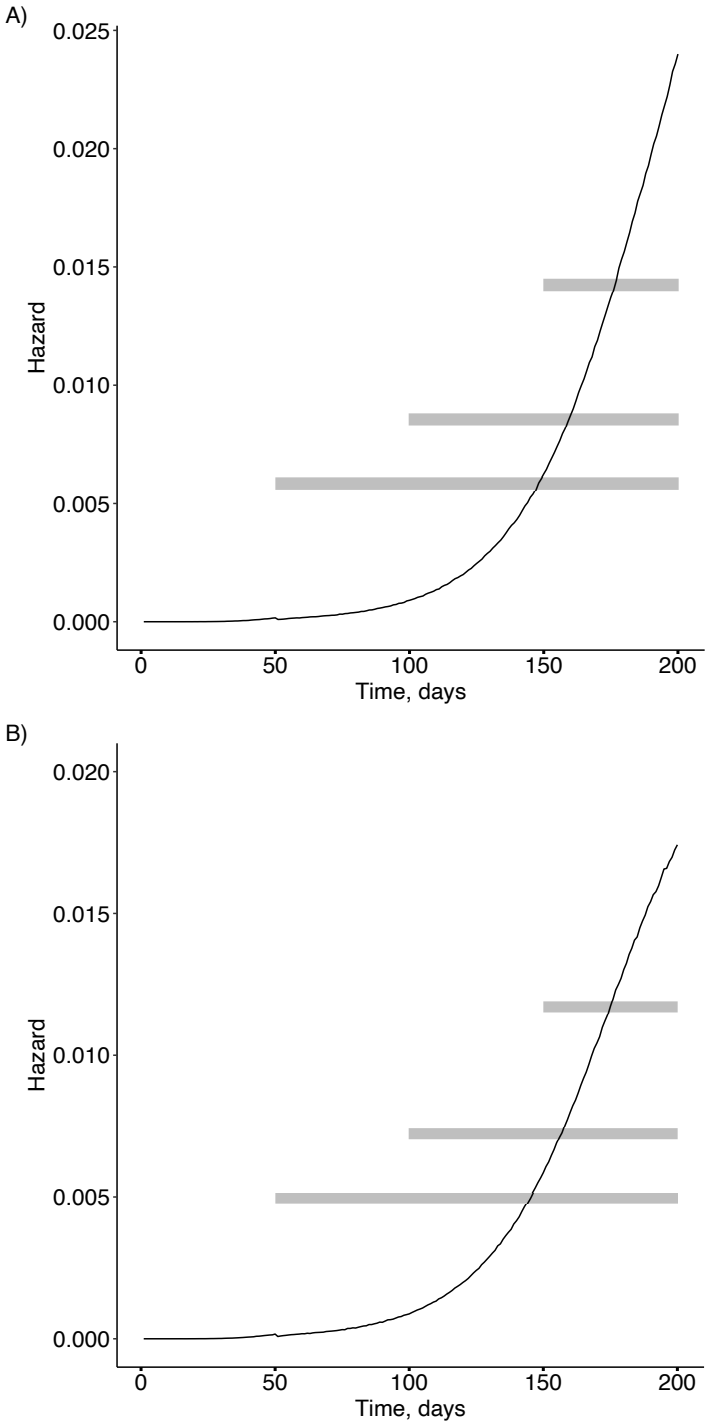


Figure 4.2. Daily Hazards continued

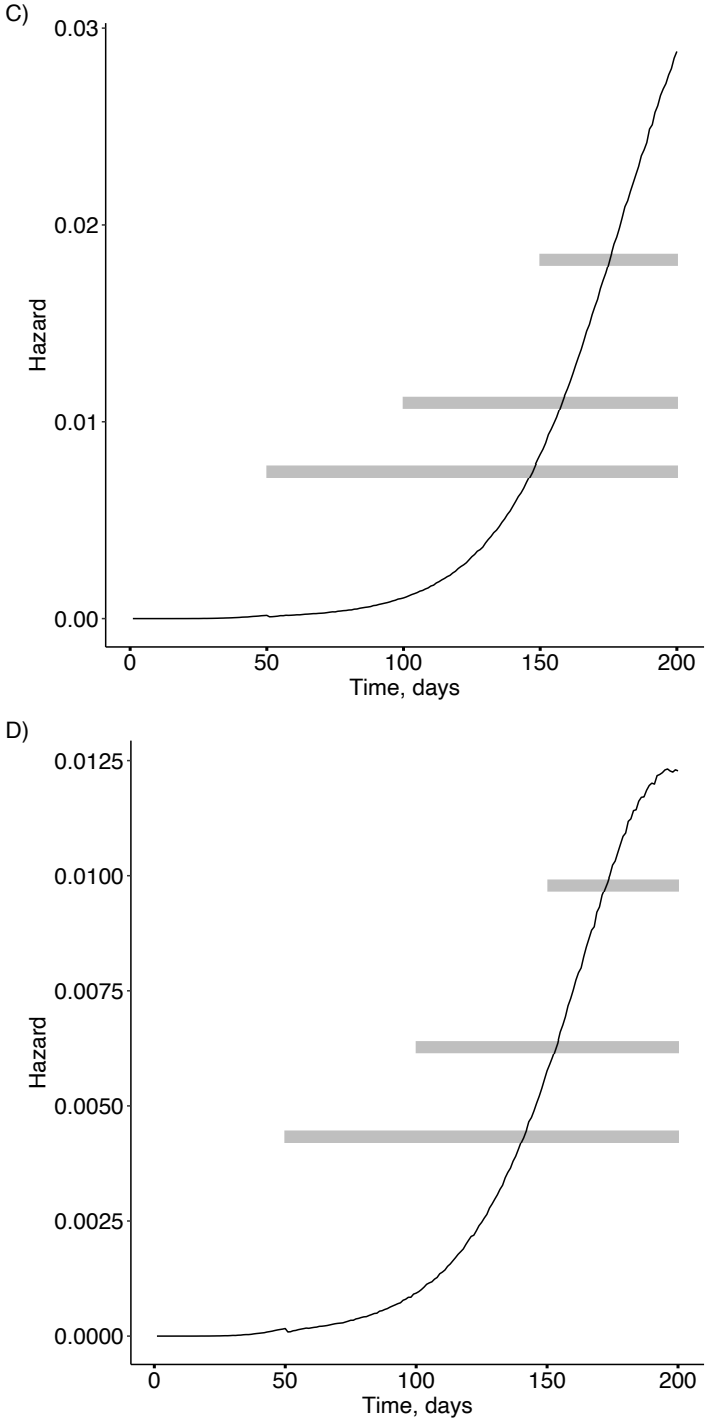


Figure 4.2. Daily Hazards continued

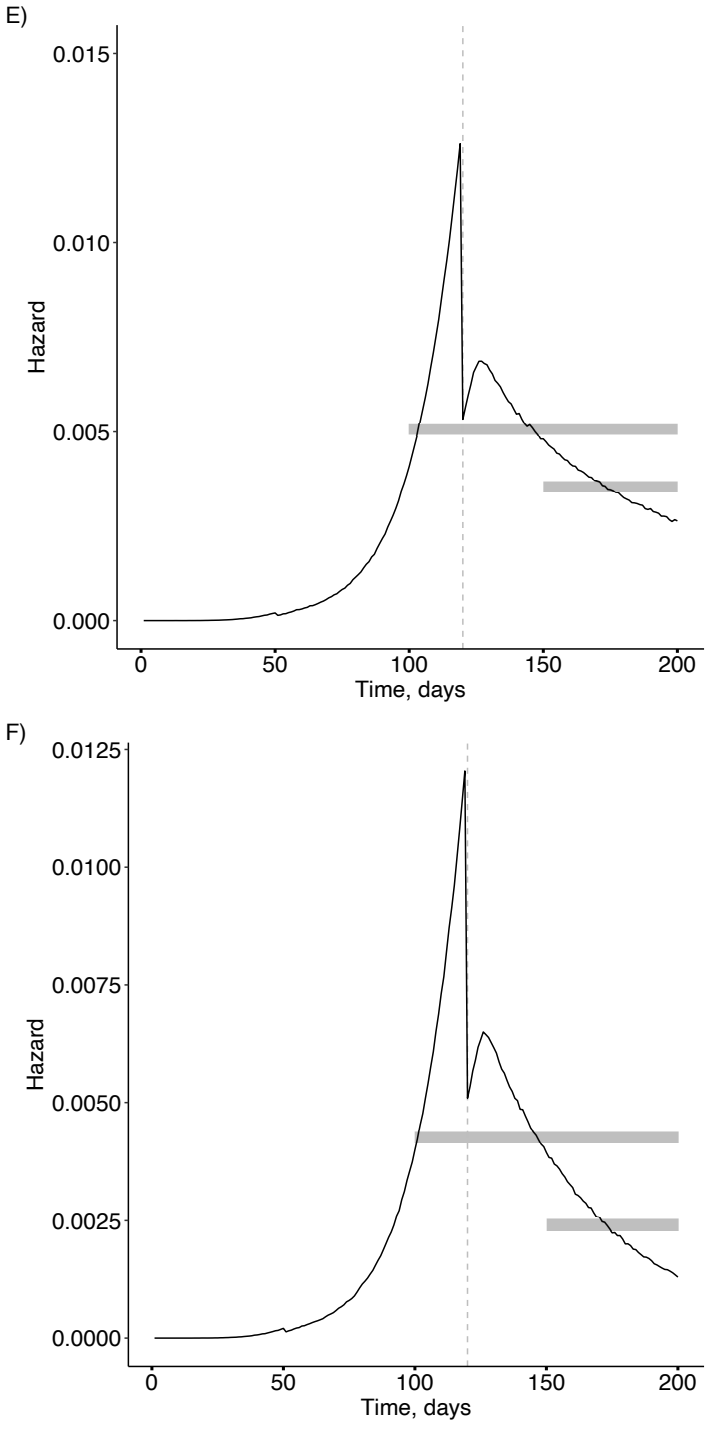


Figure 4.2. Daily Hazards continued

The average simulated daily hazard of infection for those in the initial susceptible compartment (i.e. never infected) to move to the exposed compartment in the simulations with one community: A) well mixed communities, uncontrolled, null seroprotection; B) well mixed, uncontrolled, 50% seroprotection; C) clustered communities, uncontrolled, null seroprotection; D) well mixed, uncontrolled, 95% seroprotection; E) well mixed, controlled, null seroprotection; F) well mixed, controlled, 50% seroprotection. Note the different y-axis scales. Horizontal bars show lengths of follow-up for each day of enrollment. The height of the bars indicates the average hazard for that duration of follow-up. In A–D, follow-up begins on days 50, 100, and 150, while in E and F, follow-up begins on days 100 and 150 only. Vertical grey lines denote the day control measures are implemented, which reduce the force of infection by 60% (E and F). The number of infectious individuals continues to grow beyond the day of control for approximately the average length of the latent period (5.6 days) due to those infected in the days just before control. This causes the hazard to increase again after its initial drop before declining again.

