MODELING HOUSEHOLD EPIDEMICS AND ANALYSIS OF LINKS FROM TESTING TO HOSPITALIZATIONS

by

Theresa R. Sheets

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Mathematics
The University of Utah
August 2023

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of **Theresa R. Sheets**
has been approved by the following supervisory committee members:

Frederick R. Adler ,	Chair(s)	**July 23, 2023** Date Approved
Paul C. Bressloff ,	Member	**July 21, 2023** Date Approved
Lindsay T. Keegan ,	Member	**July 21, 2023** Date Approved
Jody Renae Reimer ,	Member	**July 26, 2023** Date Approved
Damon J. A. Toth ,	Member	**July 24, 2023** Date Approved

by **Tommaso de Fernex** , Chair of the Department of **Mathematics**
and by **Darryl P. Butt** , Dean of The Graduate School.

Abstract

I develop several mathematical frameworks for investigating infectious disease transmission dynamics. In the first chapter, I develop and test statewide forecasts of weekly COVID-19 hospitalizations at the state level by incorporating public health data streams: testing, wastewater, and syndromic surveillance. The forecasts are measured against and surpass a naive benchmark and ARIMA null model. The second chapter takes a more granular view and includes a simulation model of SARS-CoV-2 transmission with the aim of quantifying the impacts of vaccination on secondary transmission within households. We find that given an importation event, vaccination of children does decrease total household infection, yet has little impact on secondary transmissions to adults. In the third chapter, I develop an analytic solution to model the spread of an infectious disease within a small population and test it against the results from the simulation offering a dramatic improvement in efficiency as compared to simulation-based methods. Together with these chapters, I aim to provide a generalizable decision-support framework to inform effective interventions for future infectious disease concerns.

For all the help along the way

Contents

List of Figures

List of Tables

List of Notation and Symbols

β	Contact rate between individuals
γ	Recovery rate
Γ	Gamma function
σ	Susceptibility and Vaccine efficacy
r	Rate parameter of the Gamma distribution
k	Shape parameter of the Gamma distribution, dispersion
AIC	Akaike's Information Criterion
ARIMA	Auto Regressive Integrated Moving Average
CDC	Centers for Disease Control
ER	Emergency Room
GAM	Generalized Additive Model
ICD-10	International Classification of Diseases Revision 10
LHS	Latin Hypercube Sampling
MAE	Mean Absolute Error
MLE	Maximum Likelihood Estimation
MSE	Mean Squared Error
NAAT	Nucleic Acid Amplification Test
P/P	Person-over-person
PCR	Polymerase Chain Reaction
PDF	Probability distribution function
PMF	Probability mass function
RSV	Respiratory syncytial virus
SAR	Secondary Attack Rate
T/T	Test-over-test
UDHHS	Utah Department of Health and Human Services

Acknowledgements

This dissertation would not have been possible without the support and encouragement of my parents, friends, supervisors, and peers.

I have been given the incredible opportunity to learn, work, and develop as a researcher in the Mathematical Biology Group at the University of Utah, surrounded and supported by an engaging and enthusiastic group of students, mentors, and researchers. I would like to express my gratitude to my advisor, Fred Adler, for his patience, the valuable perspectives shared during our meetings, and the multitude of insights whose significance only became apparent to me much later. You gave me the freedom to pursue the projects about which I am most passionate. I would also like to thank my committee: Lindsay Keegan, Damon Toth, Jodie Reimer, and Paul Bressloff. Lindsay, you have been the most incredible mentor; working with you has given me the fantastic opportunity to work to bridge the divide between math and epidemiology. Our research collaboration has opened doors for me and allowed me to pursue the questions I am most passionate about. Damon, thank you for answering all of my questions about math and patience while I reread the papers you suggested. Thank you, Jodie, for your invaluable research suggestions; your suggestions about how to structure my code enabled me to communicate what I saw in my results. Paul for reminding me to ensure my projects were grounded in math. Besides my committee, there have been many enthusiastically supportive people around campus. Tracey Lamb warmly welcomed me into her lab. In particular, thanks to the team from the Division of Epidemiology; Karim Khader, Matthew Samore, Jay Love, and Sharia Ahmed.

My experiences at the University of Maryland Baltimore County and with the Meyerhoff program gave me the foundations for graduate school. Thank you for taking a chance on me all those years ago and giving me the road map

here. In particular, Bonny Tighe, Kal Nanes, Kathleen Hoffman, and Justin Brooks. I would also like to thank the National Institute for Mathematical and Biological Synthesis for providing me with a summer research opportunity as an undergraduate and giving me a taste of math modeling research. I am incredibly grateful for Suzanne Lenhart's suggestion to apply to the U.

As a graduate student, in my first years I was supported by a Reseach Traning Grant from the National Science Foundation and I have been supported since 2020 by the Safety and Healthcare Epidemiology Prevention Research Development Grant from the Centers for Disease Control. I am honored that the Division of Epidemiology trusted me to work on these projects and support this research. This grant opened the door to my internship with the Utah Department of Health and Human Services, where I have received immense support, a fantastic perspective on the workings of Public Health, and the data to complete the last chapter of this dissertation. Thank you to Sam Lefevre, Joel Skaria, Abigail Collingwood, Randon Gruninger, and Nathan LaCross for your support and willingness to answer my questions about your work.

I want to thank my family, both of origin and chosen. My parents who have made so many sacrifices to contribute to my education. My dad is always excited to pick up the phone. My mom who wants to visit as often as possible. My brother Cole for taking care of our parents while I have been across the country. Hanna and Mac, I could not ask for better climbing and skiing partners or friends. Chris, Abe, Yasser, Jamshaid, I am so grateful to have you in my life. Wes, thank you for supporting me these past few months.

Chapter 1

Introduction

To support decision-making in emergent situations, we turn to imprecise metrics and make decisions on the fly, supported by what little evidence is available and knowledge gleaned from past experiences. Coordinating decision support, orchestrating a response, and advocating for action are monumental challenges in and of themselves while quantifying the validity of approaches, data streams, and developing scaleable methods often fall by the wayside until the crisis has passed.

In this dissertation, I present three chapters through which I aim to contribute to the body of work on modeling and assessing infectious disease transmission at various scales. The first takes a broader perspective and approaches statewide forecasting models of COVID-19 hospitalizations, incorporating public health data from testing, wastewater, and syndromic surveillance. In the second chapter, I use a simulation model to explore the vaccination of differing household members. The third chapter builds an extension to the second and proposes an analytic solution to the spread of an infectious disease in small epidemics, treating each individual independently.

In the first chapter, I develop a generalizable framework for ongoing forecasts of hospitalizations across the state. We experiment with which available data streams minimize forecast error and provide the structure for a forecasting approach to be utilized as testing, immunity, and behavioral dynamics change. The results also suggest avenues for extensions to Influenza and RSV, which can be monitored via syndromic and wastewater surveillance.

The second explores the simulation-based methods for studying household epidemics and tests multiple vaccination scenarios. Here we explore the effects

of vaccinating different household members on overall transmission within the household. This chapter also provides a sensitivity analysis for some the parameters incorporated in the following chapter. The third chapter proposes an alternative approach to small epidemics that offers a decisive advantage in computational efficiency compared to simulation-based methods. The analytic solution to household epidemics can be readily applied to test vaccine effects and quantify the on-the-ground impacts of various prevention methods. It extends theoretical work and is used to calculate final size equations for epidemics. It produces results verifiable by simulation while allowing for a much more comprehensive exploration of the probability space than simulation would allow.

These chapters aim to quantify "what if" questions about the spread of infectious disease and ultimately provide estimates and forecasts to inform decision-making and change the future we have predicted. What will happen if our vaccines are fifty percent effective? Who should be vaccinated, and what happens when they are? What will happen if individuals are more or less able to transmit? What if we include wastewater in our forecasts? Through these chapters, I aim to improve our approaches to questions of tantamount importance to the health of the people who live with and around us.

Mathematical models support decision-making and validate hypotheses in a myriad of fields; computational chemistry and pharmaceutical development,[2] econometrics,[3] biology,[17] public policy,[14] and many more. Many suggest that decision-making research is futile.[6,19] Others say that it has excellent utility[9] and propose frameworks for careful assessment.[16] Consensus is that if they are to have any utility at all, they must be carefully considered, developed, constrained, and contextuallized.[4,11] There are many formulations that models can take. Here I consider models of the purely statistical and compartmental varieties. The compartmental model is separated into a stochastic simulation presented in chapter three and an analytic approach to the deterministic formulation of the same problem presented in chapter 3.

In the late 17th and early 18th century, Jacob Bernoulli developed the law of large numbers and what would come to be known as the binomial distribution.[18] The classic example of the binomial distribution is a repeated coin flip.[15] This

model has been used and extended in many contexts. A hundred years after that, variations on chain binomial models were proposed by Lowell Reed and Wade Hampton Frost[1] in 1930 and Greenwood[8] in 1931 for modeling sequences of infectious disease transmission through a small population. These were then generalized by Becker,[5] and used widely[12] for estimating household transmission rates. These models often incorporate maximum likelihood estimation (**MLE**) to estimate transmission parameters and final outbreak size.

One of the most widely recognized compartmental models is the *susceptible-infected-recovered* (**SIR**), which Kermack and McKendric first proposed at about the same time as Greenwood's chain binomial in 1927.[10] The SIR uses a system of deterministic differential equations to model the spread of infectious disease through a population. Fitting the SIR to epidemic data can allow for estimates of the reproductive number of the epidemic, R_0, the average number of cases from one infected individual in a fully susceptible population. The SIR can also be used to predict total outbreak size, maximum infected at one time, and the peak of the epidemic,[13] parameterized by known or estimated parameters.

Mathematical models offer the opportunity to experiment with and test on systems which would be impossible to fully explore experimentally. There are many challenges with accurately constructing, parameterizing, and validating compartmental models and these models grow exponentially with any complexities added to the system.[7,11] The assumptions inherent in the SIR model are easy to rationalize, but quickly fall apart as waning immunity, vaccination, and population heterogeneity are observed. Here statistical models bridge the gap between the fine tuning necessary for compartmental models and the urgent needs of public health officials and hospital administrators.

The emergence of COVID-19 has put mathematical models of epidemic disease at the forefronts of the minds of not only mathematicians and modelers, but the rest of the world. In this dissertation, COVID-19 data are used to inform models aimed at multiple pressing questions about forecasting during, vaccination strategy, and transmission during small epidemics. Further, these models provide flexible frameworks which can be carried forward to future outbreaks and inform policy as we prepare for and eventually face new epidemic threats.

1.1 References

[1] H. Abbey, *An examination of the Reed-Frost theory of epidemics*, Hum Biol, 24 (1952), pp. 201–233.

[2] Y. A. Abramov, G. Sun, and Q. Zeng, *Emerging Landscape of Computational Modeling in Pharmaceutical Development*, Journal of Chemical Information and Modeling, 62 (2022), pp. 1160–1171. Publisher: American Chemical Society.

[3] B. H. Baltagi, *The Mathematical Aspects of Econometrics*, in Mathematics Unlimited - 2001 and Beyond, B. Engquist and W. Schmid, eds., Springer, Berlin, Heidelberg, 2001, pp. 67–81.

[4] C. T. Bauch, J. O. Lloyd-Smith, M. P. Coffee, and A. P. Galvani, *Dynamically Modeling SARS and Other Newly Emerging Respiratory Illnesses: Past, Present, and Future*, Epidemiology, 16 (2005), pp. 791–801. Publisher: Lippincott Williams & Wilkins.

[5] N. Becker and I. Marschner, *The effect of heterogeneity on the spread of disease*, in Stochastic Processes in Epidemic Theory, J.-P. Gabriel, C. Lefevre, and P. Picard, eds., Lecture Notes in Biomathematics, Berlin, Heidelberg, 1990, Springer, pp. 90–103.

[6] E. Y. Cramer, E. L. Ray, V. K. Lopez, J. Bracher, A. Brennen, A. J. Castro Rivadeneira, A. Gerding, T. Gneiting, K. H. House, Y. Huang, R. J. Jayawardena, R. Tibshirani, V. Ventura, L. Wasserman, E. B. ODea, J. M. Drake, R. Pagano, Q. T. Tran, L. S. T. Ho, H. Huynh, J. W. Walker, R. B. Slayton, M. A. Johansson, M. Biggerstaff, and N. G. Reich, *Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States*, Proceedings of the National Academy of Sciences, 119 (2022), p. e2113561119. Publisher: Proceedings of the National Academy of Sciences.

[7] J. R. Gog, L. Pellis, J. L. N. Wood, A. R. McLean, N. Arinaminpathy, and J. O. Lloyd-Smith, *Seven challenges in modeling pathogen dynamics within-host and across scales*, Epidemics, 10 (2015), pp. 45–48.

[8] M. Greenwood, *On the Statistical Measure of Infectiousness*, J Hyg Lond, 31 (1931), pp. 336–351.

[9] S.-M. Huang, D. R. Abernethy, Y. Wang, P. Zhao, and I. Zineh, *The utility of modeling and simulation in drug development and regulatory review*, Journal of Pharmaceutical Sciences, 102 (2013), pp. 2912–2923.

[10] W. O. Kermack and A. G. McKendrick, *A Contribution to the Mathematical Theory of Epidemics*, Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character, 115 (1927), pp. 700–721. Publisher: The Royal Society.

[11] M. Kretzschmar, *Disease modeling for public health: added value, challenges, and institutional constraints*, Journal of Public Health Policy, 41 (2020), pp. 39–51.

[12] I. M. Longini, J. S. Koopman, A. S. Monto, and J. P. Fox, *Estimating household and community transmission parameters for influenza*, Am J Epidemiol, 115 (1982), pp. 736–751.

[13] J. L. Mahasmara, Respatiwulan, and Y. Susanti, *Final size distribution of stochastic susceptible infected recovered SIR epidemic model*, in The third international conference on Mathematics: Education, Theory and Application, Surakarta, Indonesia, 2021, p. 020010.

[14] J. Panovska-Griffiths, C. Kerr, W. Waites, and R. Stuart, *Mathematical modeling as a tool for policy decision making: Applications to the COVID-19 pandemic*, Handbook of Statistics, 44 (2021), pp. 291–326.

[15] S. M. Ross, *A first course in probability*, Pearson Prentice Hall, Upper Saddle River, N.J, 8th ed ed., 2010. OCLC: ocn237199460.

[16] J. Thompson, R. McClure, N. Scott, M. Hellard, R. Abeysuriya, R. Vidanaarachchi, J. Thwaites, J. V. Lazarus, J. Lavis, S. Michie, C. Bullen, M. Prokopenko, S. L. Chang, O. M. Cliff, C. Zachreson, A. Blakely, T. Wilson, D. A. Ouakrim, and V. Sundararajan, *A framework for considering the utility of models when facing tough decisions in public health: a guideline for policy-makers*, Health Research Policy and Systems, 20 (2022), p. 107.

[17] C. J. Tomlin and J. D. Axelrod, *Biology by numbers: mathematical modelling in developmental biology*, Nature Reviews Genetics, 8 (2007), pp. 331–340. Number: 5 Publisher: Nature Publishing Group.

[18] J. V. Uspensky, *Introduction to Mathematical Probability*, Nature, 141 (1938), pp. 769–769. Number: 3574 Publisher: Nature Publishing Group.

[19] D. J. Weiss and J. Shanteau, *The futility of decision making research*, Studies in History and Philosophy of Science Part A, 90 (2021), pp. 10–14.

Chapter 2

Forecasting SARS-CoV-2 Hospitalizations in Utah with Multiple Public Health Metrics

Since the emergence of COVID-19 in late 2019, much effort has been spent on developing methods to forecast COVID-19 cases and hospitalizations. As we transition from epidemic to endemic and test capacity decreases, hospitalizations are increasingly a primary indicator of COVID-19 burden, and the complicated dynamics of multiple circulating respiratory infections will need to be considered and accommodated. Forecasting hospitalizations is of great interest to public health officials, hospital administrators, and the community, who are using forecasts of hospitalizations to make decisions regarding policy, interventions, and personal risk management. Hospitals make staffing decisions based on anticipated needs and can temporarily increase ICU bed capacity in response to these needs. Percent positive rate, the ratio of positive SARS-CoV-2 tests to the total number of tests administered, has been used throughout the COVID-19 pandemic as a proxy for the current level of transmission in a community.[6] Simultaneously, wastewater SARS-CoV-2 surveillance has been initiated but its direct utility has yet to be fully explored. As society shifts focus from pandemic response to endemic management, testing efforts have been reduced, so the predictive value of the test percent positive rate has been called into question, and case incidence data has diminished. To explore changing transmission dynamics, we build models incorporating SARS-CoV-2 tests, wastewater SARS-CoV-2 levels, and Syndromic surveillance from Emergency Room (**ER**) visits. We successfully forecast hospitalizations at the state level with

greater accuracy than naive or Auto Regressive Integrated Moving Averages (**ARIMA**) forecasts based on hospitalizations alone, a method used by state officials until recently.[19] Forecasts three weeks into the future are developed with sliding test windows and increasing test window cross-validation. We validate and quantify uncertainty in commonly used public health metrics and explore differences in model selection between variants. Data from the winter 2022-23 season are included as a final test for the model. This work examines how to predict hospitalizations in a changing testing environment effectively.

2.1 Introduction

At the beginning of the COVID-19 pandemic, the Utah Department of Health and Human Services (**UDHHS**) instituted COVID-19 surveillance efforts. As the pandemic progressed, the percent positive rate was emphasized in the media and published by public health as a metric for assessing community risk.[6] Simultaneously, wastewater SARS-CoV-2 monitoring was implemented, and syndromic added codes for preliminary COVID-19 diagnoses to the International Classification of Diseases Revision 10 (**ICD-10**). UDHHS and the University of Utah initiated a collaboration to improve forecasting capabilities and experiment with various intervention scenarios. Now three years into the pandemic, COVID-19 testing locations are being dismantled, and testing capacity across the state is diminishing, with home tests replacing reportable laboratory results. There are lingering questions from the UDHHS regarding whether the percent positive rate will be a good metric for future outbreaks and how to forecast hospitalizations most effectively moving forward.

The ongoing collaboration between UDHHS and the University of Utah resulted in this project which was completed over the course of an internship at UDHHS in the Division of Population Health, Office of Communicable Diseases with the Disease Response Evaluation Analytic and Monitoring (DREAM) program. The internship position is supervised by Co-PI Joel Skaria, Biostatistician within the DREAM program. Sam Lefevre, Director of the Office of Communicable Diseases, has collaborated by directing aims and revising questions for maximum public health utility. Wastewater data and insights have been pro-

vided by Nathan LaCross, supervisor of the Wastewater monitoring program. Syndromic ER COVID-19 data was contributed by Randon Gruninger. This protocol has been approved by both the UDHHS and University of Utah IRBs.

The approaches described in this chapter aim to develop better forecasts of COVID-19 hospitalizations. This study uses data collected by UDHHS for SARS-CoV-2 surveillance efforts. These data include reportable SARS-CoV-2 testing, COVID-19 hospitalizations, wastewater SARS-CoV-2 levels, and pre-diagnostic COVID-19 syndromic surveillance ER data streams from Utah. The initial objective of this project is to identify if the SARS-CoV-2 test-positive rate was or remains a leading predictor of COVID-19 hospitalizations in Utah. A secondary objective is to examine the relationship between hospitalizations and wastewater SARS-CoV-2 levels and pre-diagnostic COVID-19 syndromic surveillance from ERs. Ultimately we aim to create a forecast of COVID-19 hospitalizations in Utah to be used by public health officials.

We first establish whether the test-positive rate was and remains a leading predictor of SARS-CoV-2 hospitalizations in Utah. To do this, we first examine lags in the cross correlation[25] between various public health data streams and hospitalizations. We then fit various GAM[20] models and compare their predictive power to an ARIMA[12] forecast of hospitalizations based only on hospitalization data. We incorporate testing, pre-diagnostic COVID-19 syndromic surveillance,[17] and wastewater SARS-CoV-2[24] data to refine predictive capability and improve our forecasting capacity. This retrospective analysis seeks to improve ongoing forecasting efforts by allowing the calibrations of estimations of case burden from periods of high testing and utilizing that information to link to wastewater and syndromic data as testing diminishes.

2.2 Background

Since the emergence of SARS-CoV-2 in late 2019, there have been many challenges in disseminating guidance and validating data collected included in analysis of transmission dynamics. Heterogeneity across space, time, behavior, immunity, and policy have all been critical in shaping these transmission dynamics. It is difficult to capture this in a study spanning multiple years. Instead,

we aim to develop a generalizable approach that can be utilized for ongoing COVID-19 efforts and extended to other respiratory infections with careful tuning.

The ratio of positive SARS-CoV-2 tests to the total number of tests administered, the test positive rate, has been used throughout the pandemic as a proxy for the current level of transmission in a community,[23] with public health guidance often set based on the local percent positive rate in tandem with hospital burden. As the COVID-19 response transitions from pandemic to endemic, testing efforts have been reduced, and the predictive value of the test percent positive rate has been called into question.

Previous studies have shown that the test positive rate can lead COVID-19 hospitalizations by two weeks.[7] Test-positive rates among the non-elderly can predict case count increases for the elderly 14 days in the future.[8] As the testing landscape shifts, we ask whether testing data improves our forecasts of hospitalizations in Utah and for what time frames that is the case. We then incorporate additional data streams, such as wastewater and syndromic surveillance, to refine predictive capabilities.

Syndromic surveillance data are discharge diagnoses collected by ERs and sent to health departments. Syndromic surveillance from Switzerland has been shown to lead COVID-19 hospitalizations by 13 days.[4] In Utah, all emergency rooms participate in the UDHHS syndromic surveillance program. These data include pre-diagnostic reports of potential COVID-19 cases before laboratory testing has been processed and reported. The data tend to be low count, so people typically rely on them for aberration detection rather than prediction.[17] These data have limitations but have great potential to be used with other surveillance data to improve predictive capabilities.

SARS-CoV-2 wastewater surveillance data has been shown to be correlated with COVID-19 hospitalizations,[29] and in some cases is a leading predictor of cases, hospitalizations,[15] and ICU admittance.[9] Wastewater data are highly variable and not available for the entirety of the epidemic. However, they continue to be collected as testing capacity diminishes and more people utilize at-home rapid tests.

Additionally, as masking and other pandemic mitigation measures are re-

duced, transmission of other seasonally circulating respiratory viruses will likely increase.[16,26] As such, it will be imperative to build up our predictive capacity to be able to prepare for spikes in cases and hospitalizations for COVID-19, its newly emerging variants, and the seasonally circulating strains of Influenza and RSV that regularly strain healthcare resources.[2,3,10] Hospitalizations amongst the elderly are of particular concern, as they are the most likely to die from infection,[21] and most likely to remain hospitalized for long periods of time. Hospital burden in general remains important for other age groups, especially as the consequences of long-haul COVID become apparent.

With this project, we evaluate the reliability of the SARS-CoV-2 percent positive rate as a metric and use available public health data streams to build a predictive model of COVID-19 hospitalizations in Utah. We explore how and whether additional public health data streams lead hospitalizations and how those leads change over time in a varied transmission landscape.

2.3 Methods

Initial analyses explore whether the SARS-CoV-2 test positive rate is, or was, a leading predictor of COVID-19 hospitalizations in Utah using time-lagged cross-correlation as preliminary analysis to quantify the data synchrony in time. We then develop univariate forecasts for all data streams and subsequently add combinations of streams and wastewater SARS-CoV-2 levels for the time periods for which they were available. A forecasting flow (**Figure 2.1**) is included for clarity with components described below in 2.3.7. All forecasts are developed on 84-day training windows and projected three weeks into the future blind to any new data during the forecasting window. Variable test demand, implementation of novel wastewater monitoring techniques, and pre-diagnostic syndromic test results all interact in a highly variable landscape of behavior and disease transmission. Each of these data streams has relative strengths and weaknesses. To most effectively leverage those, we first assess the quality of each data stream individually and then incorporate them into a more complex forecasts.