In the simulations summarized in **Figure 4.1 E F**, transmission is reduced partway through the outbreak in one or more well mixed communities, representing intensified control measures (R_E=2 → 0.8). In these simulations, there are fewer reinfections, as reflected in the wider interquartile ranges. As before, the single-community estimates are unbiased when all individuals enroll on the same day, but when enrollment occurs on different days or there are multiple communities, the estimates are biased. In the single-community simulations with two different days of enrollment, the unstratified, non-matched analysis estimates are slightly biased away from the null, making seropositivity look protective. This occurs because there are more seropositives at later enrollment dates when the average hazard over the rest of the study is lowest (**Figure 4.2 E and F**).

Hence, with multiple communities or multiple enrollment dates, confounding can go in either direction depending on the dynamics of the epidemic at the times of enrollment. Matching on enrollment alleviates the different biases, as does stratification in cases where there are infections in both the seropositive and seronegative arms. If there are substantially fewer seropositive individuals than seronegative individuals and the risk of infection after enrollment is low (i.e., because of effective control measures), there can be settings with no infections among

the seropositive enrollees in some or all strata. In these cases, stratified analyses can lead to unstable results because methods to account for one arm with zero cases (e.g., adding a case to each arm) can over-correct when the zero-case arm has far fewer individuals than the other. Matched designs are thus preferable because they remove this imbalance between the two exposure arms.

We note that in the simulations under the null with limited control measures (**Figure 4.2 A and C**), the daily hazard (proportion in the S compartment moving to the E compartment) initially increases during the early spread of the virus and then begins to plateau. In simulations with controlled epidemics and/or immunity (**Figure 4.2 B, D–F**), the daily hazard increases and then decreases.

4.5 DISCUSSION

We find that in studies assessing whether seropositivity confers protection against future infection, comparing seropositive individuals to seronegative individuals with similar time-dependent patterns of exposure to infection is essential, because otherwise confounding can bias results; accounting for differential exposure among seropositive individuals and seronegative individuals is necessary to prevent bias. This bias can arise from either having multiple days of enrollment over the course of the study by design or by having multiple communities where the outbreak stochastically starts at different times. Matching in the design or stratifying in the analysis on community and day of enrollment alleviates this bias in well mixed communities. When there is clustering within communities, a slight upward bias remains, suggesting the local network structure in a study is an important factor to consider.

While most individuals are susceptible when they are enrolled into the study, it is possible for individuals to be exposed or infectious upon enrollment. Excluding individuals who are infected soon after enrollment (e.g., within the average latent period length) would remove many of these cases. For potentially asymptomatic infections, these cases would not be able to be excluded in a study without viral testing for active infection. Small biases may occur if all individuals enrolled in the study are not susceptible at enrollment.

The results shown here assume perfect specificity of the serologic test. As expected,[100] imperfect specificity causes bias towards the null (**Supplementary Text 4.3 and Supplementary Figure 4.4**). More complex interactions of immunity and infection, including immunity that wanes over the time scale of the study, viral-load dependent infection, and effects of repeated exposures, such as boosting of titers, may affect these biases as well, or introduce other potential biases. Further research is needed to understand the effects of these biological mechanisms in the specific context of SARS-CoV-2.

These simulations focus on the bias inherent in some study designs that may be considered, but do not address the feasibility of implementing these designs. In addition, we do not focus on the power of these studies; this may have important consequences in determining an adequate sample size. Sample size considerations will be particularly important in balancing the advantage of starting enrollment later, when the cumulative incidence is higher and thus the exposure arms are more likely to be balanced, and avoiding the tail of an outbreak or a setting after control measures have been implemented, which will reduce the infection risk for all participants. We have shown that matching can address these issues, but matching requires exposure status to be known at enrollment. This may be feasible if the study is designed following a serological survey, where individuals can be enrolled on the basis of their antibody

presence from the survey. If the exposure needs to be measured for the seroprotection study, however, matching may require far more serologic testing to be conducted, inflating the cost of the study. Investigators will need to consider the relative sample size requirements and testing burden of these designs in the context of their specific study.

As serologic studies begin, understanding potential sources of bias and how to alleviate them are important for accurately estimating the extent and duration of immunity to SARS-CoV-2 (**Table 4.2**). Here we have focused on the impact of epidemic dynamics on estimation of seroprotection and have assumed all individuals in the model are exchangeable and differ only in whom they contact. Future work could examine additional heterogeneity, such as behaviors or factors that increase risk of infection, which might lead to further biases.

Table 4.2. Bias Summary

Cause of bias	**Direction of bias**	**Ways to correct**
Multiple communities with different timing of epidemics	Upward	Matched design or stratified analysis (matching works better when both number of seropositives and risk of infection are low)
Different days of enrollment	Upward or downward	Matched design or stratified analysis (matching works better when both number of seropositives and risk of infection are low)
Clustered communities	Upward	Cannot correct a priori but could consider matching on household or neighborhood

Chapter 5. Conclusion

Infectious disease modeling has the potential to help us prepare for and respond to epidemics of emerging infectious diseases. In these chapters, we showed how models can be used to increase our understanding of disease dynamics during epidemics and enhance our ability to respond.

In the second chapter, we showed how analyzing the spatial and temporal spread of past outbreaks can shed light on transmission dynamics and help inform future response. Through simulations, inspired by trends observed in empirical data, we showed the impact a pathogen's incubation period can have on outbreak trajectory. While historically longer incubation periods have been thought to allow more time to prepare, this model shows they in fact can cause the outbreak to spread further, faster, and in less predictable ways than shorter incubation periods.

In the third chapter, we highlighted the importance of preparing for the design and analysis of vaccine trials in advance in order to understand and prioritize data that should be collected during a trial. We proposed and evaluated an estimator for vaccine efficacy against infectiousness and identified deep sequencing and contact tracing data as the most important for estimating this measure.

In the fourth chapter, we identified potential biases that can arise during seroprotection studies conducted during the ongoing COVID-19 pandemic. We showed that adjusting for or matching on time of enrollment and geographic location reduces confounding by epidemic dynamics. These methods will increase our understanding of immunity to this novel virus.

Models allow for flexible frameworks for evaluating key questions and testing assumptions. They can also be continuously updated as additional information becomes

available, allowing us to dynamically respond to outbreaks. Understanding how outbreaks spread and what makes them more or less predictable, whether vaccines have an impact on infectiousness, and if past infection protects against reinfection are all critical questions when working to stop outbreaks of emerging infectious diseases. The results from the studies described here can be used to improve surveillance systems amidst outbreaks, to enhance the design and analysis of trials conducted during outbreaks, and to prioritize public health resources to prepare for and prevent epidemics.

References

1. Huber C, Finelli L, Stevens W. The Economic and Social Burden of the 2014 Ebola Outbreak in West Africa. J Infect Dis [Internet]. 2018 Nov 22;218(Supplement_5):S698–704. Available from: https://doi.org/10.1093/infdis/jiy213
2. Centers for Disease Control and Prevention. 40 years of Ebola virus disease around the world. 2019.
3. Centers for Disease Control and Prevention. Zika Virus - Transmission and Risks [Internet]. 2017. Available from: https://www.cdc.gov/zika/transmission/index.html
4. Centers for Disease Control and Prevention. Coronavirus Disease 2019 (COVID-19) [Internet]. 2020. Available from: https://www.cdc.gov/coronavirus/2019-ncov/if-you-are-sick/care-for-someone.html
5. Holmdahl I, Buckee C. Wrong but useful—what covid-19 epidemiologic models can and cannot tell us. N Engl J Med. 2020;
6. Halloran ME, Auranen K, Baird S, Basta NE, Bellan SE, Brookmeyer R, et al. Simulations for designing and interpreting intervention trials in infectious diseases. BMC Med. 2017;15(1):223.
7. Rivers C, Chretien J-P, Riley S, Pavlin JA, Woodward A, Brett-Major D, et al. Using "outbreak science" to strengthen the use of models during epidemics. Nat Commun. 2019;10(1):1–3.
8. WHO Ebola Response Team. West African Ebola Epidemic after One Year — Slowing but Not Yet under Control. N Engl J Med [Internet]. 2014 Dec 24;372(6):584–7. Available from: http://dx.doi.org/10.1056/NEJMc1414992
9. Gates B. The Next Epidemic — Lessons from Ebola. N Engl J Med [Internet]. 2015 Mar 18;372(15):1381–4. Available from: https://doi.org/10.1056/NEJMp1502918
10. Desai AN, Kraemer MUG, Bhatia S, Cori A, Nouvellet P, Herringer M, et al. Real-time Epidemic Forecasting: Challenges and Opportunities. Heal Secur [Internet]. 2019 Aug 1;17(4):268–75. Available from: https://doi.org/10.1089/hs.2019.0022
11. Kahn R, Mahmud AS, Schroeder A, Aguilar Ramirez LH, Crowley J, Chan J, et al. Rapid Forecasting of Cholera Risk in Mozambique: Translational Challenges and Opportunities. Prehosp Disaster Med [Internet]. 2019/09/03. :1–6. Available from: https://www.cambridge.org/core/article/rapid-forecasting-of-cholera-risk-in-mozambique-translational-challenges-and-opportunities/3A202E7428DB72335DFA6D828CF67504
12. Scarpino S V, Petri G. On the predictability of infectious disease outbreaks. Nat Commun. 2019;10(1):898.
13. Kraemer MUG, Cummings DAT, Funk S, Reiner RC, Faria NR, Pybus OG, et al. Reconstruction and prediction of viral disease epidemics. Epidemiol Infect [Internet]. 2018/11/05. :1–7. Available from: https://www.cambridge.org/core/article/reconstruction-and-prediction-of-viral-disease-epidemics/6CC411ED2ED7A896FCFF27F51B28FFCD

14. Meyers LA, Pourbohloul B, Newman MEJ, Skowronski DM, Brunham RC. Network theory and SARS: Predicting outbreak diversity. Vol. 232, Journal of Theoretical Biology. 2005. p. 71–81.

15. Colizza V, Barrat A, Barthélemy M, Vespignani A. Predictability and epidemic pathways in global outbreaks of infectious diseases: the SARS case study. BMC Med. 2007;5(1):34.

16. Funk S, Camacho A, Kucharski AJ, Eggo RM, Edmunds WJ. Real-time forecasting of infectious disease dynamics with a stochastic semi-mechanistic model. Epidemics. 2018;22:56–61.

17. Chretien J-P, Riley S, George DB. Mathematical modeling of the West Africa Ebola epidemic. Elife. 2015;4:e09186.

18. Viboud C, Sun K, Gaffey R, Ajelli M, Fumanelli L, Merler S, et al. The RAPIDD ebola forecasting challenge: Synthesis and lessons learnt. Epidemics. 2018;22:13–21.

19. Fraser C, Riley S, Anderson RM, Ferguson NM. Factors that make an infectious disease outbreak controllable. Proc Natl Acad Sci. 2004;101(16):6146–51.

20. Peak CM, Childs LM, Grad YH, Buckee CO. Comparing nonpharmaceutical interventions for containing emerging epidemics. Proc Natl Acad Sci [Internet]. 2017 Apr 11;114(15):4023 LP – 4028. Available from: http://www.pnas.org/content/114/15/4023.abstract

21. Wallinga J, Lipsitch M. How generation intervals shape the relationship between growth rates and reproductive numbers. Proceedings Biol Sci [Internet]. 2007 Feb 22;274(1609):599–604. Available from: https://www.ncbi.nlm.nih.gov/pubmed/17476782

22. Haw DJ, Cummings DAT, Lessler J, Salje H, Read JM, Riley S. Differential mobility and local variation in infection attack rate. PLOS Comput Biol [Internet]. 2019 Jan 22;15(1):e1006600. Available from: https://doi.org/10.1371/journal.pcbi.1006600

23. Nalin DR, Hirschhorn N. Ebola and Cholera. Am J Trop Med Hyg. 2015;92(5):1081.

24. Lamontagne F, Fowler RA, Adhikari NK, Murthy S, Brett-Major DM, Jacobs M, et al. Evidence-based guidelines for supportive care of patients with Ebola virus disease. Lancet. 2017;6736(17).

25. Althaus CL. Estimating the Reproduction Number of Ebola Virus (EBOV) During the 2014 Outbreak in West Africa. Public Libr Sci Curr Outbreaks. 2014;1–9.

26. Mukandavire Z, Morris JG. Modeling the Epidemiology of Cholera to Prevent Disease Transmission in Developing Countries. Microbiol Spectr [Internet]. 2015 Jun;3(3):10.1128/microbiolspec.VE-0011–2014. Available from: https://www.ncbi.nlm.nih.gov/pubmed/26185087

27. WHO. Ebola virus disease [Internet]. 2019. Available from: https://www.who.int/news-room/fact-sheets/detail/ebola-virus-disease

28. WHO. Cholera: mechanism for control and prevention [Internet]. 2011. Available from: https://www.who.int/cholera/technical/secretariat_report/en/

29. WHO. Cholera Fact Sheet [Internet]. 2019. Available from: https://www.who.int/news-room/fact-sheets/detail/cholera

30. Victory KR, Coronado F, Ifono SO, Soropogui T, Dahl BA, Centers for Disease C, et al. Ebola transmission linked to a single traditional funeral ceremony - Kissidougou, Guinea, December, 2014-January 2015. MMWR Morb Mortal Wkly Rep [Internet]. 2015;64(14):386–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25879897

31. CDC. Global Water, Sanitation, & Hygiene (WASH) [Internet]. 2016. Available from: https://www.cdc.gov/healthywater/global/wash_statistics.html

32. Peak CM, Wesolowski A, zu Erbach-Schoenberg E, Tatem AJ, Wetter E, Lu X, et al. Population mobility reductions associated with travel restrictions during the Ebola epidemic in Sierra Leone: use of mobile phone data. Int J Epidemiol [Internet]. 2018 Jun 26;47(5):1562–70. Available from: https://doi.org/10.1093/ije/dyy095

33. Azman AS, Rudolph KE, Cummings DAT, Lessler J. The incubation period of cholera: A systematic review. J Infect [Internet]. 2013 May 29;66(5):432–8. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3677557/

34. Marvel SA, Martin T, Doering CR, Lusseau D, Newman MEJ. The small-world effect is a modern phenomenon. arXiv Prepr arXiv13102636. 2013;

35. Azman AS, Luquero FJ, Ciglenecki I, Grais RF, Sack DA, Lessler J. The Impact of a One-Dose versus Two-Dose Oral Cholera Vaccine Regimen in Outbreak Settings: A Modeling Study. PLOS Med [Internet]. 2015 Aug 25;12(8):e1001867. Available from: https://doi.org/10.1371/journal.pmed.1001867

36. Finger F, Bertuzzo E, Luquero FJ, Naibei N, Touré B, Allan M, et al. The potential impact of case-area targeted interventions in response to cholera outbreaks: A modeling study. von Seidlein L, editor. PLOS Med. 2018 Feb;15(2):e1002509.

37. Azman AS, Parker LA, Rumunu J, Tadesse F, Grandesso F, Deng LL, et al. Effectiveness of one dose of oral cholera vaccine in response to an outbreak: a case-cohort study. Lancet Glob Heal. 2016;4(11):e856–63.

38. Henao-Restrepo AM, Camacho A, Longini IM, Watson CH, Edmunds WJ, Egger M, et al. Efficacy and effectiveness of an rVSV-vectored vaccine in preventing Ebola virus disease: final results from the Guinea ring vaccination, open-label, cluster-randomised trial (Ebola Ça Suffit!). Lancet. 2017;389(10068):505–18.

39. Glynn JR, Bower H, Johnson S, Houlihan CF, Montesano C, Scott JT, et al. Asymptomatic infection and unrecognised Ebola virus disease in Ebola-affected households in Sierra Leone: a cross-sectional study using a new non-invasive assay for antibodies to Ebola virus. Lancet Infect Dis [Internet]. 2018 Apr 10;17(6):645–53. Available from: http://dx.doi.org/10.1016/S1473-3099(17)30111-1

40. Nelson EJ, Harris JB, Morris JG, Calderwood SB, Camilli A. Cholera transmission: the host, pathogen and bacteriophage dynamic. Nat Rev Microbiol [Internet]. 2009 Oct;7(10):10.1038/nrmicro2204. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3842031/

41. Weil AA, Khan AI, Chowdhury F, LaRocque RC, Faruque ASG, Ryan ET, et al. Clinical Outcomes in Household Contacts of Patients with Cholera in Bangladesh. Clin Infect Dis [Internet]. 2009;49(10):1473–9. Available from: https://academic.oup.com/cid/article-lookup/doi/10.1086/644779

42. Song C, Koren T, Wang P, Barabási A-L. Modelling the scaling properties of human mobility. Nat Phys. 2010;6(10):818.

43. WHO. Sierra Leone to begin cholera vaccination drive in disaster-affected areas. 2017;

44. Global Task Force on Cholera Control. Prevention and control of cholera outbreaks: WHO policy and recommendations. WHO.

45. Fang L-Q, Yang Y, Jiang J-F, Yao H-W, Kargbo D, Li X-L, et al. Transmission dynamics of Ebola virus disease and intervention effectiveness in Sierra Leone. Proc Natl Acad Sci [Internet]. 2016;113(16):4488–93. Available from: http://www.pnas.org/lookup/doi/10.1073/pnas.1518587113

46. World Health Organization. Case definition recommendations for Ebola or Marburg virus diseases. 2014.

47. Statistics Sierra Leone. Statistics Sierra Leone Publications [Internet]. Available from: https://www.statistics.sl/index.php/census.html

48. Kahn R. Sierra Leone Cholera & Ebola [Internet]. Available from: https://github.com/rek160/Sierra-Leone-Cholera-Ebola

49. Kulldorff M, Inc IMS. SaTScan (TM) v7. 0: Software for the spatial and space-time scan statistics. Inf Manag Serv Inc) Available http//satscan org [Verified 5 Oct 2009]. 2006;

50. Azman AS, Rumunu J, Abubakar A, West H, Ciglenecki I, Helderman T, et al. Population-level effect of cholera vaccine on displaced populations, South Sudan, 2014. Emerg Infect Dis. 2016;22(6):1067–70.

51. Azman AS, Luquero FJ, Rodrigues A, Palma PP, Grais RF, Banga CN, et al. Urban Cholera Transmission Hotspots and Their Implications for Reactive Vaccination: Evidence from Bissau City, Guinea Bissau. PLoS Negl Trop Dis. 2012;6(11).

52. Bjørnstad ON, Ims RA, Lambin X. Spatial population dynamics: Analyzing patterns and processes of population synchrony. Trends Ecol Evol. 1999;14(11):427–32.

53. Wallinga J, Teunis P. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. Am J Epidemiol. 2004;160(6):509–16.

54. White LF, Archer B, Pagano M. Estimating the reproductive number in the presence of spatial heterogeneity of transmission patterns. International Journal of Health Geographics. 2013;35.

55. Dudas G, Carvalho LM, Bedford T, Tatem AJ, Baele G, Faria NR, et al. Virus genomes reveal factors that spread and sustained the Ebola epidemic. Nature. 2017 Apr;544(7650):309–15.

56. Statistics Sierra Leone. 2015 Population and Housing Census [Internet]. 2016. Available

from: https://www.statistics.sl/images/StatisticsSL/Documents/final-results_-2015_population_and_housing_census.pdf

57. Kahn R, Rid A, Smith PG, Eyal N, Lipsitch M. Choices in vaccine trial design in epidemics of emerging infections. PLOS Med [Internet]. 2018 Aug 7;15(8):e1002632. Available from: https://doi.org/10.1371/journal.pmed.1002632

58. Halloran ME, Longini IM, Struchiner CJ. Design and Analysis of Vaccine Studies: Introduction. Vol. 36. Springer; 2009.

59. Eisinger RW, Dieffenbach CW, Fauci AS. HIV viral load and transmissibility of HIV infection: undetectable equals untransmittable. Jama. 2019;321(5):451–2.

60. Wertheim JO, Kosakovsky Pond SL, Little SJ, De Gruttola V. Using HIV Transmission Networks to Investigate Community Effects in HIV Prevention Trials. PLoS One. 2011;6(11):1–7.

61. Datta S, Halloran ME, Longini Jr IM. Efficiency of estimating vaccine efficacy for susceptibility and infectiousness: randomization by individual versus household. Biometrics. 1999;55(3):792–8.

62. Préziosi M-P, Halloran ME. Effects of pertussis vaccination on transmission: vaccine efficacy for infectiousness. Vaccine. 2003;21(17–18):1853–61.

63. Longini IMJ, Datta S, Halloran ME. Measuring Vaccine Efficacy for Both Susceptibility to Infection and Reduction in Infectiousness for Prophylactic HIV-1 Vaccines. JAIDS J Acquir Immune Defic Syndr [Internet]. 1996;13(5). Available from: https://journals.lww.com/jaids/Fulltext/1996/12150/Measuring_Vaccine_Efficacy_for_Both_Susceptibility.7.aspx

64. Campbell F, Strang C, Ferguson N, Cori A, Jombart T. When are pathogen genome sequences informative of transmission events? PLoS Pathog. 2018;14(2):e1006885.

65. Campbell F, Cori A, Ferguson N, Jombart T. Bayesian inference of transmission chains using timing of symptoms, pathogen genomes and contact data. PLoS Comput Biol. 2019;15(3):e1006930.

66. Didelot X, Fraser C, Gardy J, Colijn C. Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. Mol Biol Evol. 2017;34(4):997–1007.

67. Jombart T, Eggo RM, Dodd PJ, Balloux F. Reconstructing disease outbreaks from genetic data: a graph approach. Heredity (Edinb). 2011;106(2):383.

68. Jombart T, Cori A, Didelot X, Cauchemez S, Fraser C, Ferguson N. Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. PLoS Comput Biol. 2014;10(1):e1003457.

69. Kenah E, Britton T, Halloran ME, Longini Jr IM. Molecular infectious disease epidemiology: survival analysis and algorithms linking phylogenies to transmission trees. PLoS Comput Biol. 2016;12(4):e1004869.

70. Emmett KJ, Lee A, Khiabanian H, Rabadan R. High-resolution genomic surveillance of

2014 ebolavirus using shared subclonal variants. PLoS Curr. 2015;7.

71. Gire SK, Goba A, Andersen KG, Sealfon RSG, Park DJ, Kanneh L, et al. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. Science (80-). 2014;345(6202):1369–72.

72. Park DJ, Dudas G, Wohl S, Goba A, Whitmer SLM, Andersen KG, et al. Ebola virus epidemiology, transmission, and evolution during seven months in Sierra Leone. Cell. 2015;161(7):1516–26.