To develop forecasts, we test naive, ARIMA,[12] and Generalized Additive Models (**GAM**).[28] We calculate Akaike Information Criterion (**AIC**)[5] to assess

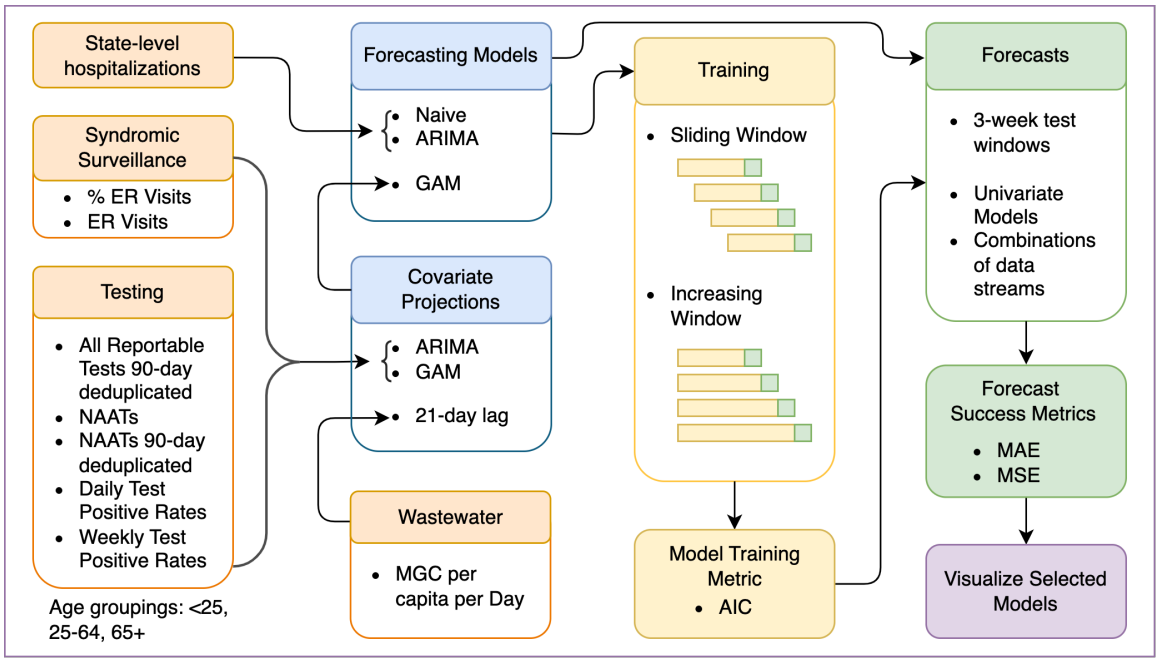


Figure 2.1: Diagram of Forecasting Methodology
To the left in orange are the data streams and their component parts which are described in detail in the data section. Then in the blue boxes are the models used for forecasts and models used for univariate projections to inform the forecasts after the available data end. The models are then trained on 12-week testing windows with forecasts calculated after the end of the training window. We compare two cross validation methods for model training: sliding window and increasing window cross validation and calculate the AIC to assess model fit. Trained models are then used on projected covariates to make forecasts three weeks in advance which are assessed using MAE and MSE. The models with lowest forecast error are then visualized.

model fit, Mean Absolute Error (**MAE**), and Mean Squared Error (**MSE**)[13] to assess forecast accuracy. Together we use these metrics to model select. All data processing and computational analysis are conducted in R version 4.2.2.[22] Methods are presented in greater detail in the following sections.

2.3.1 Data

The data streams from UDHHS include limited reportable SARS-CoV-2 testing results, COVID-19 hospitalizations, wastewater SARS-CoV-2 levels, and pre-diagnostic COVID-19 syndromic surveillance data from those who reside in the State of Utah, the records of over 3.3 million individuals. People of all ages

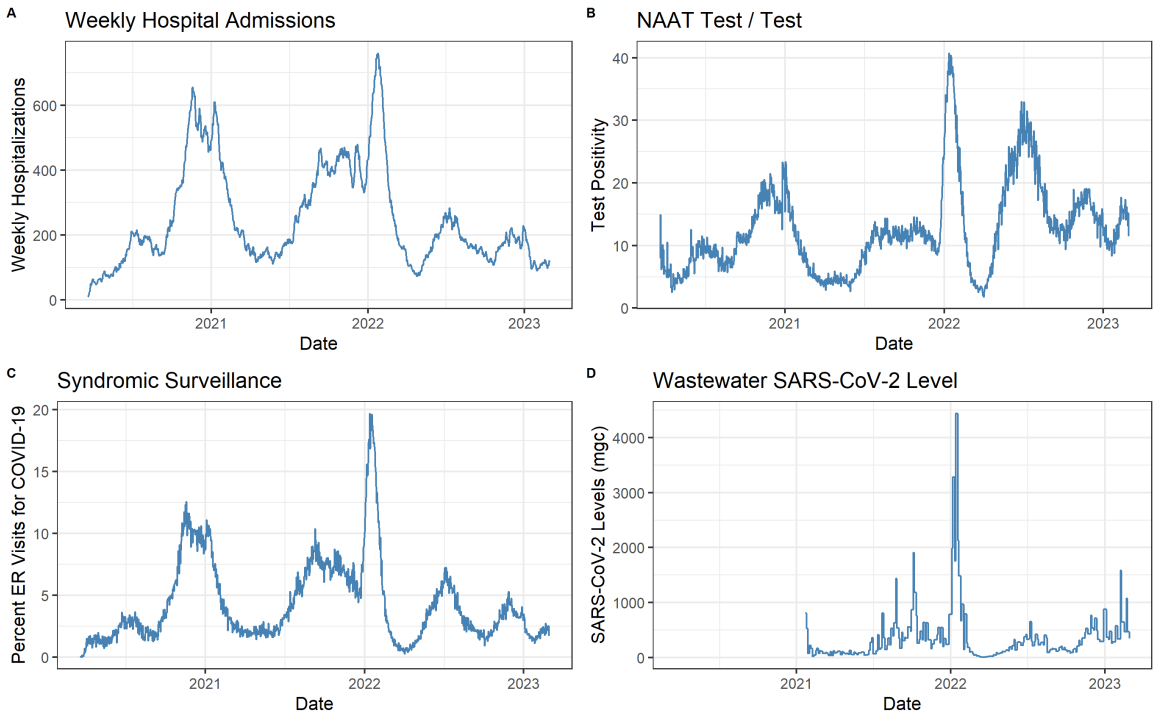


Figure 2.2: Time Series of Statewide Data Streams.
Panel A shows weekly hospital admissions, B shows NAAT T/T percent positive, C shows the percentage of ER visits with a preliminary COVID-19 diagnosis, and D shows wastewater levels in the Central Valley Sewershed in millions of gene copies per capita per day.

are at risk of infection and hospitalization from COVID-19,[21] so individuals are included in the study in the direct proportion that they occur within the existing dataset. The available data streams are described in greater detail in the following sections and summarized in the panels of **Figure 2.2**, which includes time series of Hospitalizations, Test Positive Metrics, Syndromic, and SARS-CoV-2 wastewater levels. We begin the analyses of testing, syndromic ER, and hospital admissions on March 20th, 2020 and end them on February 28th, 2023. For analyses including wastewater, we begin on January 1st, 2021, and end on February 28th, 2023.

Hospitalizations

This study uses daily hospital admission data from March 21st, 2023, to February 28, 2023. Residents of Utah are included in the hospital data at the rate at

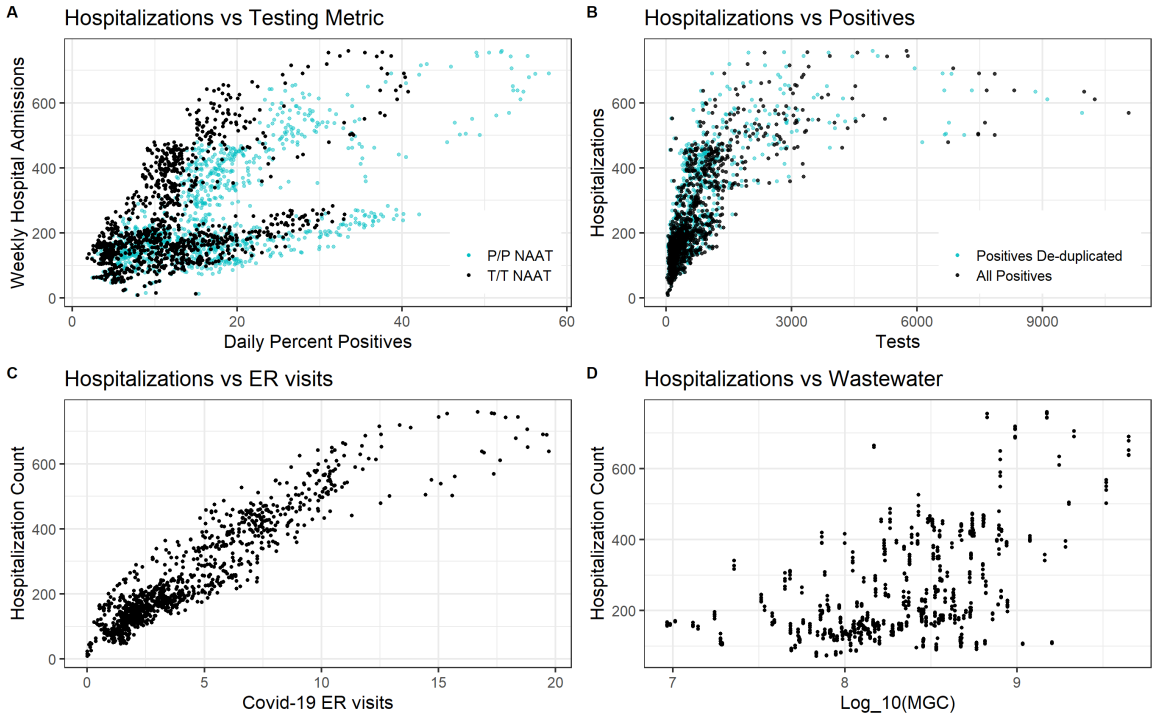


Figure 2.3: Plots of the Relationship between Hospitalizations and Selected Data Streams.
A compares hospitalizations on the y-axis to the test metrics on the x-axis and is split into all tests P/P in teal and NAAT T/T in black. B captures Hospitalizations vs number of positive NAATs, where the teal represents de-duplicate positive tests and the black represents total positive tests. C has hospitalizations on the y-axis vs COVID-19 ER visits as captured by syndromic surveillance each black point represents a day. In D, hospitalizations are plotted on the y-axis vs the wastewater parameter Log(MGC) on the x-axis each point represents a day when a sample of wastewater was collected.

which they were hospitalized with COVID-19. These data are grouped into age categories: below 25 years old, 25 to 64 years old, and above 65. A time series of weekly hospital admissions are plotted in **Figure 2.2** Panel A. Weekly hospital admissions are calculated daily using a summation the hospital admissions from the preceding seven days. In an exploratory analysis of the relationships between the data streams, we generate scatter plots between each of the data streams and hospital admissions. The data streams have visible relationships shown in the scatter plots in **Figure 2.3** with the panels described in more depth in the following subsections. The ultimate goal of this project is to forecast total hospital admissions three weeks in advance.

Tests

Various factors have complicated testing dynamics in Utah: availability of Polymerase Chain Reaction (**PCR**) sequencers for Nucleic Acid Amplification Tests (**NAAT**), accessibility of testing, speed of test results, and messaging from state officials. Colloquially, NAAT tests are often called PCR tests; we use NAAT to match UDHHS terminology. Rapid antigen tests, not reportable when administered at home, have become available and dispersed publicly as the pandemic has progressed, so the number and dynamics of reportable tests have also shifted. Negative antigen tests were no longer reportable as of March 31st, 2022.[23] State testing operations began to wind down in mid-2022 and were eventually fully closed in early 2023 as test demand diminished. Residents of Utah are included in the testing dataset at the rate with which they were tested for SARS-CoV-2. Non-residents who were tested in Utah are excluded from analysis.

While the test positive rate is a seemingly simple metric, there are various methods to calculate it, and the best is open to debate. NAAT test-over-test (**T/T**) is calculated by dividing the total number of positive tests by the total number of NAAT tests without de-duplication. In the NAAT person-over-person (**P/P**) percent positive, the number of individuals with positive NAATs is divided by the number of individuals with NAATs with a de-duplication where the individual is only counted once per day. **Figure 2.2** Panel B shows a time series of the NAAT T/T metric. In **Figure 2.3** Panel A, there is a relationship between test positive rate, but it appears to be very imperfectly correlated and there is no obviously apparent best calculation method for test positive rate. In Panel B, there is again a relationship between the positive tests and hospitalizations and it does not appear to be linear. In preliminary analyses T/T percent positive was slightly more predictive than other methods of calculating percent positive, so subsequent analyses used that metric. The test positive rate metrics all had high levels of concurvity, so it is likely that with proper parameterization they would all be similarly effective. The state also calculates a test positive metric which includes reportable antigen tests, but since the dynamics around antigen tests have changed over the course of the epidemic, they were excluded from these analyses. Tests can also be split by age and are aggregated into tests of

school-age individuals younger than 25, working-age from 25 to 64, and seniors 65 or older.

Syndromic Surveillance

Whenever patients are seen at ERs, the state receives a record of the ICD-10 codes for their diagnoses associated with a unique patient identifier.[17] The syndromic surveillance data are collected daily by ERs and reported to UDHHS. The state then uses this information to calculate the percentage of ER visits associated with influenza-like illness for public display. Our study uses data related to COVID-19-associated ER visits. The data are based on preliminary diagnoses and may be shown to be misdiagnosed as NAAT results are returned. These data are shown in **Figure 2.2** Panel C. We see a very strong linear relationship between weekly hospitalizations and COVID-19 ER visits (**Figure 2.3** Panel C). Syndromic data can also be split by age and are aggregated into tests of school-age individuals younger than 25, working-age individuals aged 25 to 64, and seniors 65 years or older. In this study, we calculate the total number of ER visits for COVID-19 and the fraction of ER visits who leave with COVID-19 diagnoses divided by the total number of ER visits for COVID-19 and those metrics for each subgroup.

Wastewater

In Utah, SARS-CoV-2 wastewater surveillance was refined throughout the pandemic. Wastewater samples are collected in Utah at 34 sewage treatment facilities covering approximately 88% of Utah's population.[6] The data are highly variable and are susceptible to variations in: seasonal temperatures, which alter the rate of viral RNA degradation; industrial chemical contamination, which can prevent amplification; and rainfall, which enters the sewershed and changes the concentration of the effluent. Since the data are collected at the sewershed-level, they are also potentially vulnerable to the effects of visitors, tourists, and commuters. These additional groups are unlikely to contribute much to the wastewater SARS-CoV-2 levels. However, infectious individuals visiting the sewershed could slightly raise levels before those infectious are captured via other surveillance efforts. The Centers for Disease Control (**CDC**) suggests

that fecal-content normalization could mitigate these effects during periods of high tourism or changing sewershed populations,[1] but this will not be explored in these analyses. Wastewater data are shown in **Figure 2.2** Panel D. The included data were collected from the Central Valley Water Reclamation Facility (**CVWRF**) sewershed. The CVWRF covers 515,494 individuals which accounts for approximately 15.4% of Utah's population.[6] The forecasts developed below incorporate wastewater data from only this sewershed. For the wastewater data streams we explore daily concentration in million gene copies per person in the sewershed per day (**MGC**), which is calculated by taking the average concentration of SARS-CoV-2 replicates measured in the wastewater effluent by the amount of water flow through the wastewater reclamation center and sewershed population. It is calculated via this equation which is used by UDHHS and matches methods encouraged by the CDC.[1]

$$MGC = \frac{(Average\ Concentration) * (Flow\ per\ Day)}{(Sewershed\ Population)} \tag{2.1}$$

We then take the $Log_{10}(MGC)$ to reduce skewness in the wastewater data. The CDC suggests that wastewater data are likely to be log-normal and so should be log transformed prior to computing trends and relevant statistics.[1] Since results are very similar between MGC and its log, we will use the logged value for the remained of this chapter.

Wastewater data were collected at various intervals over the course of the epidemic and began to consistently be collected twice a week at the beginning of 2021. The first forecast training window starts on January 1st, 2021, for the analyses incorporating wastewater data streams. Since wastewater data are collected twice weekly, they are on a different time scale than syndromic, testing, or hospitalization data. In order to include this data in the forecasting methodology, we carry the most recent value forward until there is another measurement for daily analyses and average the two weekly measurements for weekly analyses. This interpolation is chosen because of reporting delays and the high variability in wastewater measurements. It is clear that there is a relationship between wastewater and hospitalizations, but it is not as strong as those from the other test metrics (**Figure 2.3** Panel D).

2.3.2 Variants

To better understand the dynamics of SARS-CoV-2 variants and see temporal differences as the epidemic progresses, we split the data into five sections based on the dominant variant at the time. These dates shown in **Table 2.1** are based on the points in time when the new variant overtook the previous variant in sequenced NAAT results.[23] These are relatively imprecise dates and not indicative of all variants sampled at any particular time but intended to provide a rough range of what was occurring at various points throughout the epidemic.

Table 2.1: Estimated Dates when Variant Dominance Shifted.
These Dates were gathered visually from graph of variant trends from the state of Utah. In reality there are large overlaps between the occurrences of the variants, and this serves as a way to subdivide the data and explore variation in time.

Variant	Start Date	End Date
Original	2020-03-02	2021-06-01
Delta	2021-06-02	2021-12-22
Omicron 1	2021-12-23	2022-03-24
Omicron 2	2022-03-25	2022-10-12
Omicron 3	2022-10-13	2023-02-17

2.3.3 Benchmarks

In order to benchmark the forecasting results, we explore naive forecasts which carry the last value forward from the training dataset and uses a confidence interval based on the variance of the residual error of one-step forecasts on the test data, $\hat{\sigma}$, and the number of days ahead h. Here e_t is the error, which for the one-step naive method is the difference between the number of hospitalizations on a particular day and the number of hospitalizations on the previous day.

$$\hat{\sigma} = \sqrt{\frac{1}{T-1}\sum_{t=1}^{T} e_t^2}, \tag{2.2}$$

Here T is the number of days in the training window and the the confidence intervals, $\pm\hat{\sigma}_h$, for the naive forecasts are given by this equation. $\hat{\sigma}_h = \hat{\sigma}\sqrt{h}$.[11]

2.3.4 Auto Regressive Integrated Moving Averages

In this work, we take ARIMA as the null model and aim to develop forecasts with smaller forecasting errors than those developed fitting hospitalization data using the `auto.arima` method which is available as a component to the `forecast` package in R.[11] ARIMA work by leveraging autocorrelations in the data, but are best suited to stationary data[14] which are unlikely in an emerging epidemic. This method is readily available to public health officials and was used by public health officials at UDHHS during the epidemic's early stages.

2.3.5 Generalized Additive Models

GAMs provide an extension of linear models which can be interpreted as assuming a smooth prior to model parameters.[20] Here this assumption implies that there is some underlying dynamic which is represented by a smooth but potentially non-linear interaction between parameters. We use GAMs to incorporate the syndromic, testing, and wastewater data streams in multiple ways. To achieve credible intervals for these models we use the `mgcv` package and results from it take a Bayesian interpretation of the confidence intervals[28] which allows for the assumption of a smooth, if non-linear, relationship between the covariates included. These intervals have been shown to converge to the frequentist approach in simulations.[18]

2.3.6 Model Selection Metrics

For the ARIMA null model and all the variations of the GAMs, model fit is assessed using the AIC. The AIC is asymptotically equivalent to minimizing the one-step forecast error, so by construction will be lowest for the naive model,[27] but the naive model will likely run into problems when it is applied to the forecast window. AIC includes a penalty for model complexity and so might encourage the use of overly simplistic models like the naive. The AIC generated from the `mgcv` package for the GAMs and the `forecast` package for the arima

models is conditional and involves the likelihood of the model estimates at their penalized MLEs with the addition of a correction for smoothness selection. The conditional AIC allows for comparison of models across model types, i.e. GAM and ARIMA. We use AIC to assess the trained models' success over the training data, and use it to compare across model types.

2.3.7 Forecast Methodology

The forecasting methodology is presented in a flow in **Figure 2.1**. We first perform forecasts using univariate data streams to select the most effective data streams from each category: testing, syndromic, and wastewater before moving on to combinations from multiple data streams. To avoid concurvity issues, we do not combine multiple percent positive metrics but rather include combinations of percent positive rate, number of positive tests, and number of tests as parameters within the GAM forecasts.

We assess the success of models at windows of one-, two-, and three-weeks in the future, with training windows selected to be at least four times the duration of the testing window to correspond with an 80/20 training and testing split. There is significant weekly periodicity in the data, with more hospitalizations occurring on Mondays than Fridays, so we choose forecast hospitalizations aggregated weekly to avoid the impacts of those trends. We focus our effort on three-week forecasts as those are the most valuable for short-term future planning. It is possible to increase this forecast to a four-week window, however the projection of the covariates necessary for the forecasts will increase the corresponding uncertainty in the forecast.

We calculate Mean Absolute Error (**MAE**) and Mean Squared Error (**MSE**) to assess model success over the forecast windows. MAE is calculated by taking the absolute value of the differences between the predicted values and the true number of hospital admissions and dividing by the number of predictions. MSE is calculated by summing the squares of the differences between the predicted numbers of hospital admissions and the true numbers and dividing by the number of predictions. MSE penalizes outliers more heavily than MAE, but from a public health perspective, MAE is more interpretable and can be presented as the daily expected difference between the true number of hospitalizations and

the number forecast. We differentiate between model fit and forecast success. The forecast success metrics demonstrate how well the trained models are able to forecast the hospitalizations and are ultimately more important than how well the models fit the training data. Amongst models of similar forecast error, we choose the model with lower AIC as it is less likely to have overfit the data. Conversely, if there is a model with high AIC, but low forecast error, that indicates that the additional complexity of the model might be contributing to the success of the forecast.

We compare results from 2 different time series cross validation strategies, sliding test window and increasing test window, to assess the impact of the inclusion of previous data on model accuracy. For all analyses we fix the minimum training window to be 12-weeks, which corresponds to an 80/20 split of training and testing data for the three-week forecast. See **Figure 2.1** for a pictorial explanation for the training and testing split, and a diagram of the two cross validation schemes; sliding and increasing window, which are described below. We examine the results of forecasts one-, two-, and three-weeks ahead using the forecast error metrics, MAE, and MSE described above.

The sliding window method maintains the 12-week training window and slides over the dataset increasing by the length of the prediction window at every step. The metrics are then calculated for each set of testing and training windows, and averaged over the entire period of analysis, from March 20, 2020 to February 28, 2023 for testing and syndromic, and from January 1st, 2021 to February 28, 2023 for comparative analyses including wastewater.

For the increasing training window approach, the length of the training window is increased by the prediction window for every iteration of the model so the training set grows with every new prediction window. Thus, this method uses all of available past data and is potentially more susceptible to biases based on changing behavior, immunity, and data collection methodology.

Ex-post Evaluation

In order to evaluate the potential for success and establish error bounds on ex-ante forecasts, we fit GAMs to the prediction windows including the true values for the covariates over the prediction windows. This demonstrates how

errors change with the side of the prediction windows, sets optimistic goals for errors, and informs which models have the most potential for successful forecasts. We use values to see what models might be successful in an ex-ante forecasting regimen where the model does not have access to any values of the covariates.

Ex-Ante Evaluation

We develop weekly ex-ante forecasts one-, two-, and three-weeks into the future by building projections of covariates three weeks beyond the training windows. Based on the results of the ex-post forecasts, we build projections of percent ER visits for COVID-19, senior ER visits for COVID-19, all senior positive tests, and senior T/T test positive rate. We project these covariates using GAM and ARIMA models. Wastewater values are projected using a three-week lag which is maintained for both the training and testing windows. Since hospitalizations are aggregated weekly, we calculate forecasts of hospitalization for the day 21 days after the end of the training window. Since the wastewater values are not smooth, we include the lagged wastewater values in the GAM forecasts without the assumption of a smoothed basis function for the relationship between wastewater and hospitalizations. To most accurately mirror the approach which would be taken by public health in an emergent situation, we forecast a week of hospitalizations three weeks in advance, and then move the training window forward by one week and forecast ahead another three weeks. We assess the forecast success metrics calculated on that last day.

2.4 Results

2.4.1 Lags

In order to analyze the correlation between data streams and hospital admissions in time, we calculated cross correlations. For each data stream, the lags for which the correlation was highest between various data streams are presented in **Figure 2.4**. This figure is divided into panels by time periods given by the dates in **Table 2.1**. Later we forecast overall hospitalizations, but for this analysis the columns of the panels are divided into total, school-aged, working-age, and

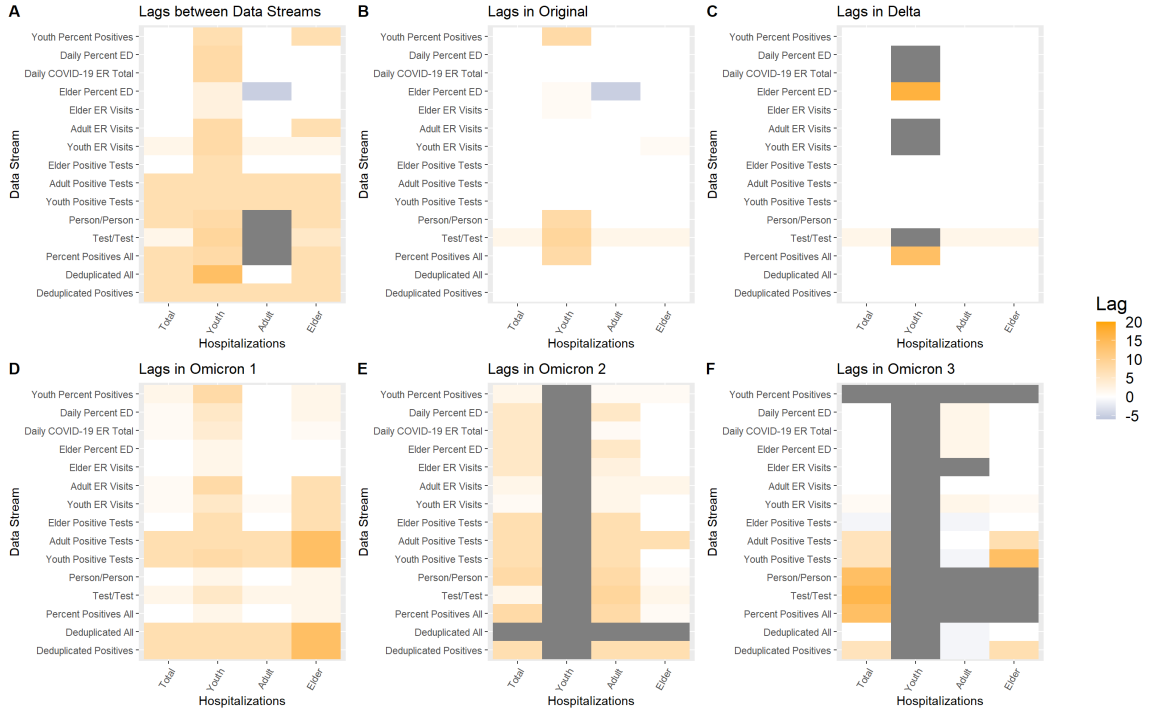


Figure 2.4: Differentiating Lags between Selected Data Streams. Total hospitalizations and age groupings of hospitalizations. These lags are calculated over the entire dataset in A, and then for the original variant in B, Delta in C, Omicron Phase 1 in D, Omicron phase 2 in E, and Omicron phase 3 in F. The scale ranges from -5 in light blue, indicating that a data stream trails hospitalizations, through 0 in white to 20 in orange indicating that a data stream leads hospitalizations. Lags with correlation values less than 0.5 are greyed out.

senior hospitalizations. Here the positive lags indicate that the data stream leads hospitalizations by the number of days given. We can see that the dynamics between hospitalizations and these highlighted data streams evolve over time. As time progresses, more of the correlations are below the threshold of 0.5. School-age hospitalizations has the most N/A values, and this is likely due to the infrequency of hospitalizations in school-age individuals. Interestingly, the only negative lag is for the senior % ER and working-age hospitalizations, meaning that working-age hospitalizations change, and then the percent of the senior group visiting the ER for COVID-19 changes. There are no N/A values in the Omicron 1 period, the period in which there was the most testing capacity in Utah.

2.4.2 Cross-validation

The MSE of ex-post predictions are shown in **Figure 2.5**. In this plot we see several selected models based on a variety of data streams for the sliding window and increasing window methods. The errors for all models tend to increase as the forecast length increases despite normalization by number of predicted values. This suggests that temporal distance from the training dataset does decrease model accuracy even when the model is not blind to the covariates. Since the forecast error metrics for the sliding window were consistently lower than those for the increasing window, we show results for the sliding window forecasts for the remainder of the results.

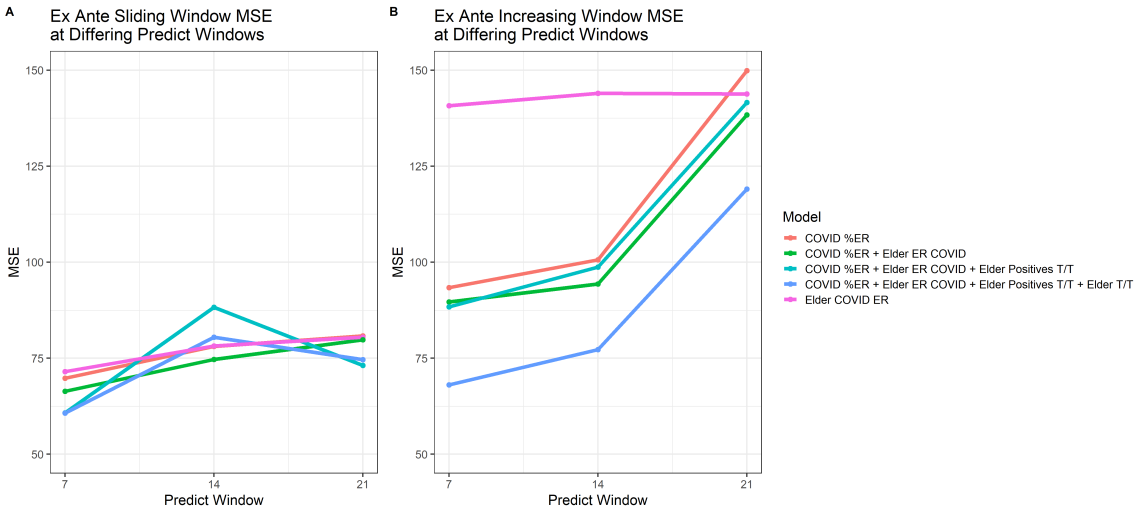


Figure 2.5: MSE over 1-, 2-, and 3-week Forecast Windows with Sliding Window and Increasing Window Training Paradigms.
A) is the MSE for the sliding window, while B) is the MSE for the increasing window. All compare the same set of 7 models: ARIMA, % of ER visits for COVID, % ER + Seniors in the ER for COVID, %ER + Senior COVID ER + Senior Positive Tests, %ER + Senior COVID ER + Senior Positive Tests + Senior T/T, and Naive.

2.4.3 Ex-post Univariate Predictions

A selection of the most successful data streams from the testing is shown in **Table 2.2**. In this table we can see that the naive benchmark actually outperforms the ARIMA null model for all but the 7-day forecasts. This is likely because there ARIMA models tend to lag their data bu only a few days. The total positive

tests for the senior group (Senior Positives) is the most successful for ex-post hospitalization prediction accuracy across all metrics. The most successful

Table 2.2: Testing Univariate Metrics

	MAE			MSE			AIC
	21-day	14-day	7-day	21-day	14-day	7-day	
Naive	9.26	8.47	7.67	168.83	137.55	110.74	379.99
ARIMA	10.17	9.01	7.54	187.33	152.33	105.02	582.06
NAAT P/P 7-Day Avg	8.78	8.6	7.9	130.15	126.67	106.6	588.27
NAAT T/T 7-Day Avg	9.02	8.4	7.81	144.55	131.08	105.48	587.6
Senior Positives De-duplicated	8.46	8.4	7.54	127.74	132.49	101.56	580.88
Senior All Positives	8.26	8.1	7.37	120.83	117.35	95.63	579.78

streams from the syndromic data are presented in **Table 2.3**. All of these streams have lower prediction error than the naive benchmark and ARIMA null models. They are also more successful than the models developed on the testing data (**Table 2.2**). These models have huge improvements in terms of both absolute error and squared error, suggesting not only that the predictions are more accurate, but also that they are less vulnerable to outliers. The single best metric from the syndromic data is the number of seniors visiting the ER for COVID-19, shown in the last line of the table.

Table 2.3: Syndromic Univariate Metrics Sliding Window

	MAE			MSE			AIC
	21-day	14-day	7-day	21-day	14-day	7-day	
Naive	9.26	8.47	7.67	168.83	137.55	110.74	379.99
ARIMA	10.17	9.01	7.54	187.33	152.33	105.02	582.06
COVID ER	7.1	6.9	6.51	89.18	84.11	74.05	566.54
COVID %ER	6.8	6.64	6.32	80.79	78.07	69.76	568.49
Senior COVID ER	6.79	6.7	6.42	80.49	78.19	71.53	576.64

2.4.4 Ex-post Multivariate Predictions

We then combine data streams from both testing and syndromic datasets, and have highlighted some of the most successful in **Table 2.4**. Usually, increasingly complicated models result in trade-offs between model fit and test accuracy.

Here we can see that there are genuine improvements in AIC as we add additional parameters to the forecast, in addition to improvements in success metrics. This demonstrates that the combinations of these multiple data streams improves overall model fit for the training data as well as forecast accuracy in the test data. We have included single streams in this table as a comparison in order to show the relative improvements of the multivariate forecasts over the single stream predictions. We include wastewater metrics in various combinations

Table 2.4: Combination Metrics Sliding Window

	MAE			MSE			AIC
	21-day	14-day	7-day	21-day	14-day	7-day	
Naive	9.26	8.47	7.67	168.83	137.55	110.74	379.99
ARIMA	10.17	9.01	7.54	187.33	152.33	105.02	582.06
COVID %ER	6.8	6.64	6.32	80.79	78.07	69.76	568.49
Senior COVID ER	6.79	6.7	6.42	80.49	78.19	71.53	576.64
COVID %ER + Senior ER COVID	6.68	6.5	6.17	79.82	74.67	66.39	565.03
COVID %ER + Senior ER COVID + Senior Positives	6.45	6.5	5.9	73.1	88.23	60.76	550.92
COVID %ER + Senior ER COVID + Senior Positives+ Senior T/T	6.37	6.33	5.8	74.61	80.45	60.73	546.57

alone and in combination with other data streams in **Table 2.5**. In this table we can see that the wastewater data streams alone are not particularly successful with forecast success metrics below those of the naive benchmark and ARIMA null models. When added as an additional component to the other successful data streams and combinations, we can see clear improvements. While alone the $MGC\ Capita$ has indeterminate success, it is combined with other metrics, those models tends to improve noticeably. Interestingly, the most successful model in the ex-post approach is one which is made up of senior ER visits for COVID-19, all positive NAAT in seniors, NAAT T/T in seniors, and $MGC\ Capita$.

2.4.5 Ex-ante Forecasts

The projections of the covariance with GAMs tended to have slightly less forecast error than those made with the ARIMA models. For the ex-ante forecasts, the relative successes of the models remained similar when the forecast was started on different days of the week, however the differences in forecast error between models were significantly smaller while both MAE and MSE was higher than in the ex-post approach. For the later time periods, the addition of wastewater data

Table 2.5: Wastewater Metrics Sliding Window

	MAE			MSE			AIC
	21-day	14-day	7-day	21-day	14-day	7-day	
Naive	8.38	7.82	7.16	147.3	121.42	94.25	379.62
ARIMA	9.17	8.05	7.11	162.73	123.1	93.97	580.01
MGC	11.04	11.09	9.35	258.15	301.95	170.61	603.86
COVID %ER	6.93	6.57	6.42	83.97	74.09	70.66	571.34
COVID %ER + MGC	6.71	6.23	6.01	77.23	64.34	61.18	565.43
COVID %ER + Senior ER COVID	6.83	6.49	6.32	83.12	73.53	67.64	567.36
COVID %ER + Senior ER COVID + MGC	6.66	6.2	6	76.27	64.63	60.93	562.3
COVID %ER + Senior ER COVID + Senior Positives + Senior T/T	6.66	6.16	5.93	101.44	72.82	63.41	545.99
COVID %ER + Senior COVID ER + Senior Positives + Senior T/T + MGC	6.61	6.04	5.82	104.42	70.53	61.89	543.16
Senior ER COVID + Senior Positives	6.47	6.59	6.05	72.17	80.81	63.72	559.42
Senior ER COVID + Senior Positives + MGC	6.41	6.12	5.87	76.39	66.21	60.55	554.05
Senior ER COVID + Senior Positives + Senior T/T	6.25	6.32	6.07	66.63	71.45	64.69	555.4
Senior ER COVID + Senior Positives+ Senior T/T+ MGC	6.18	6.26	5.98	65.42	70.52	64.29	552.82

significantly improves all of the forecast success metrics over those generated by the naive benchmark and ARIMA null models.

In **Figure 2.6** we can see the AICs for several models for sequential training windows overlaid on weekly hospital admissions. The AIC tends to increase after periods of change in the hospitalizations. This suggests that the models are struggling to appropriately fit the training data at change points, and at those times we should expect forecast error to also increase during those periods.

The MSE for several forecasts are shown in **Figure 2.7**. We can see that the forecasts all tend to be less successful during periods of high variability in hospitalizations, i.e., periods where there are peaks and changing trends. There is no one model which performs best for every prediction period, however the models do tend to be more successful than the ARIMA null forecasts. The naive benchmark does surprisingly well, yet not particularly successful during the peak of the first Omicron phase in January 2022 where the GAMs tend to have smaller values of both MSE in forecast error and AIC over the training window.

Finally, we show the forecasts as compared to weekly hospital admissions in black in **Figure 2.8**. These both have noticeable deviations from the true weekly hospitalization count which is shown in black. The naive forecast is clearly, and by construction, lagging behind the weekly hospitalization counts, while the ARIMA massively overshoots the number of hospitalizations during the largest Omicron peak and then undershoots hospitalizations in the following refractory period. The GAM model forecast presented is fit using Senior ER

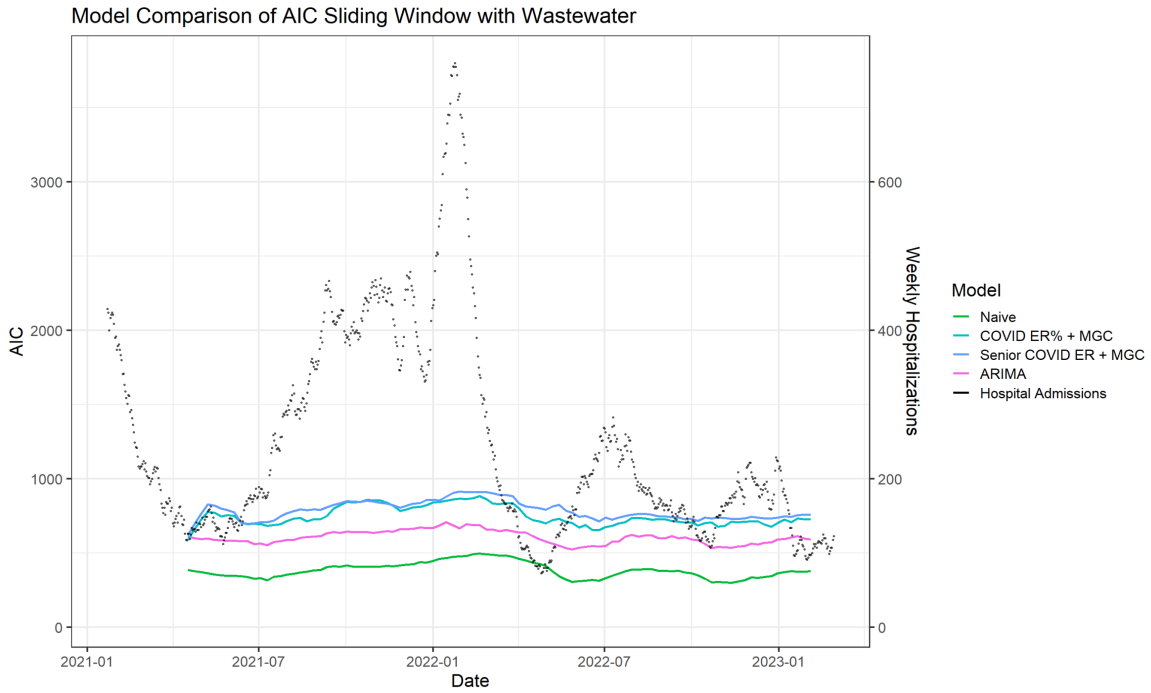


Figure 2.6: AIC Over Time for Selected Models.
On the left y-axis the AICs for sequential training windows are shown for four models with the date along the x-axis. The green line shows the AIC for the naive model. The teal line is the AIC for the COVID ER% with MGC. In blue is the senior COVID ER visits with MGC. The pink line is the AIC for the ARIMA. Weekly hospital admissions are shown in black and correspond with the right side y-axis.

visits for COVID-19 and lagged wastewater SARS-CoV-2 MGC is not a perfect fit of the data, but has better forecast success metrics with a MAE of 7.25 and MSE of 125.22 which can be compared to the naive's MAE of 8.94 and MSE of 190.83. The model also shows significant improvements over the error metrics of the ARIMA forecast with MAE of 9.55 and MSE of 224.39.

2.5 Discussion

The lag plots in **Figure 2.4** demonstrate that while there is a sometimes significant lag between the various data streams and hospitalizations, the dynamics clearly change over time.

The sliding training window tends to result in much better test accuracy than the increasing training window. The AIC from the increasing training windows

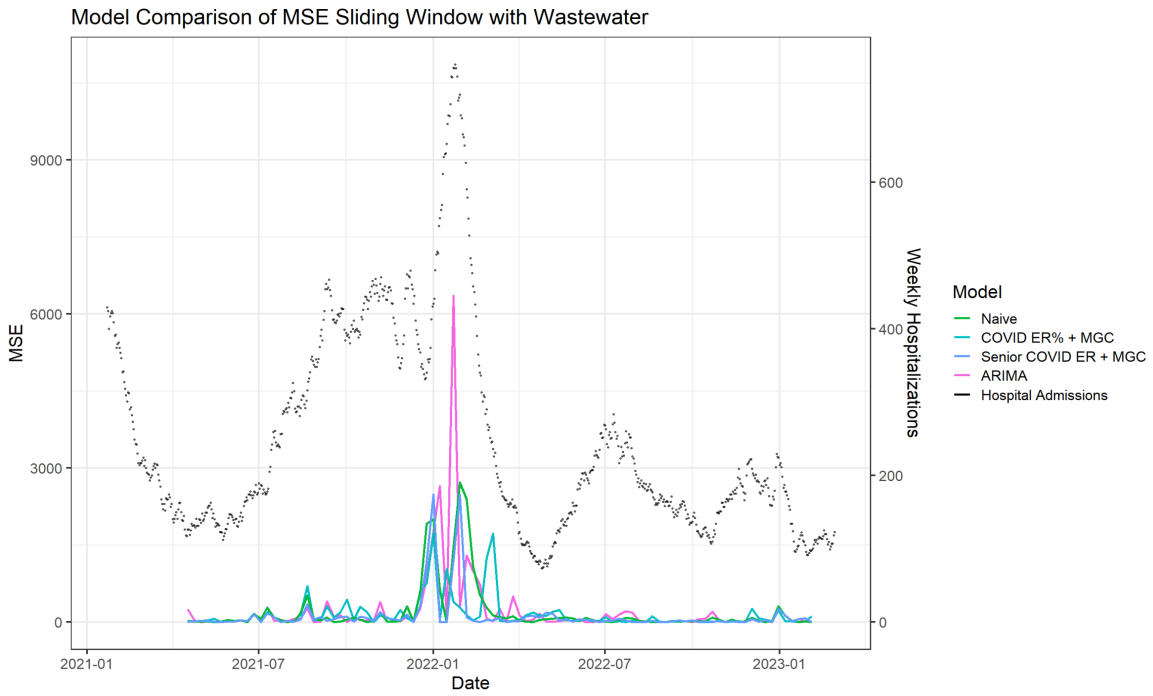


Figure 2.7: MSE Over Time for Selected Models.
On the left y-axis the MSEs for sequential training windows are shown for four models with the date along the x-axis. The green line shows the MSE for the naive model. The teal line is the MSE for the COVID ER% with MSE. In blue is the senior COVID ER visits with MSE. The pink line is the MSE for the ARIMA. Weekly hospital admissions are shown in black and correspond with the right side y-axis.

are much larger than for the sliding training window, which corresponds with the progressively longer windows and more opportunity for accumulated residuals. These AIC values are to be expected as the training data gets larger, however the increasing MAE and MSE suggest that there are changing dynamics which can not be captured by fixed parameter estimates.

The AIC in the naive model is low, however the forecast accuracy of many of the other models in both the ex-post and ex-ante schemas exceeds that of the forecast accuracy of the naive benchmark. By construction, the AIC will always be low for the naive model. Thus, the naive model will have a very good fit the training data, but ultimately performs more poorly for the actual forecasts. This is one of many reasons that AIC can not be the only metric relied upon for model selection. The accuracy of the naive model tends to beat that of the ARIMA

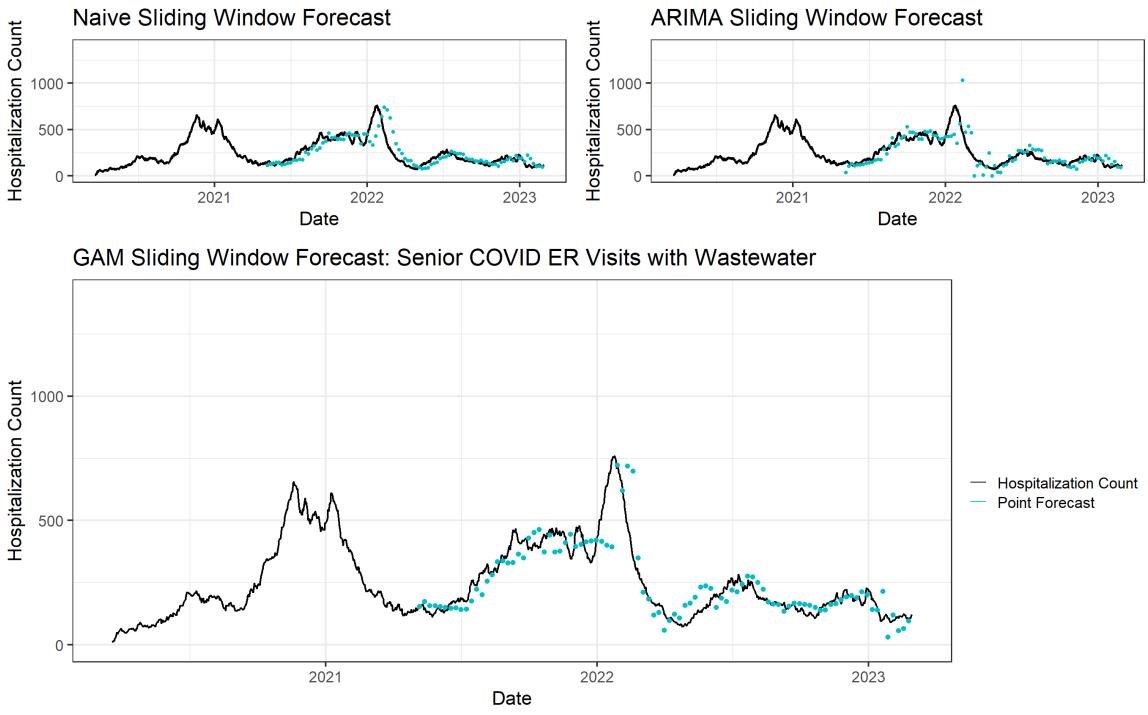


Figure 2.8: Comparison of Forecast Results.
The black lines represent are the hospital admissions, in teal are the forecast values. In the first panel we see the naive benchmark in teal, and in the second the ARIMA null model in teal. In the bottom we see the results of a forecast generated by Senior COVID ER with MGC. Along the x-axes we have the date, and the weekly hospitalization counts along the y-axes.

for all but the 7-day ex-post forecasts. In the ex-ante forecasts the naive does quite well. Inherently the naive tends to under-predict hospitalizations during increasing periods and over-predict them, but with that knowledge, one might be able to make informed considerations given the trends in hospitalizations.