73. Gonzalez-Reiche AS, Hernandez MM, Sullivan MJ, Ciferri B, Alshammary H, Obla A, et al. Introductions and early spread of SARS-CoV-2 in the New York City area. Science (80-) [Internet]. 2020 Jul 17;369(6501):297 LP – 301. Available from: http://science.sciencemag.org/content/369/6501/297.abstract

74. Villabona-Arenas CJ, Hanage WP, Tully DC. Phylogenetic interpretation during outbreaks requires caution. Nat Microbiol [Internet]. 2020;5(7):876–7. Available from: https://doi.org/10.1038/s41564-020-0738-5

75. Worby CJ, Lipsitch M, Hanage WP. Shared Genomic Variants: Identification of Transmission Routes Using Pathogen Deep-Sequence Data. Am J Epidemiol [Internet]. 2017 Nov 15;186(10):1209–16. Available from: http://dx.doi.org/10.1093/aje/kwx182

76. Martin MA, Lee RS, Cowley LA, Gardy JL, Hanage WP. Within-host Mycobacterium tuberculosis diversity and its utility for inferences of transmission. Microb genomics [Internet]. 2018/10/11. 2018 Oct;4(10):e000217. Available from: https://www.ncbi.nlm.nih.gov/pubmed/30303479

77. Lee RS, Proulx J-F, McIntosh F, Behr MA, Hanage WP. Previously undetected super-spreading of Mycobacterium tuberculosis revealed by deep sequencing. Elife. 2020;9:e53245.

78. Kahn R, Hitchings M, Bellan S, Lipsitch M. Impact of stochastically generated heterogeneity in hazard rates on individually randomized vaccine efficacy trials. Clin Trials. 2018;15(2):207–11.

79. Hitchings MDT, Lipsitch M, Wang R, Bellan SE. Competing Effects Of Indirect Protection And Clustering On The Power Of Cluster-Randomized Controlled Vaccine Trials. Am J Epidemiol. 2018 Mar 7;

80. Leavitt S V, Jenkins HE, Sebastiani P, Lee RS, Horsburgh Jr C, Tibbs A, et al. Estimation of the generation interval using pairwise relative transmission probabilities. Biostatistics. 2020;revise and resubmit.

81. Worby CJ, Read TD. 'SEEDY'(Simulation of Evolutionary and Epidemiological Dynamics): An R Package to Follow Accumulation of Within-Host Mutation in Pathogens. PLoS One. 2015;10(6):e0129745.

82. Worby CJ, Chang H-H, Hanage WP, Lipsitch M. The distribution of pairwise genetic distances: a tool for investigating disease transmission. Genetics. 2014;198(4):1395–404.

83. Field CA, Welsh AH. Bootstrapping clustered data. J R Stat Soc Ser B (Statistical

Methodol. 2007;69(3):369–90.

84. Kupferschmidt K. Mutations can reveal how the coronavirus moves—but they're easy to overinterpret. Science (80-). 2020;

85. Jary A, Leducq V, Malet I, Marot S, Klement-Frutos E, Teyssou E, et al. Evolution of viral quasispecies during SARS-CoV-2 infection. Clin Microbiol Infect [Internet]. 2020 Jul 24;S1198-743X(20)30440-7. Available from: https://pubmed.ncbi.nlm.nih.gov/32717416

86. Xu Y, Cancino-Muñoz I, Torres-Puente M, Villamayor LM, Borrás R, Borrás-Máñez M, et al. High-resolution mapping of tuberculosis transmission: Whole genome sequencing and phylogenetic modelling of a cohort from Valencia Region, Spain. PLoS Med [Internet]. 2019 Oct 31;16(10):e1002961–e1002961. Available from: https://www.ncbi.nlm.nih.gov/pubmed/31671150

87. Leavitt S V, Lee RS, Sebastiani P, Horsburgh Jr, CR, Jenkins HE, White LF. Estimating the relative probability of direct transmission between infectious disease patients. Int J Epidemiol [Internet]. 2020 Mar 24;49(3):764–75. Available from: https://doi.org/10.1093/ije/dyaa031

88. Lipsitch M. Who is Immune to the Coronavirus. New York Times. 2020;13.

89. Branswell H. CDC launches studies to get more precise count of undetected Covid-19 cases. STAT News.

90. Peeples L. News Feature: Avoiding pitfalls in the pursuit of a COVID-19 vaccine. Proc Natl Acad Sci. 2020;117(15):8218–21.

91. Metcalf CJE, Farrar J, Cutts FT, Basta NE, Graham AL, Lessler J, et al. Use of serological surveys to generate key insights into the changing global landscape of infectious disease. Lancet. 2016;388(10045):728–30.

92. Bryant JE, Azman AS, Ferrari MJ, Arnold BF, Boni MF, Boum Y, et al. Serology for SARS-CoV-2: Apprehensions, opportunities, and the path forward. Sci Immunol. 2020;5(47).

93. Organization WH. Correlates of vaccine-induced protection: methods and implications. World Health Organization; 2013.

94. Goldstein E, Pitzer VE, O'Hagan JJ, Lipsitch M. Temporally varying relative risks for infectious diseases: implications for infectious disease control. Epidemiology. 2017;28(1):136.

95. Koopman JS, Longini Jr IM. The ecological effects of individual exposures and nonlinear disease dynamics in populations. Am J Public Health. 1994;84(5):836–42.

96. Ray GT, Lewis N, Klein NP, Daley MF, Lipsitch M, Fireman B. Depletion-of-susceptibles Bias in Analyses of Intra-season Waning of Influenza Vaccine Effectiveness. Clin Infect Dis. 2020;70(7):1484–6.

97. Reed SE. The behaviour of recent isolates of human respiratory coronavirus in vitro and in

volunteers: Evidence of heterogeneity among 229E-related strains. J Med Virol. 1984;13(2):179–92.

98. Callow KA, Parry HF, Sergeant M, Tyrrell DAJ. The time course of the immune response to experimental coronavirus infection of man. Epidemiol Infect. 1990;105(2):435–46.

99. Kahn R. Serologic Studies. GitHub Repository.

100. Copeland KT, Checkoway H, McMichael AJ, Holbrook RH. Bias due to misclassification in the estimation of relative risk. Am J Epidemiol. 1977;105(5):488–95.

101. R igraph manual pages [Internet]. Available from: https://igraph.org/r/doc/sample_sbm.html

102. Hitchings MDT, Lipsitch M, Bellan S. Competing effects of indirect protection and clustering on the power of cluster-randomized controlled vaccine trials. bioRxiv [Internet]. 2017 Jan 1; Available from: https://www.biorxiv.org/content/early/2017/12/09/191163

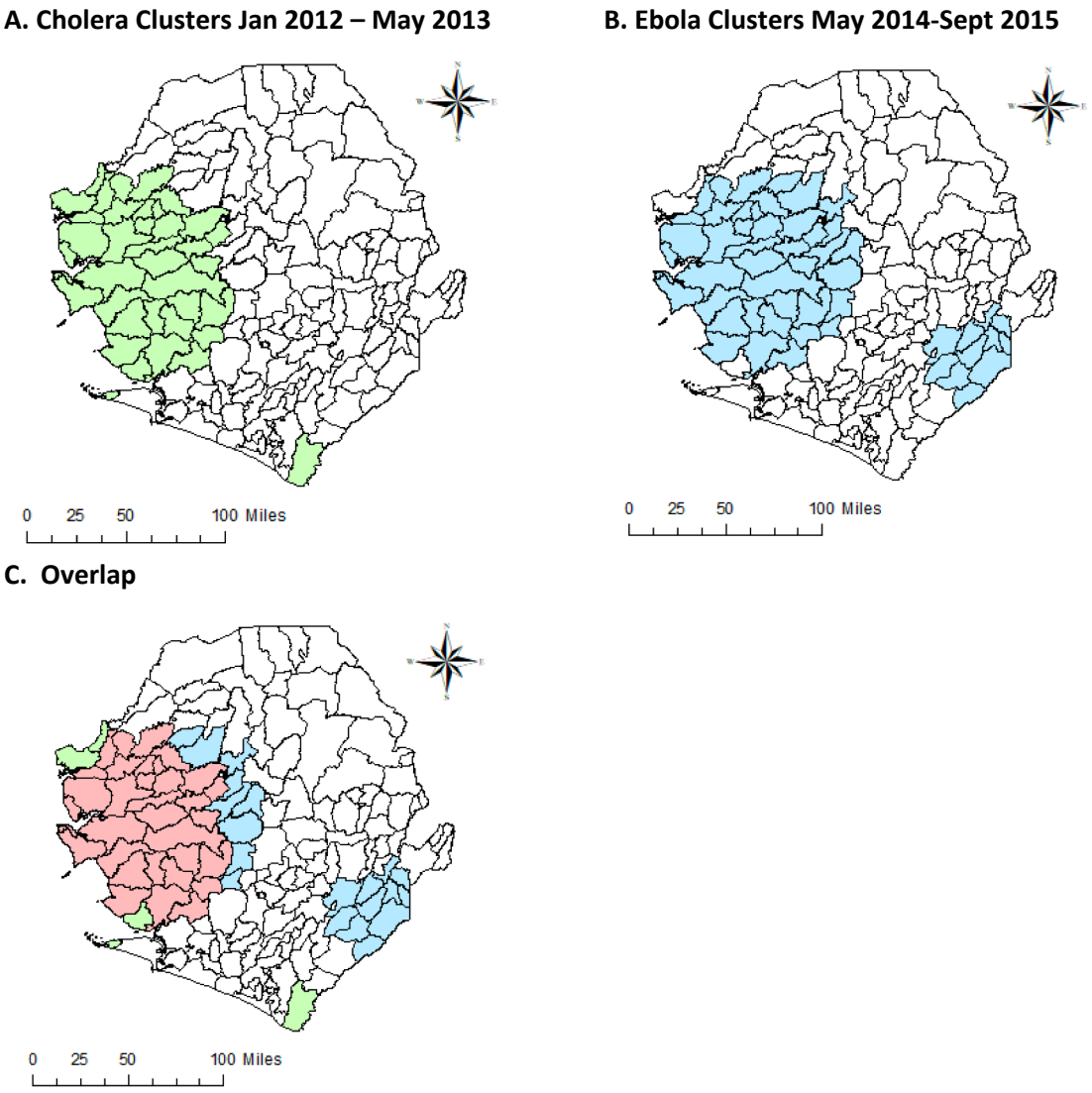
Appendix

Supplementary Movie 2.1. Spread of cholera and Ebola

See attached file.

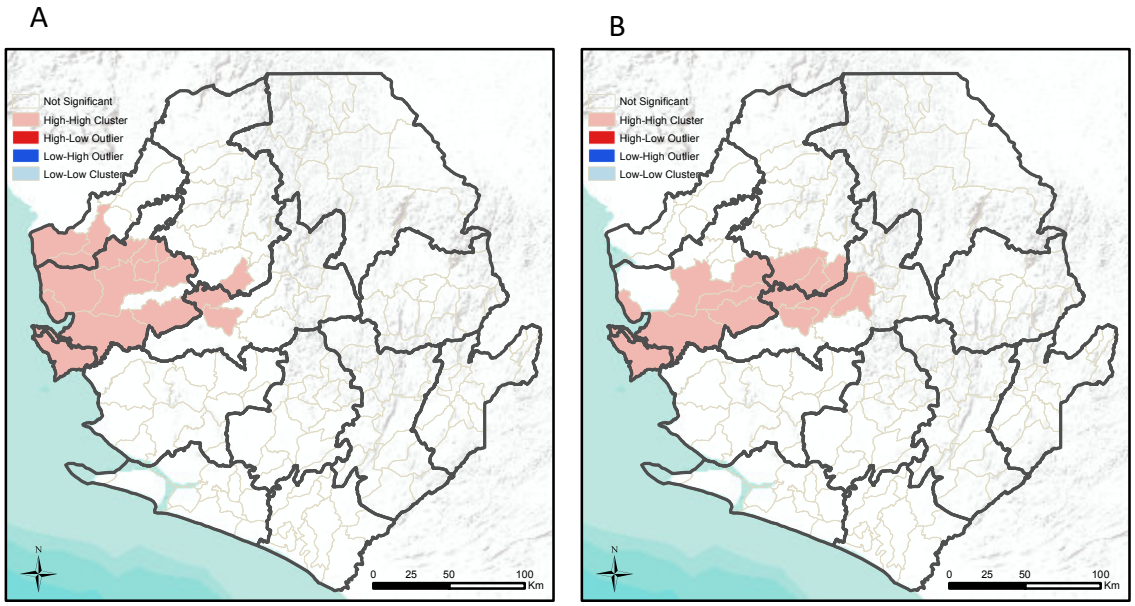
The spread of cholera (left panel) and Ebola (right panel) outbreaks across chiefdoms in Sierra Leone are shown in 14-day windows of aggregated cases (fill color). The border of each chiefdom is colored by the time since the first infection in that chiefdom (bright red indicates recent first infection). *Movie made by coauthor Juan Fernández Gracia*

Supplementary Figure 2.1. SatScan space-time analysis



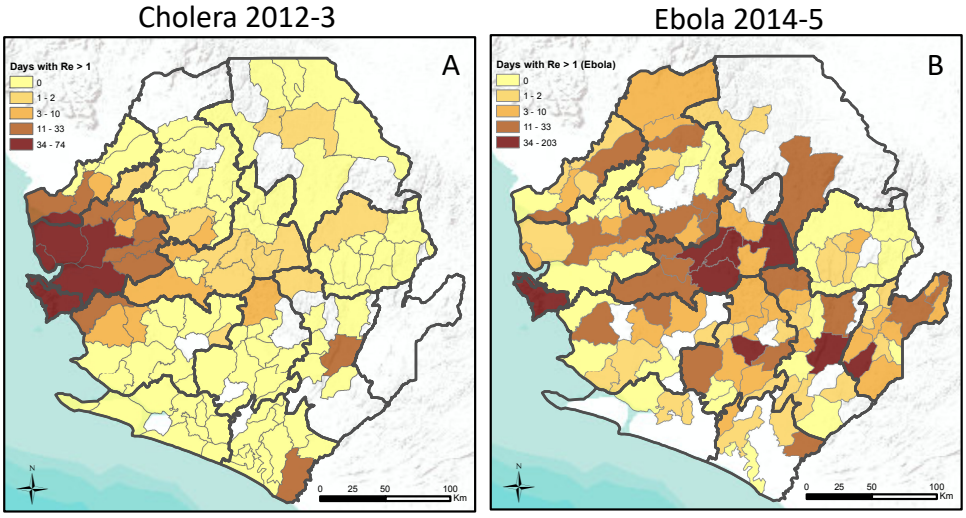
Results of SatScan space-time analysis for Cholera (A), Ebola (B) and overlap of space-time clusters (C)

Supplementary Figure 2.2. Local Moran's I attack rate



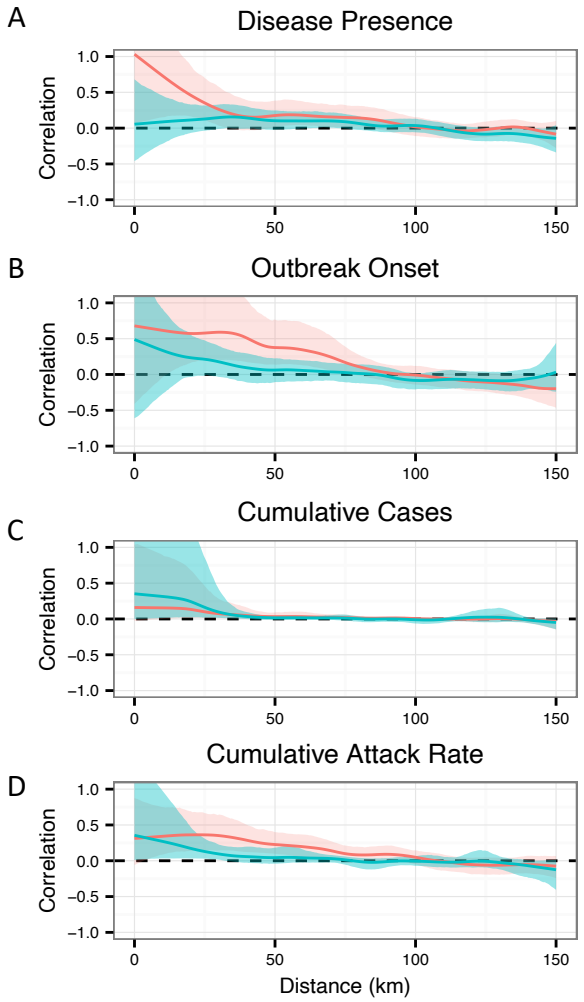
Local Moran's I Attack Rate Clustering for Cholera (A) and Ebola (B). Inverse distance squared neighborhood matrix used.

Supplementary Figure 2.3. R_t



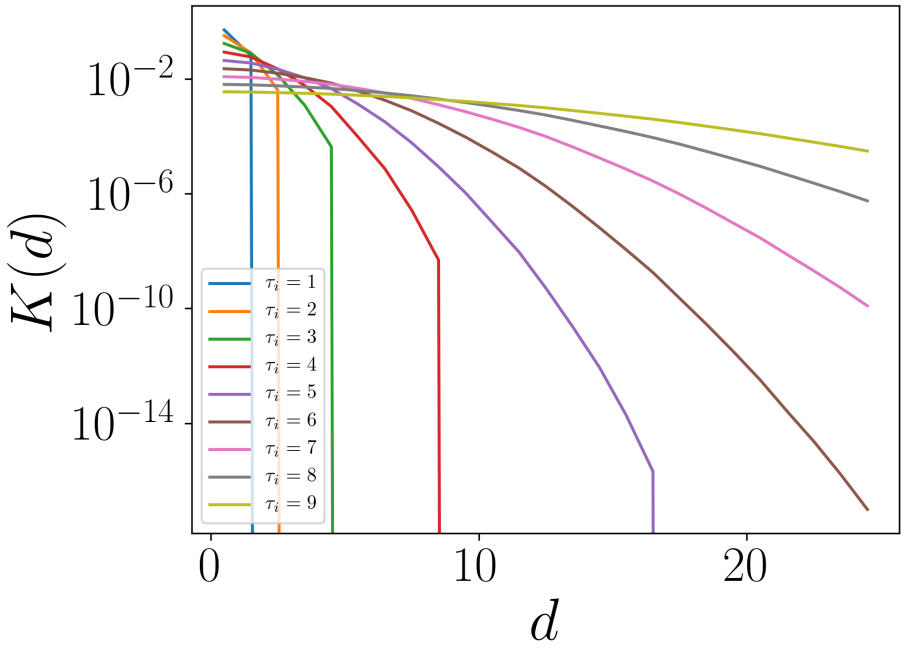
Estimated number of days with an effective reproductive number above unity for cholera (A) and Ebola (B) show darker regions sustaining more transmission. *Figure made by coauthor Corey Peak.*

Supplementary Figure 2.4. Correlation



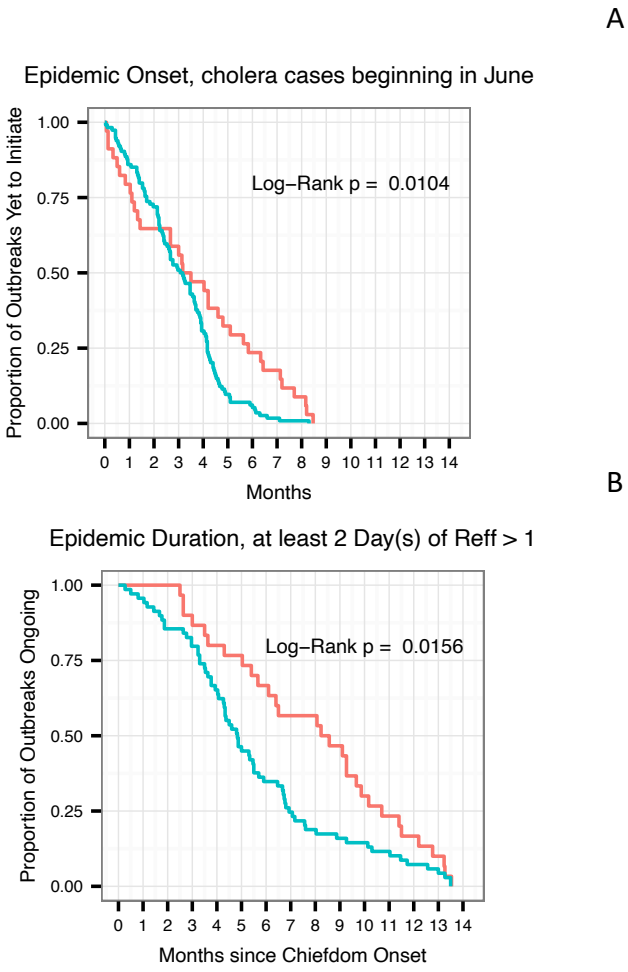
Spline correlograms showing tendency towards positive correlation between disease presence (A), outbreak onset date (B), cumulative cases (C), and cumulative attack rate (D) for both cholera (red) and Ebola (blue) as a function of distance between chiefdom pairs (x-axis). Shaded regions indicate 95% confidence intervals calculated for 1000 bootstrapped samples. *Figure made by coauthor Corey Peak.*

Supplementary Figure 2.5. Dispersion kernel



The dispersion kernel $K^x(d)$ is the probability that one an agent will end up at a position separated a distance d from the initial position after x days. For longer incubation periods (τ), the kernel is more homogeneously spread and has non-vanishing probabilities at greater distances, explaining the enhancement in sparking events for longer incubation periods. These simulations were conducted on a 50x50 lattice, with equal probability of left, right, up and down movement and 1/2 probability of not moving. *Figure made by coauthor Juan Fernández Gracia.*

Supplementary Figure 2.6. Survival curves



Survival curves for chiefdom outbreak onset (A) and outbreak duration (B) when considering a subset of cases of cholera (red) or Ebola (blue). Figure A excludes cholera cases before June. Figure B includes only chiefdoms with more than one day of R_t>1. *Figure made by coauthor Corey Peak.*

Supplementary Text 3.1.