The testing streams considered are not particularly predictive of hospitalizations. Among testing alone, the number of positive tests total and de-duplicated from the senior group is the most successful. The naive model outperforms most of the testing streams and is significantly more successful for the ex-ante forecasts than the ARIMA null model. In the 7-day ex-post predictions, the ARIMA model outperforms better than most of the testing streams.

In the ex-post univariate forecasts using syndromic data streams, percent of ER visits for COVID-19 and number of senior ER visits for COVID-19 tend to result in relatively small errors over the testing windows. The results involving

seniors might be because when the senior group is sick enough to visit the ER, they are likely to be subsequently hospitalized. This might also be because there are relatively few ER visits from the other groups. It is notable that the addition of test derived data streams, especially the senior specific streams, to the Senior ER % and the Senior ER visits does seem to improve accuracy and is relatively successful for the ex-ante forecasts as well.

The models which were most promising in the ex-post forecasts did not do as well when the covariates were projected into the future. This suggests that despite efforts against over-fitting, the relationships were not as strong as the results from the ex-post predictions suggested. It is also difficult to project these covariate values into the future with any certainty. Interestingly, but perhaps unsurprisingly, the complex models which were successful in the ex-post scenarios were significantly less successful in the ex-ante scenarios. Most of these data streams were highly correlated in time, this concurvity poses issues when relying upon multiple data streams for the forecasts and generates identifiability problems for more complicated models. In the approaches outlined above we chose not to use multiple versions of the same category of data stream in the same GAM forecasts, but even so, in the ex-ante forecasts the relative successes of all the forecasts were less consistent than those in the ex-post forecasts. This suggests that there is some kind of identifiability issue which is likely resulting in model misspecification. When multiple models with have similar forecast errors they can be differentiated by their AICs which indicate better model fit in the training data

The inclusion of wastewater metrics tends to improve model accuracy as well as model fit. The wastewater metrics alone are not particularly successful, but in tandem with other metrics they improve forecast accuracy. In the **Table 2.5**, we see that success metrics do tend to improve as increasingly complicated models have the addition of wastewater. Thus, wastewater data has the potential to be a valuable source for predicting future hospital admissions, but needs more refinement to be fully utilized.

Even without the inclusion of the wastewater measures, the forecast error tends to be smaller in the later time period than it is over the entire dataset. This may be because reporting mechanisms became more standardized as the

epidemic progressed. Population immunity also increased with progressively more of the population being vaccinated and experiencing natural exposure to the virus. Those getting tests later in the pandemic may be more likely to progress on to hospitalization, and seniors being tested even more so.

As can be seen in **Figure 2.8**, the error metrics increase over the periods of time where there are spikes in hospitalizations. This makes sense; changes in behavior are likely not captured and compound as individuals' perceptions of risk vary. This is why more mechanistic models may be valuable. The naive forecast has a MAE of 9.26 over the entire dataset, and 8.94 for the period where wastewater sampling was occurring regularly. This could correspond to as many as 64 more hospitalizations than predicted over the course of the forecast for the third week. Since it is absolute error, the likelihood of that is small, but in times where there are inflection points in the data, the implications of this error could be quite large. We see in the plots of error in time, that all the error metrics tend to increase during the spikes of the epidemic. Meanwhile, the GAM forecast using Senior ER visits for COVID-19 and wastewater SARS-CoV-2 MGC has a slightly daily forecast MAE of 7.25 which corresponds to an error of as many as 51 hospital admissions more or less than the forecast values.

2.5.1 Limitations

While this analysis proffers encouraging results across the success metrics, these streams have limitations. Wastewater is highly variable, and it would be prohibitively difficult to collect it daily. Systems go down, and the testing streams have needed to be adjusted to account for tests which were reported on a significant delay. One of the reasons that the 3-week forecast is so important is that public health will need to make forecasts using this delayed data. It is possible that other pathogens or variants could have differential age effects and the Senior ER visits will no longer be predictive of hospitalizations. The naive model is quite successful all things considered. The success metrics are all calculated as averages of averages, and thus smooth out much of the error occurring at inflection points.

2.6 Future Directions

We would like to extend this work by exploring the impacts of variants more comprehensively. There are relatively few hospitalizations at the county level, but the addition of more county-specific forecasts and spatial analyses might also improve the forecasts. More granular exploration could also contribute to our understanding of differences in transmission levels and case burden between urban and rural geographies. The UDHHS categorizes the wastewater trends by the number of sewersheds across the state with increasing concentrations of SARS-CoV-2 RNA. There is a wealth of data from counties and sewersheds, and it would be interesting to explore if the results for the CVWRF could be extended to other sewersheds. We projected wastewater MGC into the future using the lags rather than GAMs or ARIMA models as they are not smooth, so an exploration of other wastewater projections could reduce forecast error. The success of the lagged wastewater data suggests is an encouraging suggestion of the potential for wastewater data to aid forecasts. Perhaps some of the counties which lead others can be used to build a more robust forecast and validate the on-the-ground intuitions of hospital administrators and public health officials.

The models need to be tested specifically on inflection points to see not only which is most successful in a general sense, but also which is most successful at the points where hospitals are mobilizing and making carefully calculated staffing decisions. There is a wealth of data housed at UDHHS and the results presented in this chapter only begin to scratch the surface of what can be learned from this dataset. In order to more effectively assess the success of the forecasts, a new set of metrics which treat the over and underestimates generated by the model would allow for a greater understanding of its potential shortcomings. Similarly, the confidence interval coverage approach could be used to more effectively compare the errors across models and a penalty for overly large confidence intervals based on the ratio of confidence interval to forecast error could be designed to aid in model assessment and presentation.

Given all of the models considered in this chapter this work should be extended to include a more mechanistic approach. These forecasts have clear limitations, and perhaps a more mechanistic approach will illuminate the rela-

tionships between the covariates explored. Perhaps test positive rate could be explored further to develop estimates of case ascertainment rate which could be more informative. The differences between the prediction successes in the ex-post and ex-ante scenarios are a lesson in the relationship between causation and correlation. There are so many potential avenues and approaches with this data, and much to be learned from the relationships between these data streams.

2.7 Conclusions

This work will benefit the UDHHS in three ways: provide uncertainty quantification for the models currently being used, test other more complex models, and explore the transition between high testing and low testing environments. The methods developed can also be extended to other seasonally circulating respiratory viruses like Influenza and respiratory syncytial virus (**RSV**). In these results, we have succeeded in building forecasts which beat the ARIMA null model, and have shown that testing streams alone are not the best metric for forecasts of hospitalizations and often do not beat forecasts based on hospitalizations alone. Instead we can shift focus and use senior test data and syndromic surveillance in tandem with wastewater SARS-CoV-2 to improve our forecasts.

2.8 References

[1] *National Wastewater Surveillance System*, May 2023.

[2] J. Abbasi, *This Is Our COVID-What Physicians Need to Know About the Pediatric RSV Surge*, JAMA, 328 (2022), pp. 2096–2098.

[3] G. Adams, G. K. Moreno, B. A. Petros, R. Uddin, Z. Levine, B. Kotzen, K. S. Messer, S. T. Dobbins, K. C. DeRuff, C. M. Loreth, T. Brock-Fisher, S. F. Schaffner, S. Chaluvadi, S. Kanjilal, J. Luban, A. Ozonoff, D. J. Park, S. E. Turbett, K. J. Siddle, B. L. MacInnis, P. C. Sabeti, and J. E. Lemieux, *Viral Lineages in the 2022 RSV Surge in the United States*, New England Journal of Medicine, (2023). Publisher: Massachusetts Medical Society.

[4] F.-X. Ageron, O. Hugli, F. Dami, D. Caillet-Bois, V. Pittet, P. Eckert, N. Beysard, and P.-N. Carron, *Lessons from COVID-19 syndromic surveillance through emergency department activity: a prospective time series study from western Switzerland*, BMJ Open, 12 (2022), p. e054504. Publisher: British Medical Journal Publishing Group Section: Public health.

[5] H. Akaike, *On the Likelihood of a Time Series Model*, Journal of the Royal Statistical Society. Series D The Statistician, 27 (1978), pp. 217–235.

[6] D. Dowdy and G. DSouza, *COVID-19 Testing: Understanding the Percent Positive Johns Hopkins Bloomberg School of Public Health*, Aug. 2020.

[7] L. Fenga and M. Gaspari, *Predictive Capacity of COVID-19 Test Positivity Rate*, Sensors Basel, 21 (2021), p. 2435.

[8] Y. Furuse, Y. K. Ko, K. Ninomiya, M. Suzuki, and H. Oshitani, *Relationship of Test Positivity Rates with COVID-19 Epidemic Dynamics*, International Journal of Environmental Research and Public Health, 18 (2021), p. 4655. Number: 9 Publisher: Multidisciplinary Digital Publishing Institute.

[9] A. Galani, R. Aalizadeh, M. Kostakis, A. Markou, N. Alygizakis, T. Lytras, P. G. Adamopoulos, J. Peccia, D. C. Thompson, A. Kontou, A. Karagiannidis, E. S. Lianidou, M. Avgeris, D. Paraskevis, S. Tsiodras, A. Scorilas, V. Vasiliou, M.-A. Dimopoulos, and N. S. Thomaidis, *SARS-CoV-2 wastewater surveillance data can predict hospitalizations and ICU admissions*, Sci Total Environ, 804 (2022), p. 150151.

[10] G. H. Harris, K. J. Rak, J. M. Kahn, D. C. Angus, O. R. Mancing, J. Driessen, and D. J. Wallace, *US Hospital Capacity Managers Experiences and Concerns Regarding Preparedness for Seasonal Influenza and Influenza-like Illness*, JAMA Netw Open, 4 (2021), p. e212382.

[11] R. Hyndman, G. Athanasopoulos, C. Bergmeir, G. Caceres, L. Chhay, K. Kuroptev, M. OHara-Wild, F. Petropoulos, S. Razbash, E. Wang, and F. Yasmeen, *forecast: Forecasting Functions for Time Series and Linear Models*, 2023.

[12] R. J. Hyndman and Y. Khandakar, *Automatic Time Series Forecasting: The forecast Package for R*, Journal of Statistical Software, 27 (2008), pp. 1–22.

[13] R. J. Hyndman and A. B. Koehler, *Another look at measures of forecast accuracy*, International Journal of Forecasting, 22 (2006), pp. 679–688.

[14] R. Hyndmann, *Forecasting: Principles and Practice 3rd ed*, OText, 2023.

[15] N. Krivonakova, A. Soltysova, M. Tamas, Z. Takac, J. Krahulec, A. Ficek, M. Gal, M. Gall, M. Fehr, A. Krivjanska, I. Horakova, N. Belisova, P. Bimova, A. B. Skulcova, and T. Mackulak, *Mathematical modeling based on RT-qPCR analysis of SARS-CoV-2 in wastewater as a tool for epidemiology*, Sci Rep, 11 (2021), p. 19456. Number: 1 Publisher: Nature Publishing Group.

[16] J. Lu and S. Meyer, *Forecasting Flu Activity in the United States: Benchmarking an Endemic-Epidemic Beta Model*, Int J Environ Res Public Health, 17 (2020), p. 1381.

[17] K. D. Mandl, J. M. Overhage, M. M. Wagner, W. B. Lober, P. Sebastiani, F. Mostashari, J. A. Pavlin, P. H. Gesteland, T. Treadwell, E. Koski, L. Hutwagner, D. L. Buckeridge, R. D. Aller, and S. Grannis, *Implementing Syndromic Surveillance: A Practical Guide Informed by the Early Experience*, J Am Med Inform Assoc, 11 (2004), pp. 141–150.

[18] G. Marra and S. Wood, *Coverage properties of confidence intervals for generalized additive model components*, Scandinavian Journal of Statistics, 39 (2011), pp. 53–74.

[19] H. R. Meredith, E. Arehart, K. H. Grantz, A. Beams, T. Sheets, R. Nelson, Y. Zhang, R. G. Vinik, D. Barfuss, J. C. Pettit, K. McCaffrey, A. C. Dunn, M. Good, S. Frattaroli, M. H. Samore, J. Lessler, E. C. Lee, and L. T. Keegan, *Coordinated Strategy for a Model-Based Decision Support Tool for Coronavirus Disease, Utah, USA*, Emerg Infect Dis, 27 (2021), pp. 1259–1265.

[20] D. L. Miller, *Bayesian views of generalized additive modelling*, Oct. 2021.

[21] M. ODriscoll, G. Ribeiro Dos Santos, L. Wang, D. A. T. Cummings, A. S. Azman, J. Paireau, A. Fontanet, S. Cauchemez, and H. Salje, *Age-specific mortality and immunity patterns of SARS-CoV-2*, Nature, 590 (2021), pp. 140–145. Number: 7844 Publisher: Nature Publishing Group.

[22] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2022.

[23] UDEQ, *COVID-19 Wastewater Surveilance*.

[24] ——, *SARS-CoV-2 Sewage Monitoring*, July 2020.

[25] W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S*, Springer Science & Business Media, Sept. 2003.

[26] S. Venkatram, A. Alapati, A. Dileep, and G. Diaz-Fuentes, *Change in patterns of hospitalization for influenza during COVID-19 surges*, Influenza Other Respir Viruses, 16 (2022), pp. 72–78.

[27] S. I. Vrieze, *Model selection and psychological theory: A discussion of the differences between the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC)*, Psychological Methods, 17 (2012), pp. 228–243.

[28] S. Wood, *Generalized Additive Models: An Introduction with R*, CRC Press, United States, 2 ed., 2017.

[29] Y. Zhu, W. Oishi, C. Maruo, S. Bandara, M. Lin, M. Saito, M. Kitajima, and D. Sano, *COVID-19 case prediction via wastewater surveillance in a low-prevalence urban community: a modeling approach*, Journal of Water and Health, 20 (2022), pp. 459–470.

Chapter 3

Vaccination and Household Transmission

3.1 Abstract

Households are major transmission venues for infectious diseases, particularly respiratory illness.[8,12,16,21,26,28] Through this project we aim to understand the magnitude of COVID-19 vaccination's impact within households by examining four key parameters: vaccine efficacy, infectiousness of children and adults, household configuration (e.g., vaccination status and age of household members), and source of household infection importation. The goal of this chapter is to identify situations in which vaccination makes the biggest impact on household transmission, particularly to vulnerable groups. To answer these questions, we build a stochastic SIR model which contains compartments for vaccinated and non-vaccinated adults and children and simulates transmission within a household. We find that vaccination of children has minimal effect on secondary transmission within a household, but a large potential for preventing primary importation. This work provides a framework to evaluate how non-vaccinated household members impact within household transmission when data are limited, even when transmission parameters are unknown. This simulation-based experimentation can be used to explore optimal vaccine strategy given assumptions about vaccine efficacy and relative transmissibility gleaned from the literature.

3.2 Introduction

As we try to prevent the spread of SARS-CoV-2, we are faced with several challenges to preventing transmission and have the opportunity to intervene strategically to save lives. There has been much debate as to the best strategy for deploying vaccines to minimize transmission and prevent severe infection.

A complicating factor is the role of children in transmission within households. While children may be less likely to experience the severe effects of SARS-CoV-2, they have been shown to be a potential source of SARS-CoV-2 contagion.[37] Children who become infected may be asymptomatic vectors transmitting to other more vulnerable members of their households.[17] Children not only serve as potential disease vectors but are themselves a vulnerable population in need of protection. Despite lack of symptoms, immune dysregulation has been implicated in severe post-infectious cases,[11] and mounting evidence suggests that children are susceptible to long COVID and particularly lasting fatigue.[25] Understanding the role of transmission from children is critical to mitigating the spread of disease within households. In the case of the common cold, children provide a critical source of exposure for their parents[31] and it is likely that they are for SARS-CoV-2 as well, especially as it transitions from epidemic to endemic disease. In contrast to children, adults, and particularly adults with comorbidities, are more likely to experience severe disease and hospitalization.[7] Adults are likely to be exposed caring for sick children within their household. While some children have the option to stay home from school, a ten-day isolation period can place a high financial burden on working adults. It is estimated that the overall infection hospitalization rate for SARS-CoV-2 is 2.1% ranging from 0.4% for people below 40 to 9.7% for those above 60[20] and 75% of adults with long-haul COVID were not originally hospitalized for COVID-19,[11].[25]

Households provide unique arenas of transmission for their inhabitants. Households can be an inter-generational mixing ground where individuals of all ages are likely to intermingle and expose one another to disease.[30] Of particular focus, multi-generational households are a place where school age, working age, and senior individuals all might come into close contact with one

another and provide the opportunity for infectious disease transmission. At least 64 million Americans live in multi-generational households.[7]

It has been shown that there is high variability in transmission within households.[35] We examine the role of vaccinated and non-vaccinated individuals within heterogeneous households with a model which incorporates this potential for variability with the aim of informing targeted vaccination strategies optimized to have the biggest impact on transmission with the ultimate goal to reduce hospitalization and long-term symptoms. Vaccination provides an essential tool in the fight against the spread of disease. Vaccination is not a panacea, however; as novel strains emerge, our vaccines have become less effective.[2,14] Vaccines are prioritized for the most vulnerable,[5,19,22] and less likely to be quickly approved for children.[13] Even so, parents may be skeptical of giving their children access to new vaccines even when they themselves may be willing to receive the shot.[9,18] Even where vaccination is not fully effective in preventing infection, it still dramatically decreases the risks of hospitalization[10,24] and breakthrough transmission.[1,29,32] Understanding the impact of vaccination of children will provide a critically important role to preventing the spread of COVID-19 in households.

In this work, we use parameter estimates based off of the Utah Health and Economic Recovery Outreach (**HERO**) project data (described in 3.3.1) collected during the Delta wave to inform a simulated model of household SARS-CoV-2 transmission dynamics with compartments for adults and children, and heterogeneity of vaccine deployment. As new variants have emerged, the scientific community has investigated the relative infectiousness,[3] symptom profile,[34] severity of emerging strains,[36] as well as relative vaccine efficacy against those strains.[2,6,14] This model is flexible and can be updated as within-household transmission estimates improve.

3.3 Methods

We use a household model describing viral shedding dynamics to estimate the impact of vaccinating various household members on within-household SARS-CoV-2 transmission and quantify the significance of secondary transmis-

sion within households. This model facilitates the prospective assessment of transmission dynamics and quantification of the impacts of vaccination early after the emergence of a novel variant when data are limited, and behaviors are changing rapidly. The model is generalizable and based on four key parameters which can be modified to match the dynamics of future variants. This analysis was conducted in R version 4.2.2.

3.3.1 Data

This project is informed by data collected by the Utah HERO project,[27] which was designed to assess the average community seroprevalence of SARS-CoV-2 using a statistically representative population of households from several counties in Utah. All members of selected households were invited to participate in a survey and serological SARS-CoV-2 antibody testing. The survey which included a question on self-reporting prior RT-PCR test results for SARS-CoV-2, was completed by members of 3381 Utah households. Data included in this analysis were collected from January to April 2021 during the Delta wave.

In the HERO project, survey participation was opt-in and not all household members were required to participate for a household to be included, so not all households included surveys for all members. For this analysis, we chose to include only households with $\geq$75% survey coverage because mean age, total reporting COVID-19 diagnoses, and total non-reported COVID-19 cases identified by antibody tests after the fact were all dependent on survey responses. To ensure that the data set of complete households was representative, we compared the summary statistics across all households (**Table S3.1**). For households included in this study (i.e., those with $\geq$75% survey coverage), **Table 3.1** shows the mean age, household size, total infected, and total missed SARS-CoV-2 cases for all households, across households with and without children.

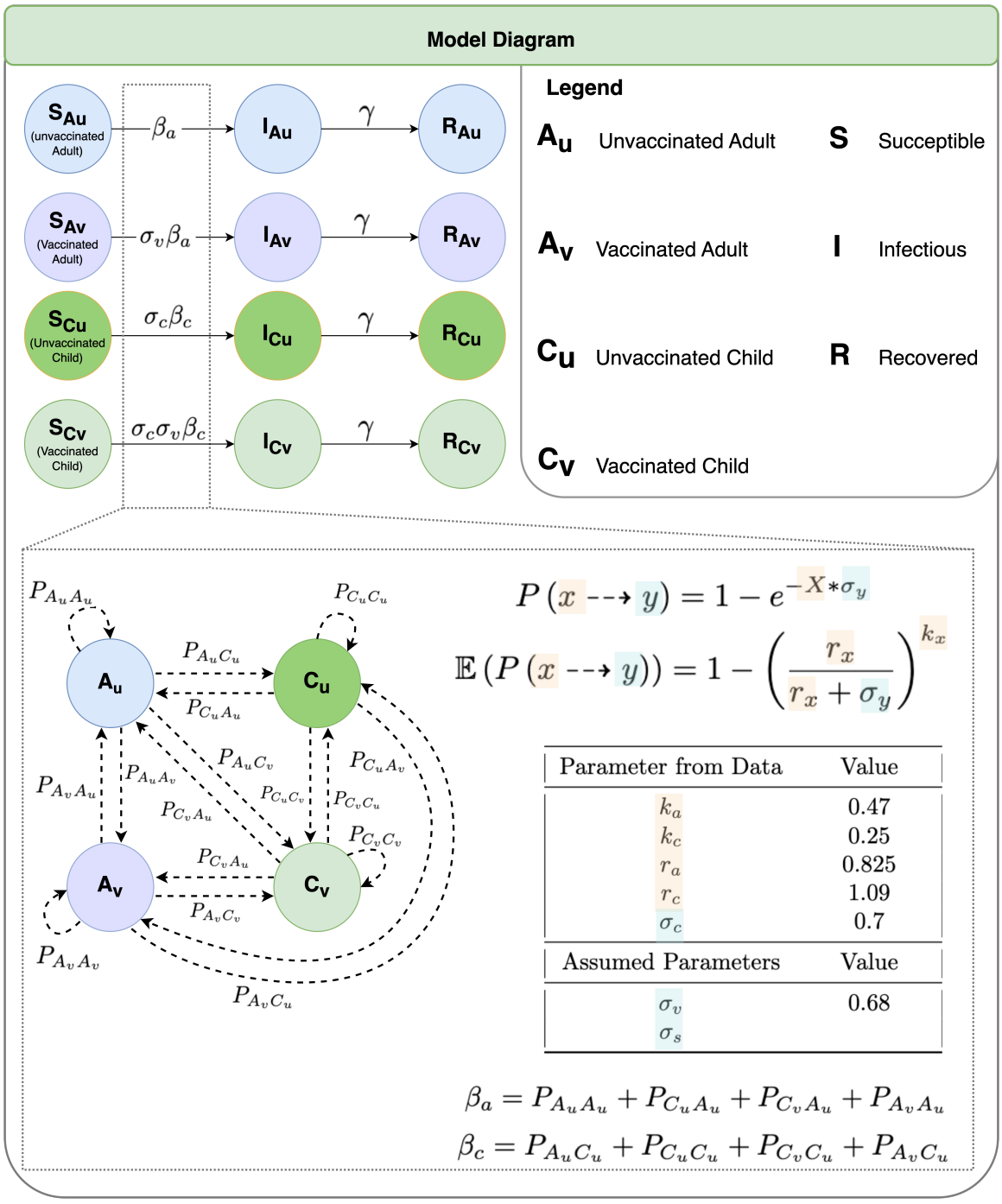
Some individuals included in our study tested positive for SARS-CoV-2 antibodies, however they do not self-report previous infection, rather, some infections were first identified by antibody test. For the purposes of this study, we define a missed infection as any infection first detected with an antibody test rather than a RT-PCR test. Consequently, individuals with missed infections were not identified while the infection was ongoing, but rather after the fact.

Table 3.1: Summary Statistics for Households with and without children with ≥75% survey coverage.