Aim

In order to estimate VE_I, we need to estimate the **ratio** of the # people infected by a vaccinated person to the # infected by a control

$$VE_I = 1 - \frac{\#\text{ infected by vacc}/\#\text{ vacc}}{\#\text{ infected by control}/\#\text{ control}} / (1 - VE_S)$$

Because # vacc = # cont:

$$VE_I = 1 - \frac{\mathbf{\#\text{ infected by vacc}}}{\mathbf{\#\text{ infected by control}}} / (1 - VE_S)$$

Analysis

The ratio of the # infected by vacc / # infected by control comes from summing probabilities based on the vaccination status of potential infectors across all infectees.

Example for 1 infected person (probability could be obtained from any of the approaches described in the Methods)

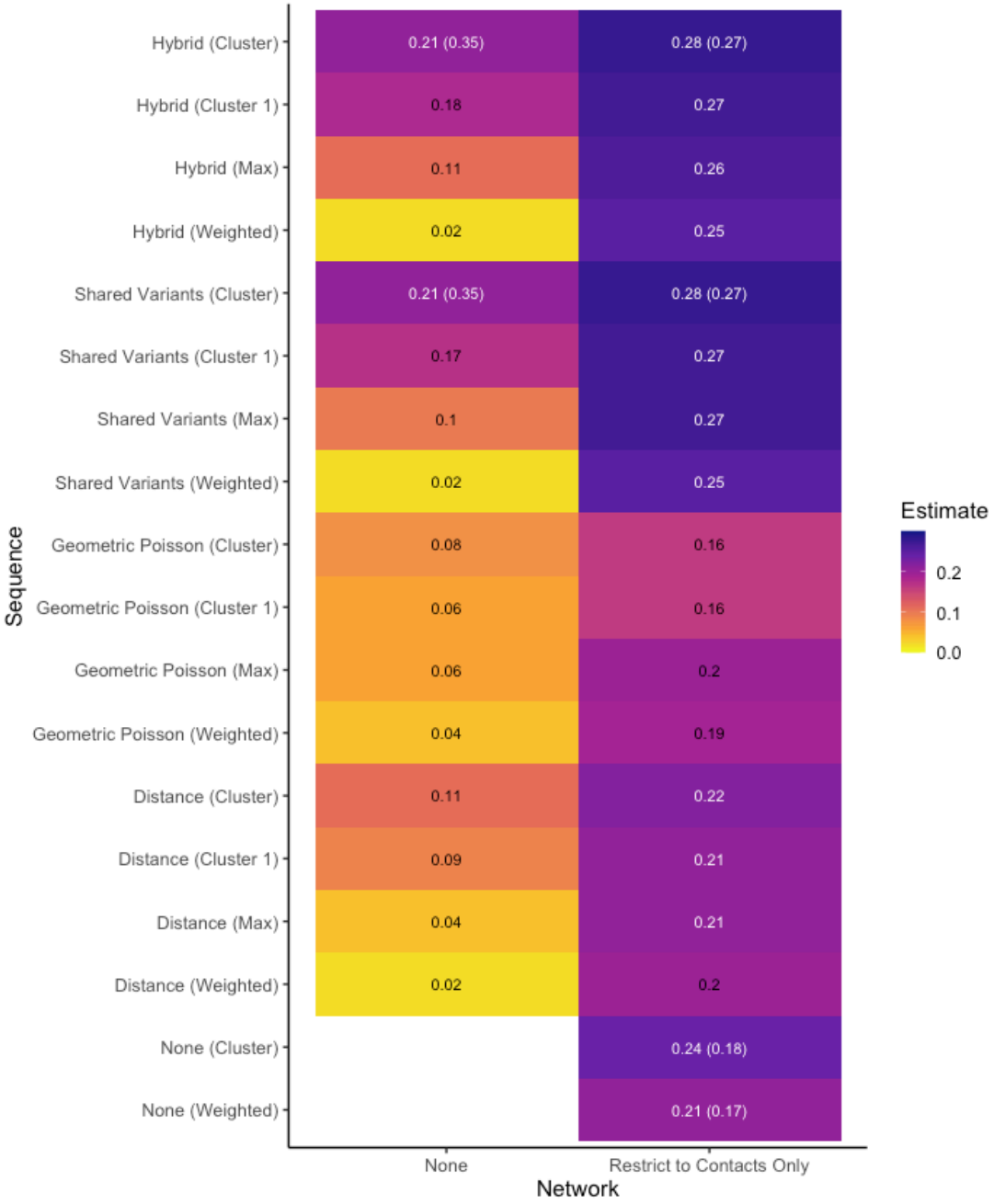
Potential Infector	Trial Status	Probability	Cluster
A	Vacc	0.05	2
B	Vacc	0.40	1
C	Control	0.35	1
D	Control	0.10	2
E	Control	0.10	2

All prob: include all potential infectors and count each as their probability. We add 0.45 [0.05+0.4] to the numerator of the ratio above and 0.55 [0.35 + 0.10 + 0.10] to the denominator.

Cluster: Divide infectors into clusters based on biggest gap in probability (here: 0.35-0.10 = 0.25). Include top cluster if gap is greater than or equal to the threshold (in simulations above, threshold=0.2 so 0.25 meets this criteria). We add 0.40/(0.40+0.35) to the numerator of the ratio above and 0.35/(0.40+0.35) to the denominator because only 0.40 and 0.35 fall into the top cluster.

Max: Include only potential infector(s) that have the highest probability and count as 1. Here, we would add 1 to the numerator of the ratio and nothing to the denominator because 0.4, the maximum probability, is a vaccinated person. If two are tied for most likely then each counts ½.