Variable	Mean	SD
All Households	(n = 3278)	
Age	43.19	18.65
Household Size	2.50	1.57
Total Sick	0.31	0.81
Total Missed Cases	0.14	0.50
Households with Children	(n = 809)	
Age	25.67	7.34
Household Size	4.47	1.39
Total Sick	0.65	1.21
Total Missed Cases	0.32	0.79
Households without Children	(n = 2469)	
Age	50.01	17.22
Household Size	1.73	0.75
Total Sick	0.19	0.54
Total Missed Cases	0.07	0.29

3.3.2 Model

We built an agent-based SIR model of within-household transmission among a household of four individuals in R and used it to run forward simulations to quantify the impact of vaccinating children on overall household infection risk. Beyond the susceptible, infected, recovered disease transmission compartments, our model contains compartments for vaccinated adults, non-vaccinated adults, vaccinated children, and non-vaccinated children. The full details of the model are shown in **Figure 3.1**. In our model, the probability of transmission between a contact is $p = 1 - e^{-\sigma X}$, where X is a gamma distributed random variable with shape k and rate r determined by the age [adult/child] of the infected individual and σ, a proxy for susceptibility, determined by the age and vaccination status of the contact. The expected transmission probability is $p = 1 - (\frac{r}{(r+\sigma)})^k$. We assume that vaccination does not change the behavior of onward transmission,



Parameter from Data	Value
k_a	0.47
k_c	0.25
r_a	0.825
r_c	1.09
σ_c	0.7
Assumed Parameters	**Value**
σ_v	0.68
σ_s	

Figure 3.1: Model Diagram
In this figure we see the general SIR format for each age compartment of the model at the top. In the lower section we see all of the interactions between individuals and routes for infection within and across groups. On the bottom left, we see the direct probability of transmission between groups and a table containing the transmission probabilities between groups for each of the scenarios. The equations at the bottom right indicate how all of the transmission probabilities are summed to get an overall β for each age category.

and only reduces a contact's susceptibility. This assumption and the calculation for the transmission probability is discussed in detail in the following chapter.

To simulate transmission within a household, we first seed the household with one infected adult or child. Then, for that infected individual we generate a transmission probability by drawing from from a gamma distribution parameterized by a rate and shape parameters which are determined by the infected individual and then moderated by the susceptible individual's relative susceptibility, σ and draw from a random binomial process to determine if transmission has occurred. The infected individual then transitions to the recovered class. If any individuals are infected, the cycle repeats for another generation and the direct transmission probabilities from them to each susceptible housemate are calculated and another random binomial draw determines if the transmission occurs. Since we simulate each time step as a generation, we move all individuals who were infected at the start of the time step to the recovered class and we repeat these steps until we run out of susceptible or infected individuals in the household.

3.3.3 Parameter Selection

In a similar study Toth et al, [in prep] estimated key transmission parameters for non-vaccinated adults and children. Here, we extend their model to include vaccinated children and adults. The full methods are described in Toth et al. [in prep]; but briefly, they fit a probabilistic model of household importation and transmission to the data from the HERO study. They calculated a maximum likelihood estimate of the importation probability, mean and variability of household transmission probability, and sensitivity and specificity of test data. They estimated transmission probabilities between non-vaccinated adults and non-vaccinated children. Based on the parameter estimates and household age groupings in the study, they categorized the populations into two scenarios, Scenario 1 where children aged 0-12 lived with adults aged 25-44, and Scenario 2 where children aged 13-24 lived with adults aged 45 and over. They then estimated between-group and within-group secondary attack rates to parameterize transmission rates in our forward simulation model (described 3.3.4). A list of these extracted parameters can be found in **Table 3.2**.

Table 3.2: Secondary Attack Rates for the Demographic Scenarios. Households are split into two scenarios; children aged 0-12 with adults aged 25-44 on the left, and children aged 13-24 with adults aged 45 and over on the right. The from column indicates the age category of the infected individual and the to columns indicate the age category of the individual infected.

Scenario 1			Scenario 2		
From	To		From	To	
	0-12	25-44		13-24	45+
0-12	15.7%	29.7%	13-24	2.5%	15.2%
25-44	29.7%	29.7%	45+	15.2%	56.2%

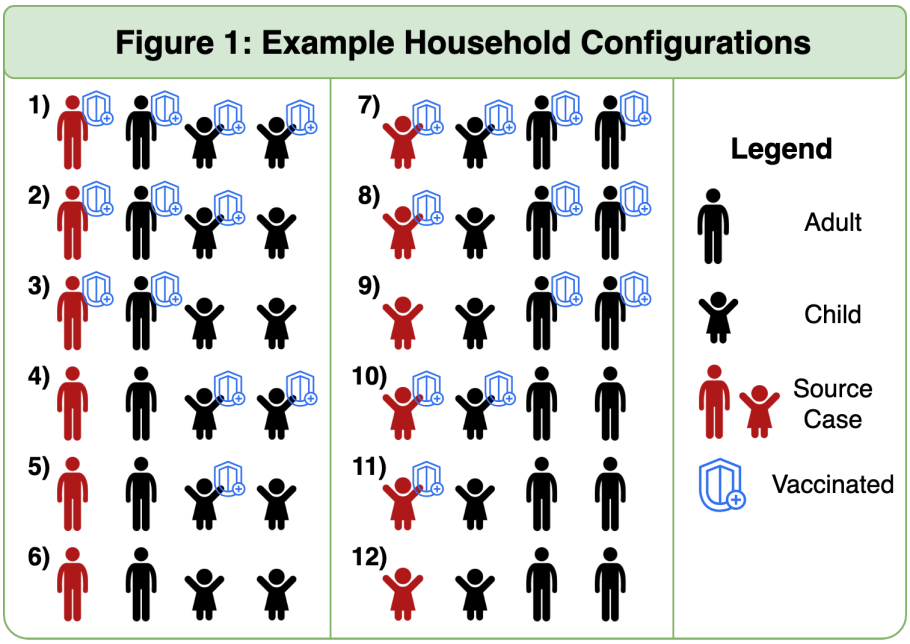


Figure 3.2: Household Configurations
A diagram of all the possible household configurations. The infected importer is shown in red, and the other household members are in black. Larger figures represent adults while smaller figures represent children. The shields represent vaccinated individuals.

3.3.4 Forward Simulation

We simulate replicates of 10,000 households for each scenario and each household configuration. In each simulation, all households have the same configuration of vaccinated and non-vaccinated school age and working age individuals. All of the household configurations explored are shown in **Figure 3.2**. A house-

hold configuration is comprised of a chosen number of non-vaccinated or vaccinated school age individuals, either 0-12 or 13-24 years old; and adults, either 25-44 or 45+ years old. For each set of replicates, we ran trials of all twelve household configurations; six with an infected adult importer, and six with a infected child importer. When there is an adult importer, we assume both adults are either non-vaccinated (Configuration 1-3) or vaccinated (Configuration 4-6) with no, one, or both children vaccinated. When the importation stems from a child, we assume both children are either non-vaccinated (Configuration 7-9) or vaccinated (Configuration 10-12) with no, one, or both children vaccinated. This is because in this model we assume that vaccination only impacts susceptibility and does nothing to moderate forward transmission. We then compare the number of infected vaccinated and non-vaccinated adults and children for each of the household configurations.

3.4 Results

3.4.1 Parameter Effects

In order to understand how the parameters interact with one another in a complex parameter space and impact infections in adults, we perform Latin Hypercube Sampling (**LHS**). We vary vaccine efficacy between 0 and 100%. Typical rates of SARS-CoV-2 vaccine efficacy estimated to be between 40 and 90%.[23,29,33] We vary child susceptibility to infection between 0 and 100% of that of adults. Here we assume that adult infectiousness is uniformly distributed between 0.135 and 4.3 (estimate 0.825), and child infectiousness is uniformly distributed between 0.18 and 2.08 (estimate 1.09). These parameter ranges cover the extremes of both scenarios.

In **Figure 3.3** we see the impacts of these parameters on distributions of adult infections when there is an infected adult importer. Here the parameters related to adult infectiousness R_a and K_a are the parameters which have the largest impact on total adult infections. There is relatively little difference between the trials with vaccinated or non-vaccinated children, but the sigmaV, vaccine efficacy still has a strong effect based on the vaccination status of the adults. This suggests that a majority of the transmissions to adults are coming from

adults, and that vaccination of children has little effect once the infection is in the household.

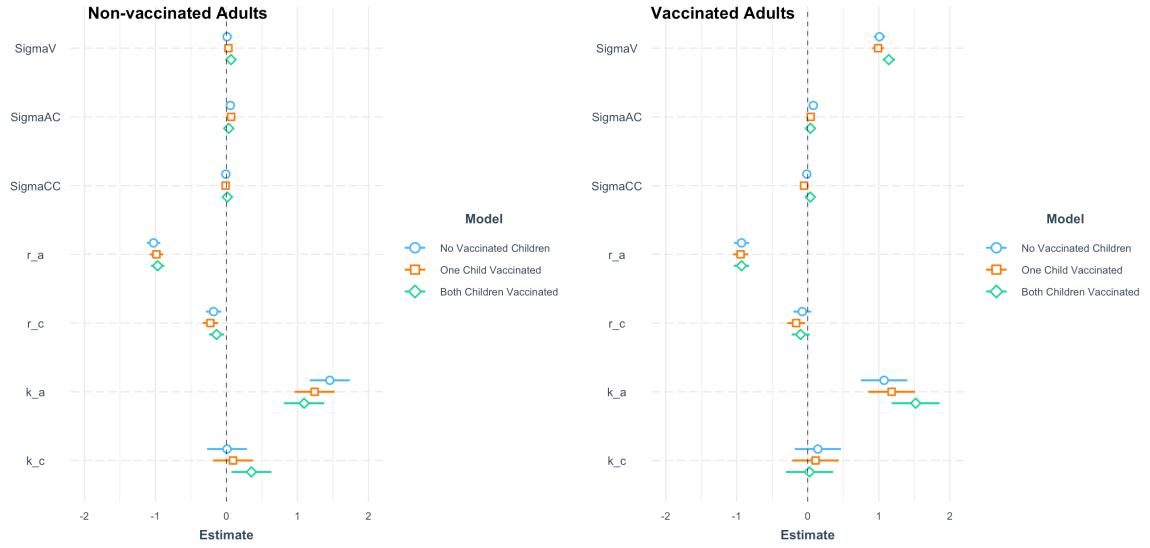


Figure 3.3: Sensitivity Analysis for Working-aged Importers. These tables show the effects of vaccinating progressively more school-aged individuals in the household with an adult importation. The blue dots represent the trials with no vaccinated school-aged individuals, the orange dots with one vaccinated school-aged individual, and the green dots with both school-aged individuals vaccinated. On the y-axis, sigmaV is the susceptibility moderated by vaccination, sigmaAC is the susceptibility moderated by transmission from adults to children, sigmaCC is the susceptibility between children. R_a is the rate parameter for the infectiousness of adults. R_c is the rate parameter for the infectiousness of school-aged individuals. K_a is the shape parameter for the infectiousness of adults. K_c is the shape parameter for the infectiousness of school-aged.

We see the sensitivity of parameters for the cases where the infected importer is a child in **Figure 3.4**. Here the parameters related to child infectiousness R_c and K_c are the parameters which have the largest impact on total adult infections. This suggests that the infectiousness of the importer in general has a large impact on the number of adult infections in the household. When the children are vaccinated, vaccine efficacy increases in impact as more of the adults in the household are vaccinated and is not significant when there are no vaccinated adults.

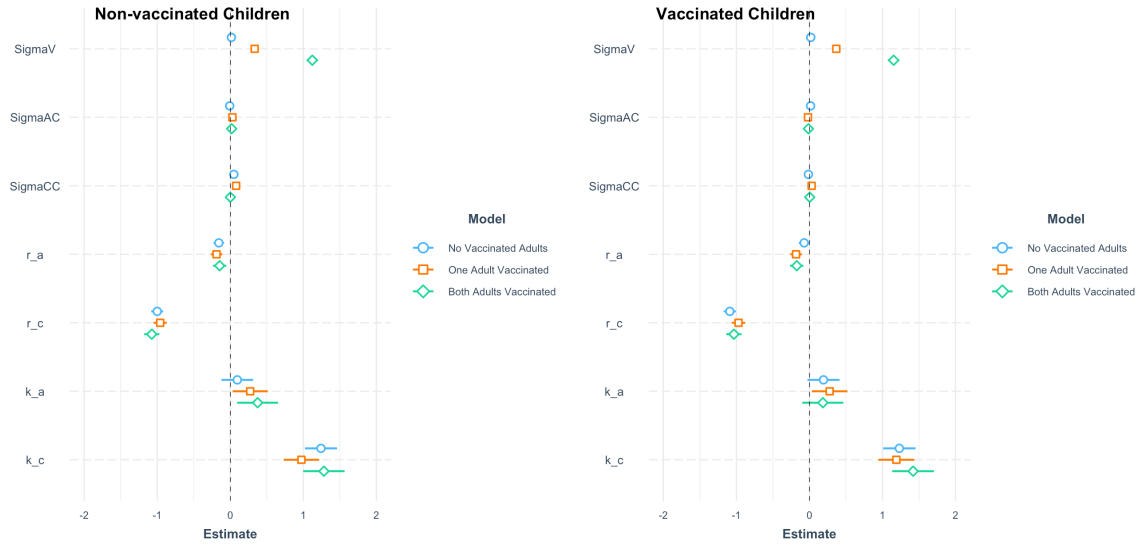


Figure 3.4: Sensitivity Analysis for School-aged Importers.
These tables show the effects of vaccinating progressively more working-age individuals in the household with an school-aged importation. The blue dots represent the trials with no vaccinated school-aged individuals, the orange dots with one vaccinated school-aged individual, and the green dots with both school-aged individuals vaccinated. On the y-axis, sigmaV is the susceptibility moderated by vaccination, sigmaAC is the susceptibility moderated by transmission from adults to children, sigmaCC is the susceptibility between children. R_a is the rate parameter for the infectiousness of adults. R_c is the rate parameter for the infectiousness of school-aged individuals. K_a is the shape parameter for the infectiousness of adults. K_c is the shape parameter for the infectiousness of school-aged.

3.4.2 Distributions of Adult Infections

We calculate final distributions of adult infection from both scenarios and various household configurations to assess the impact of vaccination on adult infections. The distribution of adult infections from Scenario 1 is presented in **Figure 3.5** where we see that in the case of an adult importer, the number of adults infected decreases slightly as the number of vaccinated children increases. There is a much more significant decrease in the number of adults infected when children are the importers and additional adults are infected. This is likely because in this parameter regimen, children are less infectious than their adult counterparts. In the bottom two panels, there is an almost imperceptible difference between the two purple columns. There is a much more significant

difference between number adult infections in the top row, suggesting that vaccination of adults has a much greater impact than the impact of secondary transmissions from children. The distribution of adult infections from Scenario

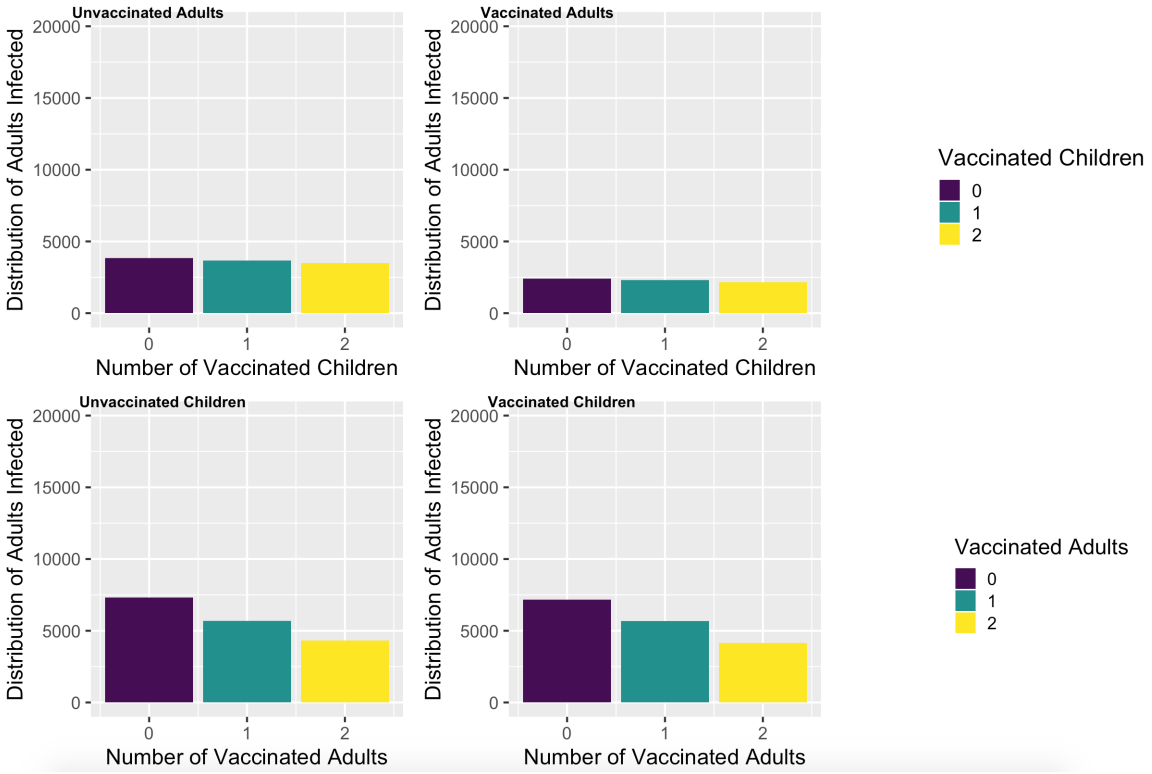


Figure 3.5: Distributions of Adult Infections from Scenario 1
Each plot contains three trials of 10,000 simulations of households with individuals aged 0-12 and 25-44. The purple bars represent the trials where the non-importer group was not vaccinated, the teal are the trials where one of the non-importer group was vaccinated, and the yellow are where both of the non-importer group were vaccinated. Along the top row are the household configurations where adults are the importer, and along the bottom row are the household configurations where the importation comes from a child. On the left hand side are the configurations where the infected importer is not vaccinated and on the right are the configurations where they are. The axes are limited from 0-20,000 for ease of comparison, but in the case where the adults are the importer there are only 10,000 adults available to be infected.

2 is presented in **Figure 3.6** where we see that in the case of an adult importer, the number of adults does not change as the number of vaccinated children increases. There is a small decrease in the number of adults infected when children are the importers and additional adults are infected. This is likely

because in this parameter regimen, children are even less infectious than their adult counterparts. The difference from left to right demonstrates the impact of vaccination in the importer age group. Here we can see that when there is a child importer vaccination of the other child does have have a slight impact on the adult infections.

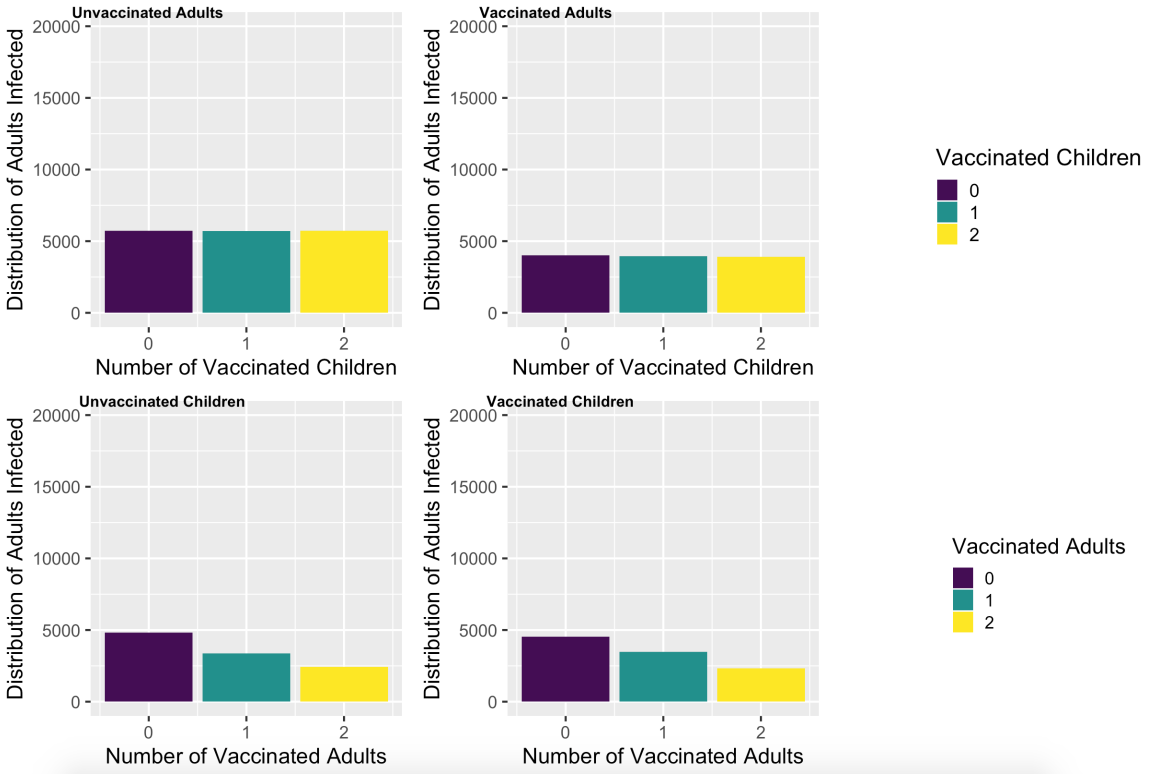


Figure 3.6: Distributions of Adult Infections from Scenario 2
Each plot contains three trials of 10,000 simulations of households with individuals aged 13-24 and 45+. The purple bars represent the trials where the non-importer group was not vaccinated, the teal are the trials where one of the non-importer group was vaccinated, and the yellow are where both of the non-importer group were vaccinated. Along the top row are the household configurations where adults are the importer, and along the bottom row are the household configurations where the importation comes from a child. On the left hand side are the configurations where the infected importer is not vaccinated and on the right are the configurations where they are. The axes are limited from 0-20,000 for ease of comparison, but in the case where the adults are the importer there are only 10,000 adults available to be infected.

3.5 Discussion

Transmission is much more likely within households than outside of them,[15,35] but for infection to enter the home it must be imported by a household member. Since children are estimated to be less infectious than adults, vaccination of children does little to decrease secondary transmission within the household. The benefit of vaccinating children comes primarily from decreases in potential importation of infection into the household and mitigated infection in children themselves.

We know they without vaccination and even with vaccination children still can get sick,[4] and the more we learn about COVID-19, the more we learn about its lingering effects.[25] Even if children are unlikely to experience severe infection or hospitalization, it would still be best to avoid infection and the potential for subsequent transmission and long-haul covid. On the other hand, vaccinating adults benefits both adults and children. This result primarily stems from the greater infectiousness of adults and is susceptible to heterogeneity based on immune response, variant virility, and vaccine efficacy.

While this model requires carefully calculated parameter estimates to be utilized, but the results were quite robust across scenarios which suggests that in the case of heterogeneous infectiousness, it is likely best to optimize around preventing infections in those who are most likely to transmit the infection and those who are most vulnerable to severe infection. This concept is commonly suggested in the cases of healthcare workers and the elderly, but working-age adults and their community and workplace exposures are less frequently considered. Vaccination is particularly important amongst adults to decrease the likelihood of severe infection and hospitalization since it is more likely for that group. Vaccinating the adults in the household also diminishes the likelihood of community spread. This result is unlikely to hold for other respiratory infections where children may have similar or even higher infectiousness and susceptibility than adults.

3.6 Conclusion and Future Directions

With this project we quantify the effects of vaccination of children on adult infections. Those effects appear to be statistically significant, but without epidemiological impact under current parameter regimes. This approach allows for experimentation to find optimal vaccination strategies within households. A limitation of this work is the impact of vaccination on importations of infection into the household. Future directions will include the quantification of the effects of importation as that is where most of the benefit of vaccination are likely to be seen.

Luckily, in Utah, where these parameters were estimated there is an adequate supply of vaccine, and we no longer have to optimize vaccine distribution. This work still has significance for other regions of the world where vaccine access is less robust and prioritization decisions need to be made. This work can also be extended with coarse estimates for SARs as new strains and novel infectious agents change transmission dynamics and relative susceptibilities of adults and children shift and with more precise SARs which are regularly estimated for influenza. Here we quantify the effects of vaccination on secondary transmissions within households and find that vaccination of children does not have large impacts on secondary transmissions. This model proffers an adaptable framework for evaluating vaccine prioritization in households.

3.7 References

[1] E. Amanatidou, A. Gkiouliava, E. Pella, M. Serafidi, D. Tsilingiris, N. G. Vallianou, I. Karampela, and M. Dalamaga, *Breakthrough infections after COVID-19 vaccination: Insights, perspectives and challenges*, Metabol Open, 14 (2022), p. 100180.

[2] N. Andrews, J. Stowe, F. Kirsebom, S. Toffa, T. Rickeard, E. Gallagher, C. Gower, M. Kall, N. Groves, A.-M. OConnell, D. Simons, P. B. Blomquist, A. Zaidi, S. Nash, N. Iwani Binti Abdul Aziz, S. Thelwall, G. Dabrera, R. Myers, G. Amirthalingam, S. Gharbia, J. C. Barrett, R. Elson, S. N. Ladhani, N. Ferguson, M. Zambon, C. N. Campbell, K. Brown, S. Hopkins, M. Chand, M. Ramsay, and J. Lopez Bernal, *Covid-*

19 Vaccine Effectiveness against the Omicron B.1.1.529 Variant, New England Journal of Medicine, 386 (2022), pp. 1532–1546. Publisher: Massachusetts Medical Society.

[3] J. M. Baker, *SARS-CoV-2 B.1.1.529 Omicron Variant Transmission Within Households - Four U.S. Jurisdictions, November 2021-February 2022*, MMWR Morb Mortal Wkly Rep, 71 (2022).

[4] R. K. Borchering, L. C. Mullany, E. Howerton, M. Chinazzi, C. P. Smith, M. Qin, N. G. Reich, L. Contamin, J. Levander, J. Kerr, J. Espino, H. Hochheiser, K. Lovett, M. Kinsey, K. Tallaksen, S. Wilson, L. Shin, J. C. Lemaitre, J. D. Hulse, J. Kaminsky, E. C. Lee, A. L. Hill, J. T. Davis, K. Mu, X. Xiong, A. Pastore y Piontti, A. Vespignani, A. Srivastava, P. Porebski, S. Venkatramanan, A. Adiga, B. Lewis, B. Klahn, J. Outten, B. Hurt, J. Chen, H. Mortveit, A. Wilson, M. Marathe, S. Hoops, P. Bhattacharya, D. Machi, S. Chen, R. Paul, D. Janies, J.-C. Thill, M. Galanti, T. Yamana, S. Pei, J. Shaman, G. Espana, S. Cavany, S. Moore, A. Perkins, J. M. Healy, R. B. Slayton, M. A. Johansson, M. Biggerstaff, K. Shea, S. A. Truelove, M. C. Runge, C. Viboud, and J. Lessler, *Impact of SARS-CoV-2 vaccination of children ages 5-11 years on COVID-19 disease burden and resilience to new variants in the United States, November 2021-March 2022: a multi-model study*, The Lancet Regional Health - Americas, 17 (2023), p. 100398.

[5] K. M. Bubar, K. Reinholt, S. M. Kissler, M. Lipsitch, S. Cobey, Y. H. Grad, and D. B. Larremore, *Model-informed COVID-19 vaccine prioritization strategies by age and serostatus*, Science, 371 (2021), pp. 916–921. Publisher: American Association for the Advancement of Science.

[6] L. Cao, J. Lou, S. Y. Chan, H. Zheng, C. Liu, S. Zhao, Q. Li, C. K. P. Mok, R. W. Y. Chan, M. K. C. Chong, W. K. K. Wu, Z. Chen, E. L. Y. Wong, P. K. S. Chan, B. C. Y. Zee, E. K. Yeoh, and M. H. Wang, *Rapid evaluation of COVID-19 vaccine effectiveness against symptomatic infection with SARS-CoV-2 variants by analysis of genetic distance*, Nat Med, 28 (2022), pp. 1715–1722. Number: 8 Publisher: Nature Publishing Group.

[7] D. Cohn and J. S. Passel, *A record 64 million Americans live in multigenerational households*.

[8] A. Emanuels, J. Heimonen, J. OHanlon, A. E. Kim, N. Wilcox, D. J. McCulloch, E. Brandstetter, C. R. Wolf, J. K. Logue, P. D. Han, B. Pfau, K. L. Newman, J. P. Hughes, M. L. Jackson, T. M. Uyeki, M. Boeckh, L. M. Starita, D. A. Nickerson, T. Bedford, J. A. Englund, and H. Y. Chu, *Remote Household Observation for Noninfluenza Respiratory Viral Illness*, Clin Infect Dis, 73 (2020), pp. e4411–e4418.

[9] J. Grose, *Theyre Not Anti-Vaccine, but These Parents Are Hesitant About the Covid Shot*, The New York Times, (2021).

[10] F. P. Havers, H. Pham, C. A. Taylor, M. Whitaker, K. Patel, O. Anglin, A. K. Kambhampati, J. Milucky, E. Zell, H. L. Moline, S. J. Chai, P. D. Kirley, N. B. Alden, I. Armistead, K. Yousey-Hindes, J. Meek, K. P. Openo, E. J. Anderson, L. Reeg, A. Kohrman, R. Lynfield, K. Como-Sabetti, E. M. Davis, C. Cline, A. Muse, G. Barney, S. Bushey, C. B. Felsen, L. M. Billing, E. Shiltz, M. Sutton, N. Abdullah, H. K. Talbot, W. Schaffner, M. Hill, A. George, A. J. Hall, S. R. Bialek, N. C. Murthy, B. P. Murthy, and M. McMorrow, *COVID-19-Associated Hospitalizations Among Vaccinated and Unvaccinated Adults 18 Years or Older in 13 US States, January 2021 to April 2022*, JAMA Intern Med, 182 (2022), pp. 1071–1081.

[11] F. Health, *A detailed study of patients with long-haul COVID: an analysis of private healthcare claims - Digital Collections - National Library of Medicine*.

[12] Y. Krishnamoorthy, K. Ezhumalai, S. Murali, S. Rajaa, M. Jose, A. Sathishkumar, G. Soundappan, C. Horsburgh, N. Hochberg, W. E. Johnson, S. Knudsen, P. Salgame, J. Ellner, S. Prakash Babu, and S. Sarkar, *Prevalence and risk factors associated with latent tuberculosis infection among household contacts of smear positive pulmonary tuberculosis patients in South India*, Tropical Medicine & International Health, 26 (2021), pp. 1645–1651.

[13] N. Kumar and S. Susan, *COVID-19 Pandemic Prediction using Time Series Forecasting Models*, in 2020 11th International Conference on Computing, Communication and Networking Technologies ICCCNT, July 2020, pp. 1–7.

[14] R. Link-Gelles, *Early Estimates of Bivalent mRNA Booster Dose Vaccine Effectiveness in Preventing Symptomatic SARS-CoV-2 Infection Attributable to Omicron BA.5- and XBB/XBB.1.5-Related Sublineages Among Immunocompetent Adults - Increasing*

Community Access to Testing Program, United States, December 2022-January 2023, MMWR Morb Mortal Wkly Rep, 72 (2023).

[15] I. M. Longini and J. S. Koopman, *Household and community transmission parameters from final distributions of infections in households*, Biometrics, 38 (1982), pp. 115–126.

[16] I. M. Longini, J. S. Koopman, A. S. Monto, and J. P. Fox, *Estimating household and community transmission parameters for influenza*, Am J Epidemiol, 115 (1982), pp. 736–751.

[17] J. F. Ludvigsson, *Children are unlikely to be the main drivers of the COVID-19 pandemic - A systematic review*, Acta Paediatrica, 109 (2020), pp. 1525–1530.

[18] S. Mallapaty, *COVID jabs for kids: theyre safe and they work - so why is uptake so patchy*, Nature, 610 (2022), pp. 246–248.

[19] S. A. Mbaeyi, *COVID-19 vaccine prioritization: Work Group considerations*.

[20] N. Menachemi, B. E. Dixon, K. K. Wools-Kaloustian, C. T. Yiannoutsos, and P. K. Halverson, *How Many SARS-CoV-2-Infected People Require Hospitalization Using Random Sample Testing to Better Inform Preparedness Efforts*, J Public Health Manag Pract, 27 (2021), pp. 246–250.

[21] P. K. Munywoki, D. C. Koech, C. N. Agoti, C. Lewa, P. A. Cane, G. F. Medley, and D. J. Nokes, *The Source of Respiratory Syncytial Virus Infection In Infants: A Household Cohort Study In Rural Kenya*, J Infect Dis, 209 (2014), pp. 1685–1692.

[22] S. Munzert, S. Ramirez-Ruiz, B. Cali, L. F. Stoetzer, A. Gohdes, and W. Lowe, *Prioritization preferences for COVID-19 vaccination are consistent across five countries*, Humanit Soc Sci Commun, 9 (2022), pp. 1–10. Number: 1 Publisher: Palgrave.

[23] T. Pilishvili, R. Gierke, K. E. Fleming-Dutra, J. L. Farrar, N. M. Mohr, D. A. Talan, A. Krishnadasan, K. K. Harland, H. A. Smithline, P. C. Hou, L. C. Lee, S. C. Lim, G. J. Moran, E. Krebs, M. T. Steele, D. G. Beiser, B. Faine, J. P. Haran, U. Nandi, W. A. Schrading, B. Chinnock, D. J. Henning, F. Lovecchio, J. Lee, D. Barter, M. Brackney, S. K. Fridkin, K. Marceaux-Galli, S. Lim, E. C. Phipps, G. Dumyati, R. Pierce, T. M. Markus, D. J. Anderson, A. K. Debes, M. Y. Lin,

J. Mayer, J. H. Kwon, N. Safdar, M. Fischer, R. Singleton, N. Chea, S. S. Magill, J. R. Verani, and S. J. Schrag, *Effectiveness of mRNA Covid-19 Vaccine among U.S. Health Care Personnel*, N Engl J Med, 385 (2021), p. e90. Publisher: Massachusetts Medical Society.

[24] K. Rahmani, R. Shavaleh, M. Forouhi, H. F. Disfani, M. Kamandi, R. K. Oskooi, M. Foogerdi, M. Soltani, M. Rahchamani, M. Mohaddespour, and M. Dianatinasab, *The effectiveness of COVID-19 vaccines in reducing the incidence, hospitalization, and mortality from COVID-19: A systematic review and meta-analysis*, Front Public Health, 10 (2022), p. 873596.

[25] M. Roessler, F. Tesch, M. Batram, J. Jacob, F. Loser, O. Weidinger, D. Wende, A. Vivirito, N. Toepfner, F. Ehm, M. Seifert, O. Nagel, C. Konig, R. Jucknewitz, J. P. Armann, R. Berner, M. Treskova-Schwarzbach, D. Hertle, S. Scholz, S. Stern, P. Ballesteros, S. Babler, B. Bertele, U. Repschlager, N. Richter, C. Riederer, F. Sobik, A. Schramm, C. Schulte, L. Wieler, J. Walker, C. Scheidt-Nave, and J. Schmitt, *Post-COVID-19-associated morbidity in children, adolescents, and adults: A matched cohort study including more than 157,000 individuals with COVID-19 in Germany*, PLOS Medicine, 19 (2022), p. e1004122. Publisher: Public Library of Science.

[26] M. A. Rolfes, H. K. Talbot, H. Q. McLean, M. S. Stockwell, K. D. Ellingson, K. Lutrick, N. M. Bowman, E. E. Bendall, A. Bullock, J. D. Chappell, J. E. Deyoe, J. Gilbert, N. B. Halasa, K. E. Hart, S. Johnson, A. Kim, A. S. Lauring, J. T. Lin, C. J. Lindsell, S. H. McLaren, J. K. Meece, A. M. Mellis, M. Moreno Zivanovich, C. E. Ogokeh, M. Rodriguez, E. Sano, R. A. Silverio Francisco, J. E. Schmitz, C. Y. Vargas, A. Yang, Y. Zhu, E. A. Belongia, C. Reed, and C. G. Grijalva, *Household Transmission of Influenza A Viruses in 2021-2022*, JAMA, 329 (2023), pp. 482–489.

[27] M. H. Samore, A. Looney, B. Orleans, T. Greene, N. Seegert, J. C. Delgado, A. Presson, C. Zhang, J. Ying, Y. Zhang, J. Shen, P. Slev, M. Gaulin, M.-J. Yang, A. T. Pavia, and S. C. Alder, *SARS-CoV-2 seroprevalence and detection fraction in Utah urban populations from a probability-based sample*, Oct. 2020. Pages: 2020.10.26.20219907.

[28] E. M. Scott, A. Magaret, J. Kuypers, J. M. Tielsch, J. Katz, S. K. Khatry, L. Stewart, L. Shrestha, S. C. LeClerq, J. A. Englund, and H. Y. Chu, *Risk factors and patterns of household clusters of respiratory viruses in rural Nepal*, Epidemiol Infect, 147 (2019), p. e288.

[29] M. C. Shamier, A. Tostmann, S. Bogers, J. d. Wilde, J. IJpelaar, W. A. v. d. Kleij, H. d. Jager, B. L. Haagmans, R. Molenkamp, B. B. O. Munnink, C. v. Rossum, J. Rahamat-Langendoen, N. v. d. Geest, C. P. Bleeker-Rovers, H. Wertheim, M. P. G. Koopmans, and C. H. GeurtsvanKessel, *Virological characteristics of SARS-CoV-2 vaccine breakthrough infections in health care workers*, Aug. 2021. Pages: 2021.08.20.21262158.

[30] A. Simpson, P. Rebala, T. Johnston, and S. Ferriss, *One home, many generations: States addressing COVID risk among families*, Mar. 2021.

[31] R. S. Sneed, S. Cohen, R. B. Turner, and W. J. Doyle, *Parenthood and Host Resistance to the Common Cold*, Psychosomatic Medicine, 74 (2012), pp. 567–573.

[32] S. T. Tan, A. T. Kwan, I. Rodriguez-Barraquer, B. J. Singer, H. J. Park, J. A. Lewnard, D. Sears, and N. C. Lo, *Infectiousness of SARS-CoV-2 breakthrough infections and reinfections during the Omicron wave*, Nat Med, 29 (2023), pp. 358–365. Number: 2 Publisher: Nature Publishing Group.

[33] S. Y. Tartof, J. M. Slezak, H. Fischer, V. Hong, B. K. Ackerson, O. N. Ranasinghe, T. B. Frankland, O. A. Ogun, J. M. Zamparo, S. Gray, S. R. Valluri, K. Pan, F. J. Angulo, L. Jodar, and J. M. McLaughlin, *Effectiveness of mRNA BNT162b2 COVID-19 vaccine up to 6 months in a large integrated health system in the USA: a retrospective cohort study*, The Lancet, 398 (2021), pp. 1407–1416. Publisher: Elsevier.

[34] M. D. Thomas Gluck, *Distinguishing COVID-19 Caused by Omicron vs. Delta*, NEJM Journal Watch, 2022 (2022). Publisher: Journal Watch.

[35] D. J. A. Toth, A. B. Beams, L. T. Keegan, Y. Zhang, T. Greene, B. Orleans, N. Seegert, A. Looney, S. C. Alder, and M. H. Samore, *High variability in transmission of SARS-CoV-2 within households and implications for control*, June 2021.

[36] L. Wang, N. A. Berger, D. C. Kaelber, P. B. Davis, N. D. Volkow, and R. Xu, *COVID infection severity in children under 5 years old before and after Omicron emergence in the US*, medRxiv, (2022), p. 2022.01.12.22269179.

[37] Y. Zhu, C. J. Bloxham, K. D. Hulme, J. E. Sinclair, Z. W. M. Tong, L. E. Steele, E. C. Noye, J. Lu, Y. Xia, K. Y. Chew, J. Pickering, C. Gilks, A. C. Bowen, and K. R. Short, *A Meta-analysis on the Role of Children in Severe Acute Respiratory Syndrome Coronavirus 2 in Household Transmission Clusters*, Clinical Infectious Diseases, 72 (2021), pp. e1146–e1153.

S3.1 Supplementary Tables

Table S3.1: Summary Statistics for All Households

All Households with Surveys			75% Survey Complete		
Variable	Mean	SD	Variable	Mean	SD
All Households			All Households		
Age	43.08	18.19	Age	43.19	18.65
Household Size	2.63	1.56	Household Size	2.50	1.57
Total Sick	0.30	0.77	Total Sick	0.31	0.81
Total Missed Cases	0.13	0.48	Total Missed Cases	0.14	0.50
Households with Children			Households with Children		
Age	25.75	7.32	Age	25.67	7.34
Household Size	4.53	1.4	Household Size	4.47	1.39
Total Sick	0.65	1.20	Total Sick	0.65	1.21
Total Missed Cases	0.32	0.79	Total Missed Cases	0.32	0.79
Households without Children			Households without Children		
Age	48.75	17.06	Age	50.01	17.22
Household Size	2.01	0.98	Household Size	1.73	0.75
Total Sick	0.18	0.51	Total Sick	0.19	0.54
Total Missed Cases	0.07	0.28	Total Missed Cases	0.07	0.29

Chapter 4

Analytic Solution to Small Epidemics

In this chapter, I propose a method to analytically compute household transmission probabilities. We are able to directly solve for these probabilities, and can then use this closed form solution to explore the transmission parameter space and numerically solve for distributions of infected individuals within households. This project builds off of the rich history of chain binomial models proposed by Reed-Frost[2] in 1930 and Greenwood[6] in 1931, and their subsequent extensions by Longani[11] in 1982.

There are many limitations to simulation-based approaches. Even as computational power has increased, the complexities of and processor requirements for agent-based models increase exponentially as detail is added. This project grew from repeated examinations of the survival probability and increasing run-times as I added progressively more detail to what started as relatively simple household models. The traditional methods for calculating the effects of epidemics like the SIR model[9] fail for small populations when the impacts of heterogeneity that are averaged out in large populations are ignored at smaller scales.

Many methods of survival analysis have been explored, and each aim to answer slightly different questions about a system. Many incorporate maximum likelihood, and bootstrapping methods to estimate parameters.[4] Some incorporate transmission trees and random forest algorithms.[7] Most make approximations of the probability space, and many like the Kaplan-Meier[8] approach are non-parametric estimators. Here I aim not to estimate transmission parameters, but to generate robust final size estimates given transmission parameters.

In this project, we consider a four-person household where one housemate is infected and able to import an infection to the rest of the occupants. In order to explore age structure and vaccination status, we consider the possibility that every individual might have their own gamma distributed infectiousness, and susceptibility moderated by σ (4.2.1). We first consider the probability that said importer infects one susceptible housemate; the probability of direct transmission. Then we consider the case where the importer infects m susceptible housemates of n total. We extend the calculation to include assortments of multiple types of individuals, i.e. children and adults, vaccinated and non-vaccinated. Then we describe a scheme where all possible transmission trees are computed, and we are able to find the probabilities of infection for each individual. Using these solutions we are able to calculate a number of additional characteristics of interest like frequency of secondary transmissions and can experiment with the effects of vaccination on transmission. We then compare the results of these analytically acquired final size expectations to those generated by simulation from the previous chapter and find strong agreement between the results from simulation and those analytically derived.

4.1 SIR

4.1.1 Standard Formulation

The SIR model proposed in 1937 by Kermack and McKendrick[9] has been widely utilized for infectious disease modeling. Here S refers to the population of susceptible individuals, I infected, and R recovered. The equations concerning the recovered individuals are often omitted as the dynamics are encapsulated in the other equations.

$$\frac{dS}{dt} = -\beta SI \tag{4.1}$$

$$\frac{dI}{dt} = \beta SI - \gamma I \tag{4.2}$$

$$\frac{dR}{dt} = \gamma I \tag{4.3}$$

In these equations, β is the disease transmission rate, γ is the recovery rate, and the basic reproductive number, $R_0 = \frac{\beta}{\gamma}$. These equations can be explored using a dynamical systems approach and phase plane analysis, but adding complexities

to the model such as population heterogeneity quickly becomes difficult and is often solved with either numerical approaches or simulation-based methods.

4.1.2 Assumptions

The SIR assumes a fixed population N and homogeneous mixing. In the extension below, we also assume a fixed population N, but we remove the homogeneous mixing assumption. The individuals are given their own identities based on their age (adult/child) and vaccination status. In the analytic approach we derive an extension of a chain binomial model to calculate direct transmission probabilities. This approach is able to be extended to any number of groups and utilized to calculate distributions of final epidemic size much more rapidly than the more common simulation-based approaches. Unlike the SIR, this approach assumes distinct generations, and is likely not tractable for large infectious disease epidemics, but could be extended to other chains of events without overlapping generations.

4.2 Calculation of Transmission Probabilities

4.2.1 Survival Function

We assume the probability of transmission to a susceptible housemate is,

$$p = 1 - e^{-\sigma x} \tag{4.4}$$

Here the $x \sim \Gamma(k, r)$ represents the infectiousness of the infectious individual who's infectiousness is parameterized by a gamma distribution specified by the rate parameter r and shape parameter k. The expectation of that distribution is k/r with variance k/r^2. The σ moderates the potential infectee's susceptibility. This probability takes this form of the arrival time for a Poisson counting process. We can see that when $x = 0$, the transmission probability, $p = 0$. As the infectiousness, x increases, since it is multiplied by -1, the exponential approaches 0 and the probability of transmission approaches 1. The shape of this equation might seem counter-intuitive, but the exponential term $e^{-\sigma x}$ is the probability that we see no infections. That probability is then subtracted from the 1 to get the probability of at least one infection. This can be interpreted as

the probability that at least one virion infects an individual and the exponential functional form is used frequently in dose-response models for both viral[5] and bacterial infections.[12,13] The traditional derivation of the dose-response model takes the form of an expected amount of exposure, i.e. the number of organisms that one is exposed to, our gamma distributed x, multiplied by the probability that you are infected by a single organism, in this case σ. Here the gamma distributed x is convenient because we can vary both the mean and variance. This formulation differs from that of both the Greenwood and Reed-Frost models where the transmission probability is assumed to be some fixed p, perhaps varying by generation.

4.2.2 Transmission from Importer

Assume the probability of transmission to a susceptible housemate is $p = 1 - e^{-\sigma x}$ (4.2.1) , where $x \sim \Gamma(k, r)$ and σ defines the vaccine's effect on susceptibility to acquisition. We assume that x is a gamma distributed random variable with rate r and shape k. We can use σ as a proxy for any number of factors which might moderate susceptibility to infection.

To calculate the expected transmission probability from the importer to a single household member, we take this integral which has the product of the survival probability and the gamma distribution's probability density function (**PDF**) in the integrand.

$$p = \int_0^\infty \left(1 - e^{-\sigma x}\right) \frac{x^{k-1} e^{-rx} r^k}{\Gamma(k)}\, dx \tag{4.5}$$

The details of this calculation involve a substitution and the gamma function and are presented in S4.1.1. The expectation of transmission from one infected individual to a single individual with susceptibiltiy moderated by σ is,

$$p = 1 - \left(\frac{r}{\sigma + r}\right)^k \tag{4.6}$$

If we seek the expectation for the non-vaccinated individual, we take $\sigma = 1$ and,

$$p = 1 - \left(\frac{r}{1 + r}\right)^k \tag{4.7}$$

4.2.3 Multiple Transmissions

To calculate the probability of multiple transmissions from one infected individual we need to set up a slightly different integral, in this case we need to add the binomial probability mass function (**PMF**). For these calculations, we must restrict the $n \geq m$. These transmissions do not occur independently of one another, and so can not just be multiplied together in the way of independent probabilities. Here we have the survival function as the probability in the binomial distribution which contains the gamma distributed x so the gamma PDF is included in the integral.

$$p(n,m) = \int_0^\infty \frac{n!}{m!(n-m)!} \left(1-e^{-\sigma x}\right)^m \left(1-(1-e^{-\sigma x})\right)^{n-m} \frac{x^{k-1}e^{-rx}r^k}{\Gamma(k)}\, dx \tag{4.8}$$

Let $m = 2$ to indicate that there are two successful transmissions, rather than the singular one described by the calculation for a direct transmission from one infected individual to another.

$$p(n,2) = \int_0^\infty \frac{n!}{2!(n-2)!} \left(1-e^{-\sigma x}\right)^2 \left(e^{-\sigma x}\right)^{n-2} \frac{x^{k-1}e^{-rx}r^k}{\Gamma(k)}\, dx \tag{4.9}$$

The details of this calculation are again presented in S4.1.4. The solution to the $m = 2$ case is,

$$p(n,2) = \frac{n!r^k}{2!(n-2)!}\left[\frac{1}{(\sigma n - 2\sigma + r)^k} - \frac{2}{(\sigma n - \sigma + r)^k} + \frac{1}{(\sigma n + r)^k}\right] \tag{4.10}$$

We can do the same for the $m = 3$ case, which results in:

$$p(n,3) = \frac{n!r^k}{3!(n-3)!}\left[\frac{1}{(\sigma n - 3\sigma + r)^k} - \frac{3}{(\sigma n - 2\sigma + r)^k} + \frac{3}{(\sigma n - \sigma + r)^k} + \frac{1}{(\sigma n + r)^k}\right] \tag{4.11}$$

And for $m = 4$,

$$p(n,4) = \frac{n!r^k}{4!(n-4)!}\left[\frac{1}{(\sigma n - 4\sigma + r)^k} - \frac{4}{(\sigma n - 3\sigma + r)^k} + \frac{6}{(\sigma n - 2\sigma + r)^k} + \frac{4}{(\sigma n - \sigma + r)^k} + \frac{1}{(\sigma n + r)^k}\right] \tag{4.12}$$

Thus we can hypothesize a solution for the general m. This solution is verified inductively in S4.1.3.

$$p(n,m) = \frac{n!r^k}{m!(n-m)!}\left[\sum_{i=0}^{m}(-1)^i\frac{\binom{m}{i}}{(\sigma n-(m-i)\sigma+r)^k}\right] \tag{4.13}$$

4.3 Multiple Groups

Repeated transmissions are relatively simple to model, but the interesting part is when you begin to add complexity to the system. In this case, the complexity comes in the form of heterogeneity between individuals, i.e.children and adults, vaccinated and non-vaccinated.