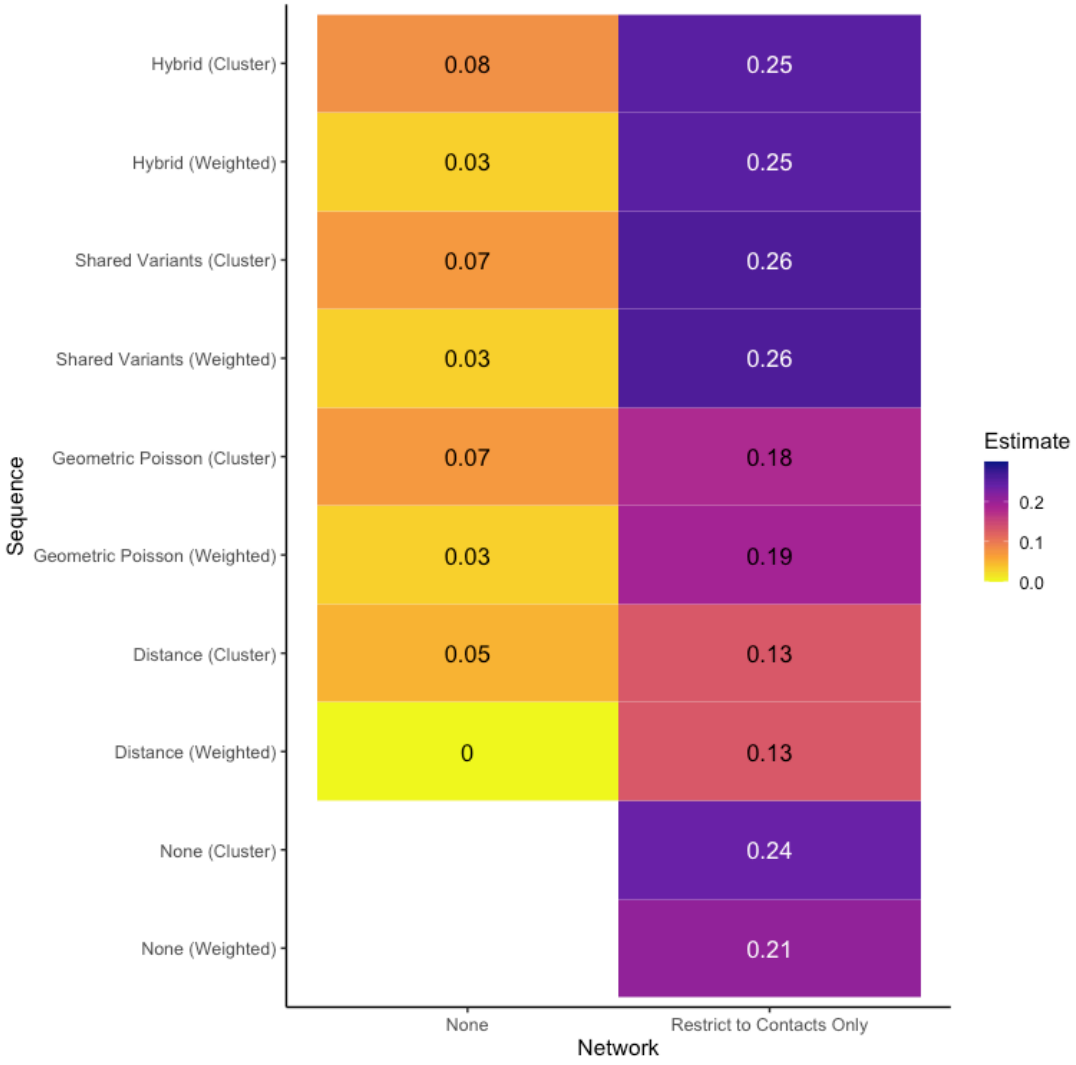
Supplementary Figure 3.1. All methods



*Cluster: 0.2 threshold; Cluster 1: 0.1 threshold used

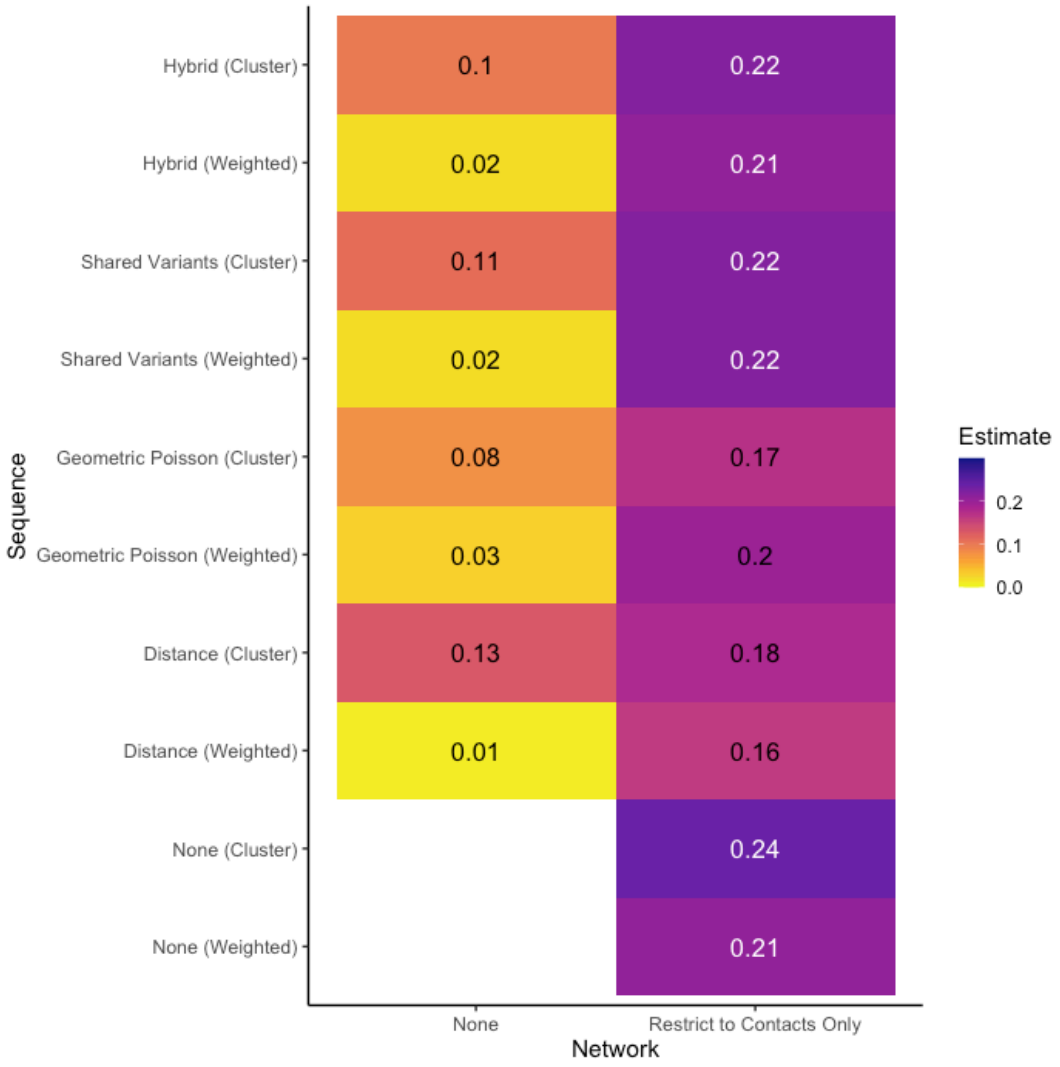
The median VE_I estimates from 500 simulations with the baseline parameters, with a true VE_I of 0.3. All approaches, including multiple clustering thresholds and the approach using the vaccination status of only the most likely infector(s) ("Max"), are shown, as well as the average standard errors for the best performing approaches.

Supplementary Figure 3.2. Mutation rate = 0.003



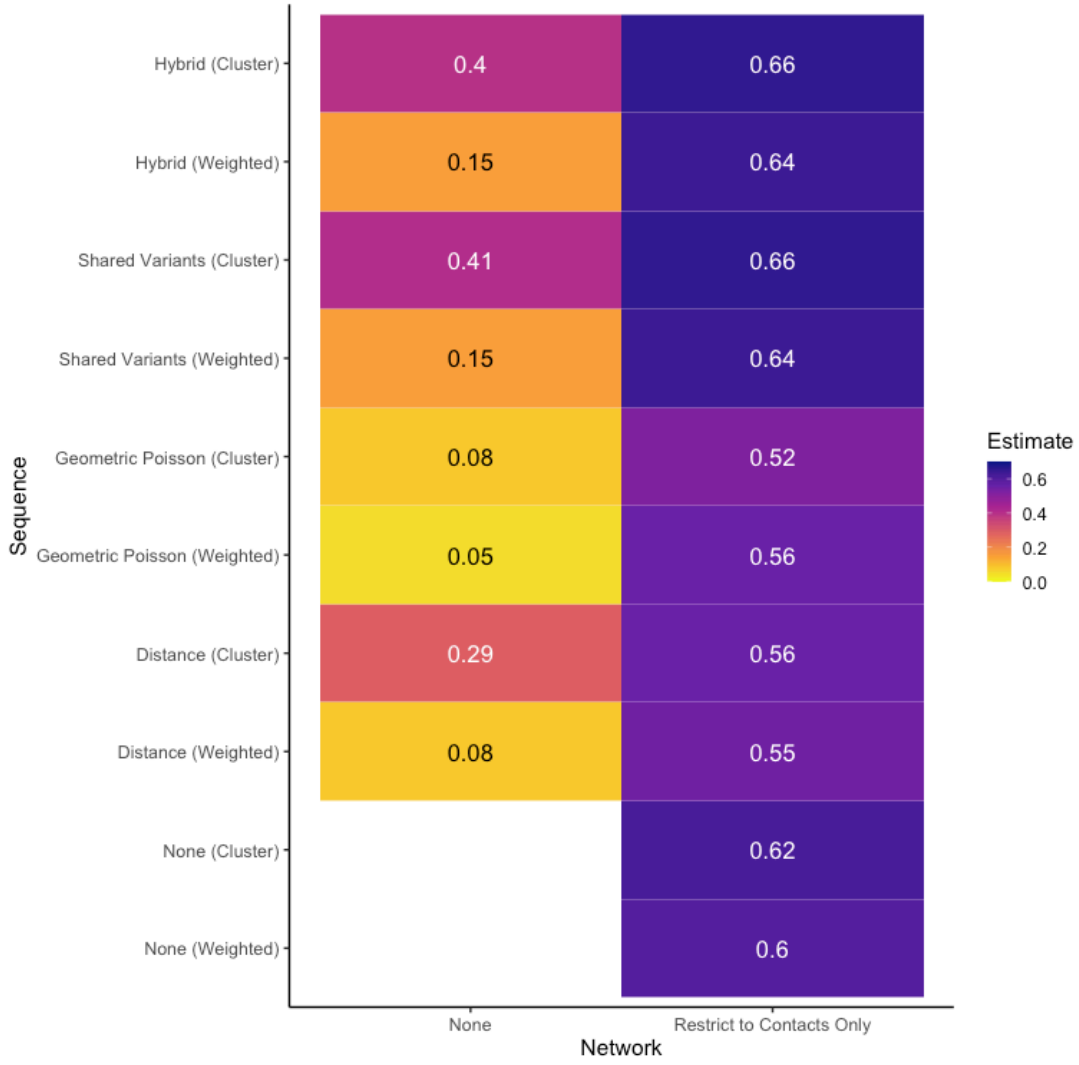
The median VE_I estimates from 500 simulations with a lower mutation rate of 0.003, with a true VE_I of 0.3.

Supplementary Figure 3.3. Bottleneck = 2



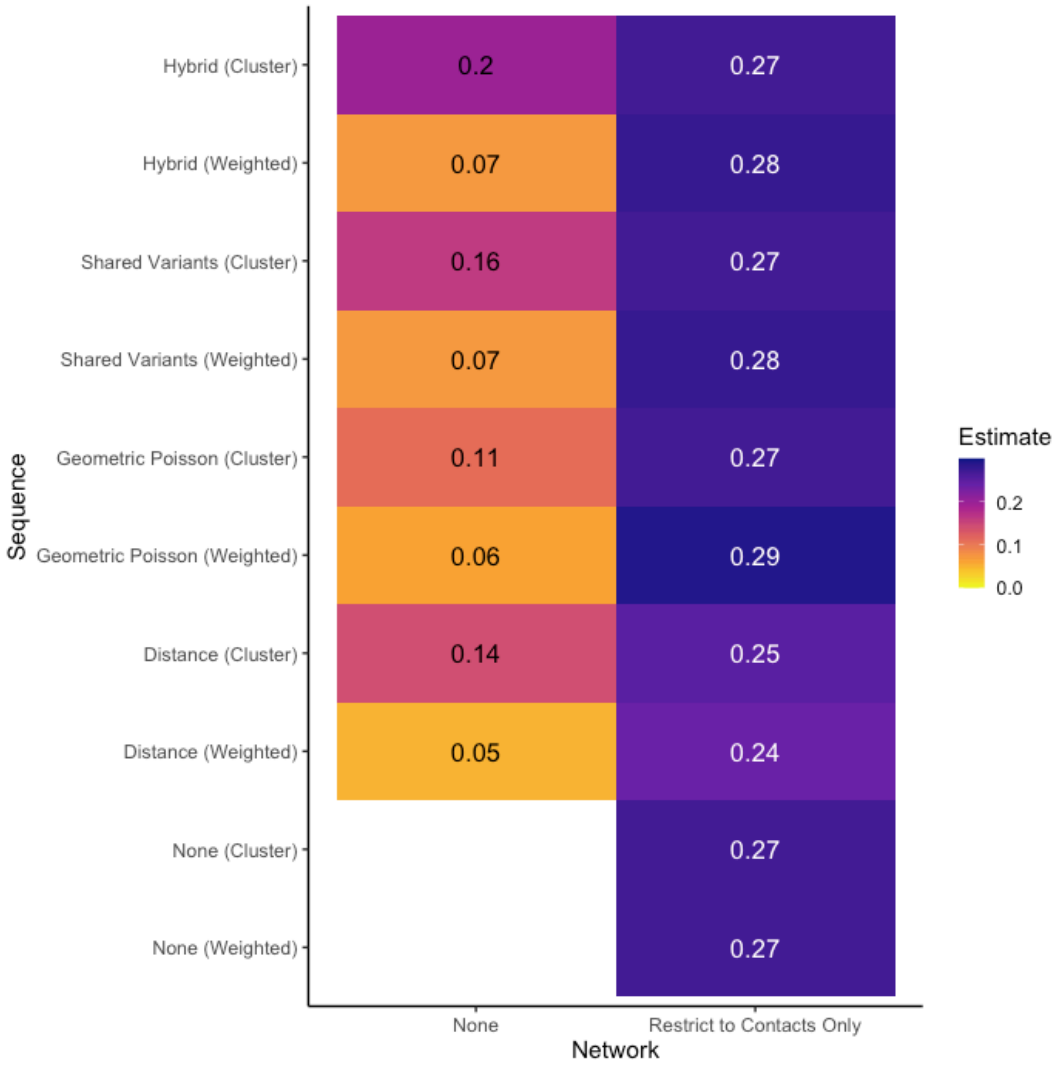
The median VE_I estimates from 500 simulations with a lower bottleneck size of 2, with a true VE_I of 0.3.

Supplementary Figure 3.4. $VE_I = 0.7$



The median VE_I estimates from 500 simulations, with a true VE_I of 0.7.

Supplementary Figure 3.5. $VE_S = 0.8$



The median VE_I estimates from 500 simulations with a VE_S of 0.8, with a true VE_I of 0.3.

Supplementary Text 4.1. Data generating details – network and outbreak

Generate a network

We use a stochastic block model to generate a network graph, using the sample_sbm function in the R package igraph.[101] In our simulations, we create networks with one single community and networks with 10 communities. We keep the total population across simulations constant at 10,000. Therefore in simulations with 1 community, there are 10,000 nodes in that one community, and for simulations with 10 communities, there are 1,000 people in each community.

The sample_sbm function conducts a Bernoulli trial for each potential edge in the graph. In our "well mixed" communities, the probability for each edge within the same community is 1, meaning all nodes in the community are connected. The probability of connection for nodes in different communities is 0, meaning there are no edges between communities. For "clustered" communities, the probability of an edge between nodes in different communities remains 0. While in the well mixed communities, the probability of an edge between nodes in the same community was 1, here it is greatly reduced. For the single community simulations, the probability of an edge is 0.002, and for the 10 community simulations, the probability of a within-community edge is 0.02, meaning the expected number of edges for each node is approximately 20. This creates smaller, overlapping communities, or clusters, within each larger discrete community.

In order to keep R_E constant in all simulations, we use the following formula[14,78] to calculate β (the force of infection):

$$R_E = T * (\frac{\langle k^2 \rangle}{k} - 1)$$

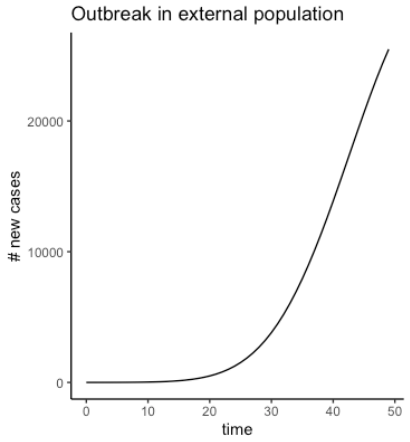
$$T = 1 - (\frac{\gamma}{\gamma+\beta})^{\alpha}$$

where k is the mean degree of the network, k^2 is the mean of the distribution of the square of the number of connections an individual has in the network, γ is the infectious period rate, α is the infectious period shape (see Table 1).

Seeding outbreak

The outbreak in the communities is seeded by introductions from an outbreak in an external population[102] of one million individuals. All introductions occur between day 1 and 50. The number of nodes infected externally on a given day is based on a binomial distribution, with the probability equal to $1 - e^{-F_i * I}$ where F_i is the proportionality constant for the amount of contact between the external population and a node in community *i*, and I is the number of infected individuals in the external population, which has an exponentially growing deterministic outbreak from day 1 to day 50. $F_i = \frac{-\ln(1-\frac{(\varphi)}{\sqrt{C_i}*\sum_i \sqrt{C_i}})}{\int_1^{50} I_t}$ where φ = expected number of introductions, C_i = size of community i and I_t is the number of infected individuals in the external population on day t. The probability of introduction for a given node scales with the size of that node's community; however in our simulations, all communities have the same size so the probability is equal.

Supplementary Figure 4.1. Outbreak in external population



Outbreak in communities

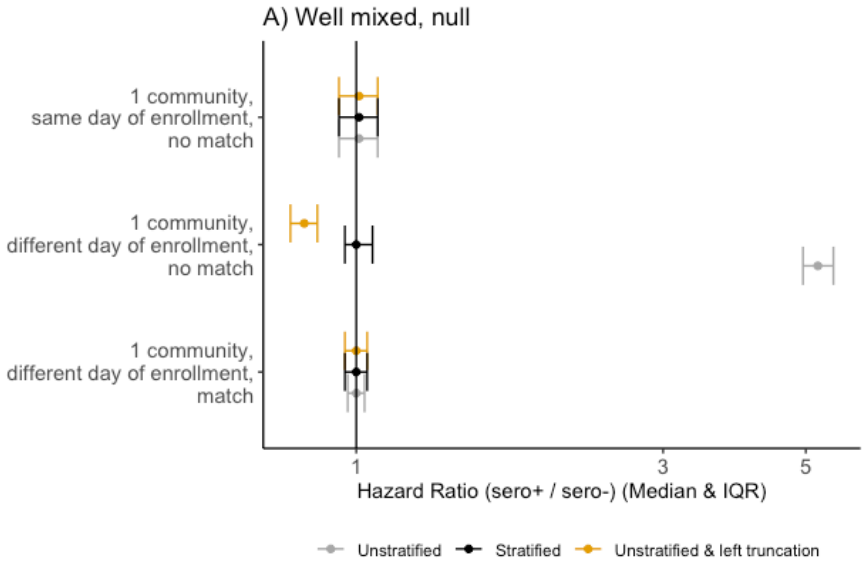
When a node is infected, either from the external population or from an infected node in the community, they move from the susceptible compartment to the exposed compartment. Their latent period, the time between exposure and onset of infectiousness, is drawn from a gamma distribution with mean 5.6 days, independent of the period for any other node. After the latent period ends, individuals progress to the infectious compartment. Their infectious period is drawn from a gamma distribution with a mean of 10 days, independent of the period for any other node. After their infectious period ends, they move into the susceptible' (S') compartment.

On each day, an infectious node has a daily probability of infecting each of the susceptible (seronegative) nodes they are connected to of $1 - e^{-\beta^-}$ where β^- is the force of infection for those initially susceptible, and a daily probability of infecting each of the susceptible' (seropositive) nodes they are connected to of $1 - e^{-\beta^+}$ where β^+ is the force of infection for those who have been infected previously. The nodes they infect then move into the exposed compartment and the steps above repeat.

For an uninfected node, the probability of infection on any given day is equal to $1-(e^{-\beta^+})^n$if the node is seropositive and $1-(e^{-\beta^-})^n$ if the node is seronegative, where *n* is the number of infectious contacts of that uninfected node. For small *n*, these probabilities are approximately equal to $n\beta^+$ and $n\beta^-$, respectively, so the ratio of these probabilities (i.e., the hazard ratio due to seroprotection) is approximately equal to the ratio β^+/β^-. When *n* is not small, however, this simplification no longer holds. Since in our data, we only record total daily new infections (in seropositives and seronegatives), Cox model software [that uses (conditional) logistic regression to deal with tied failure times] estimates the parameter $[1-(e^{-\beta^+})^n]/[1-(e^{-\beta^-})^n]$, which is closer to the null than the instantaneous (i.e. continuous time) hazard ratio.β^+/β^-. If we had recorded the number of new infections occurring hourly rather than daily, the same Cox model software would have again outputted an estimate approximately equal to β^+/β^- as the number of hourly contacts, n_hour, is 1/16 of the daily contacts (assuming 8 hours sleep without contact).

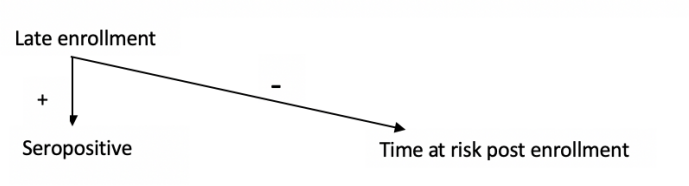
Supplementary Text 4.2. Left truncation

Supplementary Figure 4.2. Left truncation results



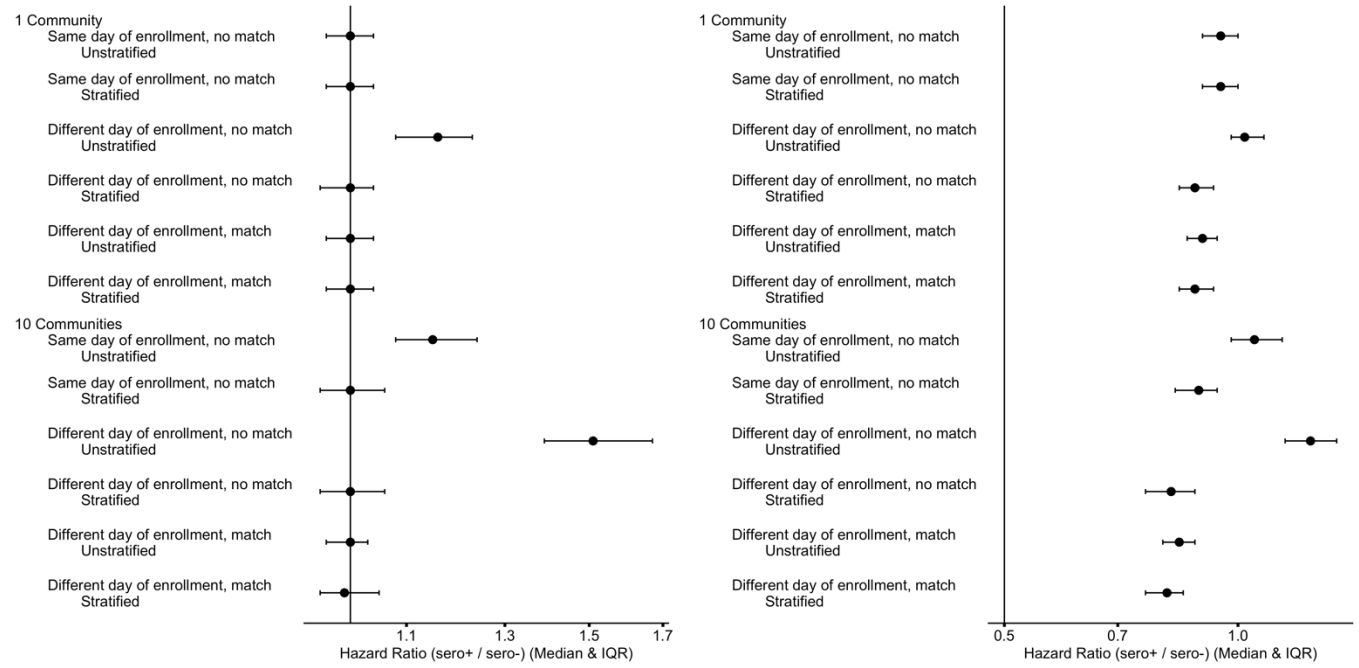
We compare the median and IQR of the estimated hazard ratios, comparing seropositives to seronegatives, for simulations in one well mixed community in an uncontrolled epidemic with R_E=2. We consider three sampling designs: enrolling individuals on a single day without matching, enrolling individuals on multiple days without matching, and enrolling individuals on multiple days with matching. In the matched designs, for each seropositive individual enrolled on each enrollment day, a seronegative individual from the same community is also enrolled on that day. We compare analyses stratified by enrollment day and community (black) to both unstratified analyses as described above (grey) and to unstratified analyses accounting for left truncation by enrollment time (orange). When individuals are enrolled in an unmatched design on different days, the left truncated analysis, which uses calendar time instead of time since enrollment, is biased down in a null setting with no effect of seropositivity. This occurs because the distribution of seropositives and seronegatives enrolled is not constant across days of enrollment. First, the proportion of seropositives enrolled on day 150 is greater than on day 50. Additionally, those enrolled at day 150 from the S and S' compartments cannot move into compartment I on day 151 as they must go through E first (which has an expected duration greater than 1 day), while a fraction of those enrolled on day 50 can be in the exposed state on day 150 already and thus can become infected on day 151. Thus a higher percentage of seronegatives are infected on day 151 compared to seropositives since these seronegatives are overrepresented in those enrolled on day 50. See below for a directed acyclic graph representing this bias. Matching removes this imbalance as equal numbers of seropositives and seronegatives are enrolled on each day. In settings with lower R_E, this bias from left truncation is imperceptible. Overall, there are settings where stratified analyses are unbiased and unstratified analyses accounting for left truncation retain bias, so the former analysis approach is preferable.

Supplementary Figure 4.3. Left truncation directed acyclic graph



Supplementary Text 4.3. 90% specificity

Supplementary Figure 4.4. 90% specificity results



The median and IQR of estimated hazard ratios, comparing seropositives to seronegatives, with 90% specificity for simulations with well mixed communities with an uncontrolled epidemic under settings of no seroprotection (A) and 50% seroprotection (B). We consider three sampling designs for each simulation setting: enrolling individuals on a single day without matching, enrolling individuals on multiple days without matching, and enrolling individuals on multiple days with matching. In the matched designs, for each seropositive individual enrolled on each enrollment day, a seronegative individual from the same community is also enrolled on that day. We compare analyses stratified by enrollment day and community (black) to unstratified analyses (grey). As expected in settings with seroprotection, imperfect specificity biases results towards the null.