4.3.1 Two Groups

The previous result only holds when all individuals have the same σ. We can take a similar approach when there are mixtures of individuals, ie, household members of different susceptibilities. The susceptibilities are relative and could be based on immune features, contact rates, and vaccination status. In this case we consider m infections in n adults with susceptibility σ and q infections in p children with susceptibility ξ. We again have the σ moderated survival probability for the probability transmission, and now we have multiple binomial PMFs for each of the groups. The details of this calculation are presented in S4.1.4.

$$\begin{aligned} p(n,m,q,p) = \int_0^\infty \binom{n}{m}\binom{q}{p}\left(1-e^{-\sigma x}\right)^m\left(e^{-\sigma x}\right)^{n-m} \\ \left(1-e^{-\xi x}\right)^p\left(e^{-\xi x}\right)^{q-p}\left(\frac{x^{k-1}e^{-rx}r^k}{\Gamma(k)}\right)dx \end{aligned} \tag{4.14}$$

$$p(n,m,q,p) = r^k\binom{n}{m}\binom{q}{p}\sum_{i=0}^{m}\left(\sum_{j=0}^{p}\binom{m}{i}\binom{p}{j}\frac{(-1)^{i+j}}{(\sigma(n-m+i)+\xi(q-p+j)+r)^k}\right) \tag{4.15}$$

If you set the $\sigma = \xi$, this equation does not readily simplify to the one group solution. You will find that you are off by a coefficient since, $\binom{n}{m}\binom{q}{p} \neq \binom{n_{tot}}{m_{tot}}$.

This is not an error in the equations, but is coming from the difference in the underlying structure of the system. There are more ways to achieve the m_{tot} infections in n_{tot} potential hosts than there are in two groupings. Thus, to get this to match the one group case, you must consider all the ways to achieve $m+p$ infections in $n+q$ infectees. So in order to get the coefficient correct, you would have to combine several binomial coefficients. In the case of $n_{tot}=5$, $m_{tot}=3$, $n=3$, $q=2$, the binomial coefficient would be $\binom{3}{3}\binom{2}{0}+\binom{3}{2}\binom{2}{1}+\binom{3}{1}\binom{2}{2}$ because if the individuals are the same, there are multiple ways to achieve the same number of infected individuals and you would have to separately account for all of the possible combinations of transmission events which have been missed by fixing m and p in the two group solution.

4.3.2 Three Groups

We now switch to notation involving many subscripts. Groups n_1, n_2, and n_3 with m_1, m_2, and m_3 infections with susceptibilities σ_1, σ_2, and σ_3. The full calculation to achieve this solution is presented in S4.1.4. The integral is set up as follows,

$$p(n_1,m_1,...,n_3,m_3)=\int_0^\infty\left[\prod_{i=1}^{3}\binom{n_i}{m_i}(1-e^{-x\sigma_i})^{m_i}(e^{-x\sigma_i})^{n_i-m_i}\right]\frac{x^{k-1}e^{-rx}r^k}{\Gamma(k)}dx \tag{4.16}$$

with its corresponding solution:

$$\begin{aligned}p(n_1,m_1,...,n_3,m_3)=r^k\prod_{i=1}^{3}\binom{n_i}{m_i}\Bigg[\sum_{j_1=0}^{m_1}\sum_{j_2=0}^{m_2}\sum_{j_3=0}^{m_3}\binom{m_1}{j_1}\binom{m_2}{j_2}\binom{m_3}{j_3}\\ \left(\frac{(-1)^{j_1+j_2+j_3}}{[\sigma_1(n_1-m_1+j_1)+\sigma_2(n_2-m_2+j_2)+\sigma_3(n_3-m_3+j_3)+r]^k}\right)\Bigg]\end{aligned} \tag{4.17}$$

This is the result we need to calculate the probabilities of transmissions for the four person households presented in the previous chapter, but it also gives an intuition for the solution in even more complicated cases.

4.3.3 Any Number of Groups

We are able to identify a pattern, and can repeat this procedure for q groups n_1, n_2...n_q, with their respective number of infected individuals m_1, m_2,...,m_q, and their relative susceptibilities σ_1, σ_2,...,σ_q.

$$p(n_1, m_1, ..., n_q, m_q) = \int_0^\infty \left[\prod_{i=1}^{q} \binom{n_i}{m_i} \left((1 - e^{-x\sigma_i})^{m_i} (e^{-x\sigma_i})^{n_i - m_i} \right) \right] \frac{x^{k-1} e^{-rx} r^k}{\Gamma(k)} \, dx \tag{4.18}$$

The solution for q groups is,

$$\begin{aligned} p(n_1, m_1, ..., n_q, m_q) = r^k & \left[\prod_{i=1}^{q} \binom{n_i}{m_i} \right] \\ & \cdot \left[\sum_{j_1=0}^{m_1} \cdots \sum_{j_q=0}^{m_q} \left[\prod_{i=1}^{q} \binom{m_i}{j_i} \right] \left(\frac{(-1)^{(\sum_{i=1}^{q} j_i)}}{[\sum_{i=1}^{q} \sigma_i (n_i - m_i + j_i) + r]^k} \right) \right] \end{aligned} \tag{4.19}$$

The formal inductive proof to verify this solution is presented in S4.1.4.

Note that this formulation is for transmission from a single person with a specific gamma distributed infectiousness. If you have multiple types of infected individuals capable of transmission, you will need to calculate this equation again for the transmissions from them based on the remaining active infections in the household.

4.4 Application to Data

There are many applications for the calculation above. It provides a robust framework to calculate the probabilities of any number of events, given knowledge of the underlying parameters. There is also likely an extension which uses this calculation as a component in an optimization or maximum likelihood function to approach inverse problems. This calculation calculation was developed in order to populate transmission trees which could be used derive final size estimates for small epidemics and test the effects of vaccination. The binomial models and their extensions[2,6,11] propose methods to estimate transmission probabilities based on data, and the method proposed here could be extended to estimate transmission probabilities from data rather than generate final size distributions given transmission parameter estimates. In **Figure 4.1**, I show non-exhaustive examples of four-person transmission trees. If you fix the importer as a particular individual and do not allow for multiple importations, there are six ways for an epidemic to occur in two susceptible housemates, and thirty-two potential ways for an epidemic to occur in the remaining three

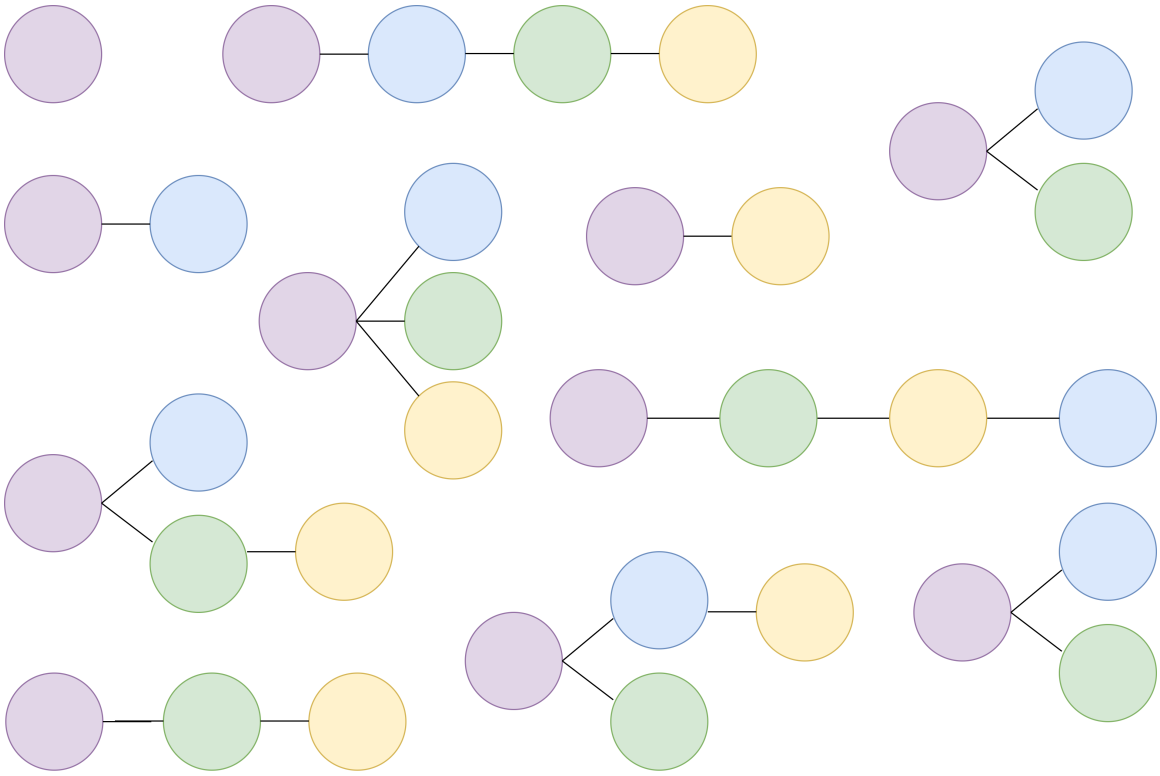


Figure 4.1: Example Trees
between four housemates where each circle represents a person in the household. The importer in each of the example scenarios is represented by a purple circle. Lines between circles from left to right indicate a transmission event. We show 11 of the 32 possible trees in this figure. While this list is not exhaustive of all the potential combinations of trees, it does contain all the potential forms the trees may take.

susceptible housemates. In order to verify the transmission trees, placeholder parameters are assigned to the probabilities for each of the events and multiplied the chains of events, then the values for these chains summed over all the transmission chains to represent the total probability space. This calculation symbolically calculation summed to one and was verified in Maple. Now, to generate final size estimates, it is a matter of populating the transmission trees with probabilities derived from (4.17). Simulation results are computed based on the simulation described in detail in the previous chapter and pictured via the model diagram **Figure 3.1**.

In **Table 4.1** and **Figure 4.2** we see variations of the importer's identity and household vaccination coverage in the first column. The next two columns

Table 4.1: Comparison between Analytic Solution and Simulation Results In the columns labeled adult, we have the probability of transmission to one adult. In the columns labeled total, we have the total expected number of transmissions within the household. Each row represents one configuration where the importer's age category and vaccination status is specified and the number of vaccinated individuals is specified. Since we assume that vaccination does not impact infectiousness merely susceptibility, the vaccination status of the other individual in the same age category as the importer is assumed to be the same as that of the importer. It is possible to alter the distribution of infectiousness of any individual in the analytic solution, so this assumption was made for convenience and could easily be modified.

Household Configuration	Analytic Adult	Simulation Adult	Total Analytic	Simulation Total
Adult Importer, 0 vax kid	0.39661	0.3903	2.18983	2.1662
Adult Importer, 1 vax kid	0.37281	0.3738	1.99769	1.9891
Adult Importer, 2 vax kid	0.35261	0.3509	1.82697	1.8187
Vax Adult Importer, 0 vax kid	0.25525	0.2572	2.00406	2.0164
Vax Adult Importer, 1 vax kid	0.23622	0.2311	1.82313	1.8109
Vax Adult Importer, 2 vax kid	0.21982	0.2214	1.65947	1.6567
Child Importer, 0 vax Adult	0.79322	0.7766	2.18983	2.1577
Child Importer, 1 vax Adult	0.63889	0.6024	2.02253	1.9676
Child Importer, 2 vax Adult	0.47994	0.4681	1.83812	1.8174
Vax Child Importer, 0 vax Adult	0.74562	0.7480	1.99769	2.0006
Vax Child Importer, 1 vax Adult	0.58979	0.5844	1.82697	1.8153
Vax Child Importer, 2 vax Adult	0.43965	0.4409	1.65947	1.6537

are comparisons between the probability that one or both of the adults in each scenario is infected. In the last two columns, we see the total expected infections based on the analytic calculation and the simulations. In order to generate these results for the analytic solution, we compile all of the potential transmission tress, and calculate the probability of each transmission using (4.17) parameterized based on the infectiousness of the infected individual and the susceptibility of each susceptible housemate.

4.5 Discussion

This framework provides a relatively straightforward method for generating final size equations for epidemics in small households. Even in Utah, where families tend to be larger on average than many other parts of the country and world, the average household is made up of only 3.08 people.[1] As long as you are careful about the number of groups you include, it can easily be

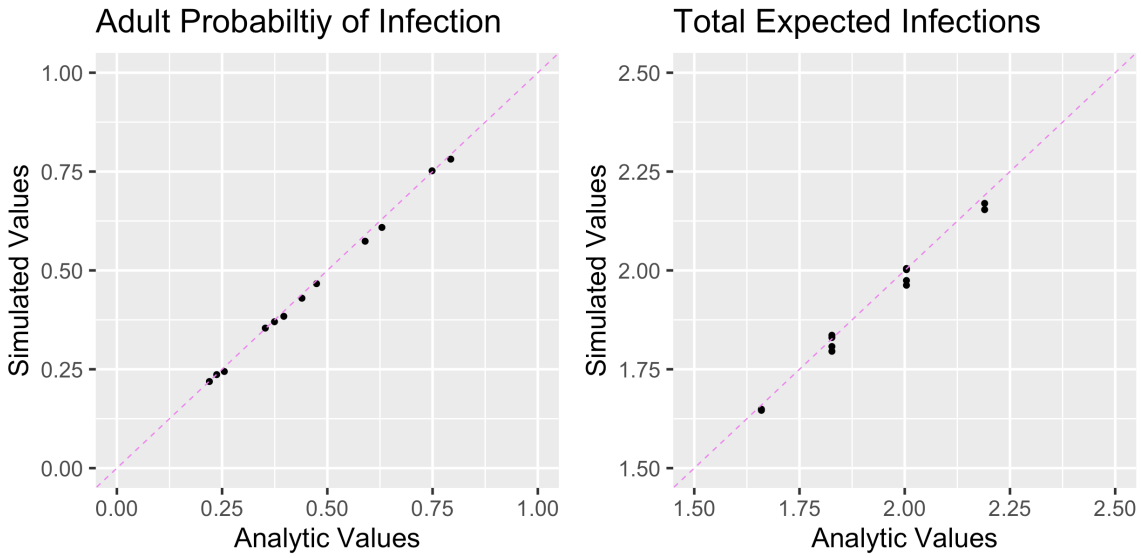


Figure 4.2: Results of Analytic Solution Compared to Simulation
Here we compare the results for each household configuration from scenario one which models households with two children aged 0-13 and two adults aged 25-44 from the previous chapter with various household members vaccinated. Each black dot represents a single household configuration. The pink dashed line passes through the origin with slope one.

extended to larger households. The difficulty in inclusion of multiple groups is two-fold: the first is that the between-group parameter estimates become progressively more difficult to gather and believe, and the second is that the addition of groups makes the compilation of transmission trees exponentially more difficult. By construction, (4.17) will be able to handle the multiple ways multiple individuals can be infected when they are equally susceptible, using a combination of contact rates and immune functions. There is a recursive approach to this problem which resolves many of the bookkeeping difficulties posed by this approach.

If, rather than using the approach outlined above, to estimate the transmission probabilities we instead populate the transmission trees with the direct transmission probabilities, we end up with significantly more infections than we see in reality. This is because the infections are not occurring independently. To accurately account for this, the probabilities are drawn from a binomial distribution rather than the Poisson.

A major limitation of this approach is the assumption of non-overlapping

generations. There was a long incubation period for the original COVID-19 variants, later variants seem to have increased in transmissibility and decreased in incubation time.[3] In small households, like the ones explored in this project, this assumption is fairly reasonable, but in larger small epidemics, like those seen in nursing homes, there is a large potential for generational overlap. Interestingly, this is also a limitation of the assumptions for the simulation as well, so this approach still has improvements over some of the more traditional methods.

4.6 Future Directions and Conclusion

To more fully understand the implications of this solution, a classical sensitivity analysis should be employed. An approach from bifurcation analysis could be used to establish thresholds for the combinations of vaccine efficacy and infectiousness which are required to see significant vaccine effects on secondary transmissions. A formalized approach to generating not only these transmission probabilities, but also the transmission trees recursively would aid functionality of this approach. In order to generate confidence intervals, the variance of these transmission probabilities could be developed.

In this approach, the transmitter has a gamma distributed infectiousness, parameterized by r and k. The individuals have their own susceptibility, moderated by σ. There is a complicated interplay between a transmitter and their potential infectee, moderated by immune factors, contact rate, and strain specific virulence. There are many factors beyond those as well, which means that while the factors affecting transmission can be approximated, they likely change in time and are not constant over the course of either an epidemic or an individual's lifetime. To align this model with previous work[2,11] which estimate not only household SARs, but also community probability of infections this model could be extended to include the influence of community transmission as well. Previous work involving chain binomials suggests that the SARs will be overestimated without that critical step.[10] This would be necessary for appropriate estimation of SAR using this calculation for transmission probabilities. Still, this is an exciting result, and a relatively straightforward extension of chain

binomial models for infectious disease which generates results in line with simulations at dramatically reduced computational efforts and run times.

4.7 References

[1] *U.S. Census Bureau QuickFacts: Utah.*

[2] H. Abbey, *An examination of the Reed-Frost theory of epidemics*, Hum Biol, 24 (1952), pp. 201–233.

[3] E. Amanatidou, A. Gkiouliava, E. Pella, M. Serafidi, D. Tsilingiris, N. G. Vallianou, I. Karampela, and M. Dalamaga, *Breakthrough infections after COVID-19 vaccination: Insights, perspectives and challenges*, Metabol Open, 14 (2022), p. 100180.

[4] P. K. Andersen, O. Borgan, R. D. Gill, and N. Keiding, *Statistical Models Based on Counting Processes*, Springer Series in Statistics, Springer US, New York, NY, 1993.

[5] A. B. Beams, R. Bateman, and F. R. Adler, *Will SARS-CoV-2 Become Just Another Seasonal Coronavirus*, Viruses, 13 (2021), p. 854. Number: 5 Publisher: Multidisciplinary Digital Publishing Institute.

[6] M. Greenwood, *On the Statistical Measure of Infectiousness*, J Hyg Lond, 31 (1931), pp. 336–351.

[7] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, *Random survival forests*, The Annals of Applied Statistics, 2 (2008), pp. 841–860. Publisher: Institute of Mathematical Statistics.

[8] E. L. Kaplan and P. Meier, *Nonparametric Estimation from Incomplete Observations*, Journal of the American Statistical Association, 53 (1958), pp. 457–481.

[9] W. O. Kermack and A. G. McKendrick, *A Contribution to the Mathematical Theory of Epidemics*, Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character, 115 (1927), pp. 700–721. Publisher: The Royal Society.

[10] I. M. Longini and J. S. Koopman, *Household and community transmission parameters from final distributions of infections in households*, Biometrics, 38 (1982), pp. 115–126.

[11] I. M. Longini, J. S. Koopman, A. S. Monto, and J. P. Fox, *Estimating household and community transmission parameters for influenza*, Am J Epidemiol, 115 (1982), pp. 736–751.

[12] B. T. Mayer, J. S. Koopman, E. L. Ionides, J. M. Pujol, and J. N. S. Eisenberg, *A dynamic dose-response model to account for exposure patterns in risk assessment: a case study in inhalation anthrax*, Journal of the Royal Society Interface, 8 (2011), pp. 506–517.

[13] D. J. A. Toth, A. V. Gundlapalli, W. A. Schell, K. Bulmahn, T. E. Walton, C. W. Woods, C. Coghill, F. Gallegos, M. H. Samore, and F. R. Adler, *Quantitative Models of the Dose-Response and Time Course of Inhalational Anthrax in Humans*, PLoS Pathogens, 9 (2013), p. e1003555.

[14] C. Walck, *Hand-book on statistical distributions for experimentalists*, Applied Mathematics, 12 1996.

S4.1 Supplemental Materials

S4.1.1 Transmission from Importer

Assume the probability of transmission to a susceptible, non-vaccinated housemate is $p_u = 1 - e^{-x}$ (the survival probability), and to a susceptible, vaccinated housemate is $p_v = 1 - e^{-\sigma x}$, where $x \sim \Gamma(k, r)$ and σ defines the vaccine's effect on susceptibility to acquisition. We assume that x is a gamma distributed random variable with rate r and shape k. Gamma distributed random variables are used to model the time between events.[14] We can use σ as a proxy for any number of factors which might moderate susceptibility to infection.

To calculate the expected transmission probability from the importer to a single household member, we take this integral which has the product of the survival probability and the gamma distribution's probability density function (**PDF**) in the integrand.

$$p = \int_0^\infty \left(1 - e^{-\sigma x}\right) \frac{x^{k-1} e^{-rx} r^k}{\Gamma(k)}\, dx \tag{S1}$$

The integral is not dependant on r or k, so we pull out those coefficients.

$$= \frac{r^k}{\Gamma(k)} \int_0^\infty \left(1 - e^{-\sigma x}\right) x^{k-1} e^{-rx}\, dx \tag{S2}$$

We then expand via multiplication.

$$= \frac{r^k}{\Gamma(k)} \int_0^\infty x^{k-1} e^{-rx} - x^{k-1} e^{-(\sigma+r)x}\, dx \tag{S3}$$

Perform a substitution. Let $u = rx \to x = \frac{u}{r}, dx = \frac{du}{r}$
Let $v = (\sigma + r)x \to x = \frac{v}{\sigma+r}, dx = \frac{dv}{\sigma+r}$

$$p = \frac{r^k}{\Gamma(k)} \left[\int_0^\infty \left(\frac{u}{r}\right)^{k-1} e^{-u} \frac{1}{r} du - \int_0^\infty \left(\frac{v}{\sigma + r}\right)^{k-1} e^{-v} \frac{1}{\sigma + r} dv \right] \tag{S4}$$

Combine the denominators and move them out of the integral since they are not dependant on the variable of integration.

$$= \frac{r^k}{\Gamma(k)} \left[\frac{1}{r^k} \int_0^\infty u^{k-1} e^{-u} du - \frac{1}{(\sigma + r)^k} \int_0^\infty v^{k-1} e^{-v} dv \right] \tag{S5}$$

Use the definition of the gamma function to solve the integrals.

$$= \frac{r^k}{\Gamma(k)} \left[\frac{1}{r^k} \Gamma(k) - \frac{1}{(\sigma + r)^k} \Gamma(k) \right] \tag{S6}$$

Collect the gamma terms and cancel them throughout the expression for the solution.

$$= r^k \left[\frac{1}{r^k} - \frac{1}{(\sigma + r)^k} \right] \tag{S7}$$

Thus the expectation of transmission to a vaccinated individual is,

$$p = 1 - \left(\frac{r}{\sigma + r} \right)^k \tag{S8}$$

If we seek the expectation for the non-vaccinated individual, we take $\sigma = 1$ and find,

$$p = 1 - \left(\frac{r}{1 + r} \right)^k \tag{S9}$$

S4.1.2 Multiple Transmissions

To calculate the probability of multiple transmissions from one infected individual we need to set up a slightly different integral, in this case we need to add the binomial probability mass function (**PMF**). For these calculations, we must restrict the $n \geq m$. These transmissions do not occur independently of one another, and so can not just be multiplied together in the way of independent probabilities. Here we have the survival function as the probability in the binomial distribution which contains the gamma distributed x so the gamma PDF is included in the integral.

$$p(n, m) = \int_0^\infty \frac{n!}{m!(n-m)!} \left(1 - e^{-\sigma x}\right)^m \left(1 - (1 - e^{-\sigma x})\right)^{n-m} \frac{x^{k-1} e^{-rx} r^k}{\Gamma(k)} \, dx \tag{S10}$$

Cancel the 1s in the third term.

$$= \int_0^\infty \frac{n!}{m!(n-m)!} \left(1 - e^{-\sigma x}\right)^m \left(e^{-\sigma x}\right)^{n-m} \frac{x^{k-1} e^{-rx} r^k}{\Gamma(k)} \, dx \tag{S11}$$

Let $m = 2$ to indicate that there are two successful transmissions, rather than the singular one described by the calculation for a direct transmission from one infected individual to another.

$$p(n, 2) = \int_0^\infty \frac{n!}{2!(n-2)!} \left(1 - e^{-\sigma x}\right)^2 \left(e^{-\sigma x}\right)^{n-2} \frac{x^{k-1} e^{-rx} r^k}{\Gamma(k)} \, dx \tag{S12}$$

Use the constant multiple rule to pull the coefficients independent of the variable of integration out of the integral.

$$p(n,2) = \frac{n!r^k}{\Gamma(k)2!(n-2)!}\int_0^\infty \left(1-e^{-\sigma x}\right)^2\left(e^{-\sigma x}\right)^{n-2}x^{k-1}e^{-rx}\,dx \tag{S13}$$

Exponentiate the first term of the integral.

$$= \frac{n!r^k}{\Gamma(k)2!(n-2)!}\int_0^\infty \left(1-2e^{-\sigma x}+e^{-2\sigma x}\right)\left(e^{-\sigma x}\right)^{n-2}x^{k-1}e^{-rx}\,dx \tag{S14}$$

Use the power rule for exponents to expand the second term of the integral.

$$= \frac{n!r^k}{\Gamma(k)2!(n-2)!}\int_0^\infty \left(1-2e^{-\sigma x}+e^{-2\sigma x}\right)\left(e^{2\sigma x}\right)\left(e^{-\sigma xn}\right)x^{k-1}e^{-rx}\,dx \tag{S15}$$

Multiply the first and second terms of the integral.

$$= \frac{n!r^k}{\Gamma(k)2!(n-2)!}\int_0^\infty \left(e^{2\sigma x}-2e^{\sigma x}+1\right)\left(e^{-\sigma xn}\right)x^{k-1}e^{-rx}\,dx \tag{S16}$$

Multiply the first and second terms of the integral.

$$= \frac{n!r^k}{\Gamma(k)2!(n-2)!}\int_0^\infty \left(e^{2\sigma x-\sigma xn}-2e^{\sigma x-\sigma xn}+e^{-\sigma xn}\right)x^{k-1}e^{-rx}\,dx \tag{S17}$$

Multiply the first and last terms of the integral.

$$= \frac{n!r^k}{\Gamma(k)2!(n-2)!}\int_0^\infty \left(e^{2\sigma x-\sigma xn-rx}-2e^{\sigma x-\sigma xn-rx}+e^{-\sigma xn-rx}\right)x^{k-1}\,dx \tag{S18}$$

Factor -1 and x from the exponent in the first term of the integral.

$$= \frac{n!r^k}{\Gamma(k)2!(n-2)!}\int_0^\infty \left(e^{-(\sigma n-2\sigma+r)x}-2e^{-(\sigma n-\sigma+r)x}+e^{-(\sigma n+r)x}\right)x^{k-1}dx \tag{S19}$$

Multiply the terms in the integral and use the sum rule to separate this into three integrals.

$$\begin{aligned}= \frac{n!r^k}{\Gamma(k)2!(n-2)!}\Bigg(&\int_0^\infty e^{-(\sigma n-2\sigma+r)x}x^{k-1}dx\\ &-\int_0^\infty 2e^{-(\sigma n-\sigma+r)x}x^{k-1}dx+\int_0^\infty e^{-(\sigma n+r)x}x^{k-1}dx\Bigg)\end{aligned} \tag{S20}$$

Perform a similar substitution as above.

Let $u=(\sigma n-2\sigma+r)x \rightarrow x=\frac{u}{\sigma n-2\sigma+r} \rightarrow dx=\frac{du}{\sigma n-2\sigma+r}$.

Let $v=(\sigma n-\sigma+r)x \rightarrow x=\frac{v}{\sigma n-\sigma+r} \rightarrow dx=\frac{dv}{\sigma n-\sigma+r}$.

Let $w=(\sigma n+r)x \rightarrow \frac{w}{\sigma n+r} \rightarrow dx=\frac{dw}{\sigma n+r}$.

$$
\begin{aligned}
=&\frac{n!r^k}{\Gamma(k)2!(n-2)!}\left(\int_0^\infty e^{-u}\left(\frac{u}{\sigma n-2\sigma+r}\right)^{k-1}\frac{du}{\sigma n-2\sigma+r}\right.\\
&\left.-\int_0^\infty 2e^{-v}\left(\frac{v}{\sigma n-\sigma+r}\right)^{k-1}\frac{dv}{\sigma n-\sigma+r}+\int_0^\infty e^{-w}\left(\frac{w}{\sigma n+r}\right)^{k-1}\frac{dw}{\sigma n+r}\right)
\end{aligned}
\tag{S21}
$$

Combine the matching denominators.

$$
\begin{aligned}
=&\frac{n!r^k}{\Gamma(k)2!(n-2)!}\left(\int_0^\infty e^{-u}u^{k-1}\frac{du}{(\sigma n-2\sigma+r)^k}\right.\\
&\left.-\int_0^\infty 2e^{-v}v^{k-1}\frac{dv}{(\sigma n-\sigma+r)^k}+\int_0^\infty e^{-w}w^{k-1}\frac{dw}{(\sigma n+r)^k}\right)
\end{aligned}
\tag{S22}
$$

Factor out coefficients.

$$
\begin{aligned}
=&\frac{n!r^k}{\Gamma(k)2!(n-2)!}\left(\frac{1}{(\sigma n-2\sigma+r)^k}\int_0^\infty e^{-u}u^{k-1}du\right.\\
&\left.-\frac{2}{(\sigma n-\sigma+r)^k}\int_0^\infty e^{-v}v^{k-1}dv+\frac{1}{(\sigma n+r)^k}\int_0^\infty e^{-w}w^{k-1}dw\right)
\end{aligned}
\tag{S23}
$$

We now have three integrals which are recognizably the gamma function.

$$
=\frac{n!r^k}{\Gamma(k)2!(n-2)!}\left[\frac{\Gamma(k)}{(\sigma n-2\sigma+r)^k}-\frac{2\Gamma(k)}{(\sigma n-\sigma+r)^k}+\frac{\Gamma(k)}{(\sigma n+r)^k}\right] \tag{S24}
$$

Then we can cancel out the $\Gamma(k)$ terms.

$$
=\frac{n!r^k}{2!(n-2)!}\left[\frac{1}{(\sigma n-2\sigma+r)^k}-\frac{2}{(\sigma n-\sigma+r)^k}+\frac{1}{(\sigma n+r)^k}\right] \tag{S25}
$$

We can use the same approach for the $m=3$ case, which results in:

$$
\begin{aligned}
p(n,3)=\frac{n!r^k}{3!(n-3)!}&\left[\frac{1}{(\sigma n-3\sigma+r)^k}-\frac{3}{(\sigma n-2\sigma+r)^k}+\right.\\
&\left.\frac{3}{(\sigma n-\sigma+r)^k}+\frac{1}{(\sigma n+r)^k}\right]
\end{aligned}
\tag{S26}
$$

And for $m=4$,

$$
\begin{aligned}
p(n,4)=\frac{n!r^k}{4!(n-4)!}&\left[\frac{1}{(\sigma n-4\sigma+r)^k}-\frac{4}{(\sigma n-3\sigma+r)^k}+\frac{6}{(\sigma n-2\sigma+r)^k}+\right.\\
&\left.\frac{4}{(\sigma n-\sigma+r)^k}+\frac{1}{(\sigma n+r)^k}\right]
\end{aligned}
\tag{S27}
$$

Thus we can hypothesize a solution for the general m.

$$p(n,m) = \frac{n!r^k}{m!(n-m)!}\left[\sum_{i=0}^{m}(-1)^i\frac{\binom{m}{i}}{(\sigma n-(m-i)\sigma+r)^k}\right] \tag{S28}$$

If we choose $n = m = 1$, we can use this equation to verify the expectation for primary transmission from above.

$$p(1,1) = \frac{1!r^k}{1!(1-1)!}\left[\frac{1}{(\sigma-(1)\sigma+r)^k} - \frac{1}{(\sigma-(0)\sigma+r)^k}\right]. \tag{S29}$$

Simplify.

$$= r^k\left[\frac{1}{(r)^k} - \frac{1}{(\sigma+r)^k}\right] \tag{S30}$$

Multiply r^k through the expression.

$$= \frac{r^k}{(r)^k} - \frac{r^k}{(\sigma+r)^k} \tag{S31}$$

Simplify.

$$= 1 - \left(\frac{r}{\sigma+r}\right)^k \tag{S32}$$

S4.1.3 Induction for Multiple Infections

In order to formalize and verify (4.13) by way of induction that

$$\begin{aligned} p(n,m) &= \int_0^\infty \frac{n!}{m!(n-m)!}\left(1-e^{-\sigma x}\right)^m\left(e^{-\sigma x}\right)^{n-m}\frac{x^{k-1}e^{-rx}r^k}{\Gamma(k)}\,dx \\ &= \frac{n!r^k}{m!(n-m)!}\left[\sum_{i=0}^{m}(-1)^i\frac{\binom{m}{i}}{(\sigma n-(m-i)\sigma+r)^k}\right] \end{aligned}$$

holds for all $m \leq n$ where $\{n,m,k\} \in \mathbb{Z}$ and $\{n,m,k,r\} > 0$.

Consider the base case where there is one successful transmission in a susceptible population of n individuals $m = 1$:

$$p(n,1) = \int_0^\infty \frac{n!}{1!(n-1)!}\left(1-e^{-\sigma x}\right)^1\left(e^{-\sigma x}\right)^{n-1}\frac{x^{k-1}e^{-rx}r^k}{\Gamma(k)}\,dx \tag{S33}$$

Factor out coefficients not dependant on the variable of integration.

$$= \frac{n!r^k}{\Gamma(k)(n-1)!}\int_0^\infty\left(1-e^{-\sigma x}\right)\left(e^{-\sigma x}\right)^{n-1}x^{k-1}e^{-rx}\,dx \tag{S34}$$

Expand the second term of the integral.

$$= \frac{n! r^k}{\Gamma(k)(n-1)!} \int_0^\infty \left(1 - e^{-\sigma x}\right) \left(e^{\sigma x}\right) \left(e^{-\sigma x n}\right) x^{k-1} e^{-rx} \, dx \tag{S35}$$

Multiply the first two terms of the integral.

$$= \frac{n! r^k}{\Gamma(k)(n-1)!} \int_0^\infty \left(e^{\sigma x} - 1\right) \left(e^{-\sigma x n}\right) x^{k-1} e^{-rx} \, dx \tag{S36}$$

Multiply the first and second terms of the integral.

$$= \frac{n! r^k}{\Gamma(k)(n-1)!} \int_0^\infty \left(e^{-(\sigma n - \sigma)x} - e^{-\sigma x n}\right) x^{k-1} e^{-rx} \, dx \tag{S37}$$

Multiply the first and last terms of the integral, factoring -1 and x in the exponents.

$$= \frac{n! r^k}{\Gamma(k)(n-1)!} \int_0^\infty \left(e^{-(\sigma n - \sigma + r)x} - e^{-(\sigma n + r)x}\right) x^{k-1} \, dx \tag{S38}$$

Multiply the last term and use the sum rule to split the integral into two.

$$= \frac{n! r^k}{\Gamma(k)(n-1)!} \left[\int_0^\infty e^{-(\sigma n - \sigma + r)x} x^{k-1} \, dx - \int_0^\infty e^{-(\sigma n + r)x} x^{k-1} \, dx\right] \tag{S39}$$

We perform a substitution by letting $u = (\sigma n + r)x \to x = \frac{u}{\sigma n + r} \to dx = \frac{du}{\sigma n + r}$ and $v = (\sigma n - \sigma + r)x \to x = \frac{v}{\sigma n - \sigma + r} \to dx = \frac{dv}{\sigma n - \sigma + r}$,

$$\begin{aligned} = \frac{n! r^k}{\Gamma(k)(n-1)!} \Bigg[&\int_0^\infty e^{-v} \left(\frac{v}{\sigma n - \sigma + r}\right)^{k-1} \frac{dv}{(\sigma n - \sigma + r)} \\ &- \int_0^\infty e^{-u} \left(\frac{u}{\sigma n + r}\right)^{k-1} \frac{du}{(\sigma n + r)}\Bigg] \end{aligned} \tag{S40}$$

Combine the matching denominators.

$$= \frac{n! r^k}{\Gamma(k)(n-1)!} \left[\int_0^\infty e^{-v} (v)^{k-1} \frac{dv}{(\sigma n - \sigma + r)^k} - \int_0^\infty e^{-u} (u)^{k-1} \frac{du}{(\sigma n + r)^k}\right] \tag{S41}$$

Factor coefficients independent of the variable of integration.

$$= \frac{n! r^k}{\Gamma(k)(n-1)!} \left[\frac{1}{(\sigma n - \sigma + r)^k} \int_0^\infty e^{-v} (v)^{k-1} dv - \frac{1}{(\sigma n + r)^k} \int_0^\infty e^{-u} (u)^{k-1} du\right] \tag{S42}$$

By the definition of the gamma function,

$$= \frac{n! r^k}{\Gamma(k)(n-1)!} \left[\frac{\Gamma(k)}{(\sigma n - \sigma + r)^k} - \frac{\Gamma(k)}{(\sigma n + r)^k}\right] \tag{S43}$$

Cancel the $\Gamma(k)$ terms.

$$= \frac{n!r^k}{(n-1)!}\left[\frac{1}{(\sigma n-\sigma+r)^k}-\frac{1}{(\sigma n+r)^k}\right] \tag{S44}$$

Utilize $1!=1$, the definition of the $\binom{m}{i}$, and algebra in the denominators.

$$= \frac{n!r^k}{1!(n-1)!}\left[\frac{\binom{1}{0}}{(\sigma n-(1-0)\sigma+r)^k}-\frac{\binom{1}{1}}{(\sigma n(1-1)\sigma+r)^k}\right] \tag{S45}$$

Convert to sigma notation.

$$= \frac{n!r^k}{1!(n-1)!}\left[\sum_{i=0}^{1}(-1)^i\frac{\binom{1}{i}}{(\sigma n-(1-i)\sigma+r)^k}\right] \tag{S46}$$

Thus we have demonstrated the base case holds.
Now, by way of induction, since $p(n,m)$ holds for $m=1$, assume for some fixed but arbitrary $m>1$, $p(n,m)$ is true. So we consider $p(n,m+1)$.

$$p(n,m+1)=\int_0^\infty \frac{n!}{(m+1)!(n-(m+1))!}\left(1-e^{-\sigma x}\right)^{m+1}\left(e^{-\sigma x}\right)^{n-(m+1)}\frac{x^{k-1}e^{-rx}r^k}{\Gamma(k)}\,dx \tag{S47}$$

Factor out terms independent of the variable of integration.

$$= \frac{n!r^k}{\Gamma(k)(m+1)!(n-m-1)!}\int_0^\infty \left(1-e^{-\sigma x}\right)^{m+1}\left(e^{-\sigma x}\right)^{(n-m-1)}x^{k-1}e^{-rx}\,dx \tag{S48}$$

By the definition of the binomial expansion we have,

$$= \frac{n!r^k}{\Gamma(k)(m+1)!(n-m-1)!}\int_0^\infty \left[\sum_{i=0}^{m+1}\binom{m+1}{i}1^{m+1-i}(-e^{-\sigma x})^i\right]e^{-\sigma x(n-m-1)}x^{k-1}e^{-rx}\,dx \tag{S49}$$

We use the power of one rule.

$$= \frac{n!r^k}{\Gamma(k)(m+1)!(n-m-1)!}\int_0^\infty \left[\sum_{i=0}^{m+1}\binom{m+1}{i}\left(-e^{-\sigma x}\right)^i\right]e^{-\sigma x(n-m-1)}x^{k-1}e^{-rx}\,dx \tag{S50}$$

Factor the -1 out of the second term in the summation.

$$= \frac{n!r^k}{\Gamma(k)(m+1)!(n-m-1)!}\int_0^\infty \left[\sum_{i=0}^{m+1}\binom{m+1}{i}(-1)^i\left(e^{-\sigma xi}\right)\right]e^{-\sigma x(n-m-1)}x^{k-1}e^{-rx}\,dx \tag{S51}$$

Combine the last two exponential terms of the integral.

$$= \frac{n!r^k}{\Gamma(k)(m+1)!(n-m-1)!} \int_0^\infty \left[\sum_{i=0}^{m+1} \binom{m+1}{i} (-1)^i \left(e^{-\sigma x i}\right) \right] \left(e^{-(\sigma n - \sigma m - \sigma + r)x}\right) x^{k-1}\, dx \tag{S52}$$

Multiply the terms in the integral outside of the summation into the summation.

$$= \frac{n!r^k}{\Gamma(k)(m+1)!(n-m-1)!} \int_0^\infty \left[\sum_{i=0}^{m+1} \binom{m+1}{i} (-1)^i \left(e^{-\sigma x i} e^{-(\sigma n - \sigma m - \sigma + r)x}\right) x^{k-1} \right] dx \tag{S53}$$

Use the power rule for exponents to combine the exponential terms.

$$= \frac{n!r^k}{\Gamma(k)(m+1)!(n-m-1)!} \int_0^\infty \left[\sum_{i=0}^{m+1} \binom{m+1}{i} (-1)^i \left(e^{-(\sigma n - \sigma m + \sigma i - \sigma + r)x}\right) x^{k-1} \right] dx \tag{S54}$$

Since $r, x > 0$,$k \in \mathbb{Z}^+$, and $m, n, i \in \mathbb{Z}$, by Fubini's theorem we can take the summation out of the integral.

$$= \frac{n!r^k}{\Gamma(k)(m+1)!(n-m-1)!} \left[\sum_{i=0}^{m+1} \binom{m+1}{i} (-1)^i \int_0^\infty \left(e^{-(\sigma n - \sigma m + \sigma i - \sigma + r)x}\right) x^{k-1}\, dx \right] \tag{S55}$$

Let $v = ((n-m+i-1)\sigma + r)x \to x = \frac{v}{((n-m+i-1)\sigma+r)} \to dx = \frac{dv}{(n-m+i-1)\sigma+r)}$.

$$\begin{aligned} = \frac{n!r^k}{\Gamma(k)(m+1)!(n-m-1)!} \Bigg[& \sum_{i=0}^{m+1} \binom{m+1}{i} (-1)^i \\ & \cdot \int_0^\infty e^{-v} \left(\frac{v}{((n-m+i-1)\sigma + r)} \right)^{k-1} \frac{dv}{((n-m+i-1)\sigma + r)} \Bigg] \end{aligned} \tag{S56}$$

Combine the matching denominators.

$$= \frac{n!r^k}{\Gamma(k)(m+1)!(n-m-1)!} \left[\sum_{i=0}^{m+1} \binom{m+1}{i} (-1)^i \int_0^\infty e^{-v} v^{k-1} \frac{dv}{((n-m+i-1)\sigma + r)^k} \right] \tag{S57}$$

Factor out the denominator under the dv.

$$= \frac{n!r^k}{\Gamma(k)(m+1)!(n-m-1)!} \left[\sum_{i=0}^{m+1} \binom{m+1}{i} \frac{(-1)^i}{((n-m+i-1)\sigma + r)^k} \int_0^\infty e^{-v} v^{k-1} dv \right] \tag{S58}$$

By the definition of the gamma function,

$$= \frac{n! r^k}{\Gamma(k)(m+1)!(n-m-1)!} \left[\sum_{i=0}^{m+1} \binom{m+1}{i} \frac{(-1)^i \Gamma(k)}{((n-m+i-1)\sigma + r)^k} \right] \tag{S59}$$

Cancel out the gamma function terms and rearrange the denominator terms to match the notation from (4.13).

$$= \frac{n! r^k}{(m+1)!(n-(m+1))!} \left[\sum_{i=0}^{m+1} \binom{m+1}{i} \frac{(-1)^i}{(\sigma n - [(m+1)-i]\sigma + r)^k} \right] \tag{S60}$$

So, by the principle of mathematical induction, we have shown that

$$\begin{aligned} p(n,m) &= \int_0^\infty \frac{n!}{m!(n-m)!} \left(1-e^{-\sigma x}\right)^m \left(e^{-\sigma x}\right)^{n-m} \left(\frac{x^{k-1} e^{-rx} r^k}{\Gamma(k)} \right) dx \\ &= \frac{n! r^k}{m!(n-m)!} \left[\sum_{i=0}^{m} (-1)^i \frac{\binom{m}{i}}{(\sigma n - (m-i)\sigma + r)^k} \right] \end{aligned} \tag{S61}$$

S4.1.4 Multiple Groups

Two Groups

The previous result only holds when all individuals have the same σ. We can take a similar approach when there are mixtures of individuals, i.e., household members of different susceptibilities. In this case we consider m infections in n adults with susceptibility σ and q infections in p children with susceptibility ξ.

$$\begin{aligned} p(n,m,q,p) = \int_0^\infty & \binom{n}{m} \binom{q}{p} \left(1-e^{-\sigma x}\right)^m \left(e^{-\sigma x}\right)^{n-m} \\ & \cdot \left(1-e^{-\xi x}\right)^p \left(e^{-\xi x}\right)^{q-p} \left(\frac{x^{k-1} e^{-rx} r^k}{\Gamma(k)} \right) dx \end{aligned} \tag{S62}$$

Factor out coefficients independent of the variable of integration.

$$\begin{aligned} = \frac{r^k}{\Gamma(k)} \binom{n}{m} \binom{q}{p} \int_0^\infty & \left(1-e^{-\sigma x}\right)^m \left(e^{-\sigma x}\right)^{n-m} \\ & \cdot \left(1-e^{-\xi x}\right)^p \left(e^{-\xi x}\right)^{q-p} \left(x^{k-1} e^{-rx} \right) dx \end{aligned} \tag{S63}$$

By the binomial theorem applied twice,

$$= \frac{r^k}{\Gamma(k)}\binom{n}{m}\binom{q}{p}\int_0^\infty \left[\sum_{i=0}^{m}\binom{m}{i}\left(1\right)^{m-i}\left(-e^{-\sigma x}\right)^i\right]\left(e^{-\sigma x}\right)^{n-m} \left[\sum_{j=0}^{p}\binom{p}{j}\left(1\right)^{p-j}\left(-e^{-\xi x}\right)^j\right]\left(e^{-\xi x}\right)^{q-p}\left(x^{k-1}e^{-rx}\right)\,dx \tag{S64}$$

By the power of one rule,

$$= \frac{r^k}{\Gamma(k)}\binom{n}{m}\binom{q}{p}\int_0^\infty \left[\sum_{i=0}^{m}\binom{m}{i}\left(-e^{-\sigma x}\right)^i\right]\left(e^{-\sigma x}\right)^{n-m} \left[\sum_{j=0}^{p}\binom{p}{j}\left(-e^{-\xi x}\right)^j\right]\left(e^{-\xi x}\right)^{q-p}\left(x^{k-1}e^{-rx}\right)\,dx \tag{S65}$$

We factor the -1 out of the exponential terms in the summation.

$$= \frac{r^k}{\Gamma(k)}\binom{n}{m}\binom{q}{p}\int_0^\infty \left[\sum_{i=0}^{m}\binom{m}{i}\left(-1\right)^i\left(e^{-\sigma x}\right)^i\right]\left(e^{-\sigma x}\right)^{n-m} \cdot\left[\sum_{j=0}^{p}\binom{p}{j}\left(-1\right)^j\left(e^{-\xi x}\right)^j\right]\left(e^{-\xi x}\right)^{q-p}\left(x^{k-1}e^{-rx}\right)\,dx \tag{S66}$$

Multiply the exponents with like variables into the summations, they will all be combined in the next step.

$$= \frac{r^k}{\Gamma(k)}\binom{n}{m}\binom{q}{p}\int_0^\infty \left[\sum_{i=0}^{m}\binom{m}{i}\left(-1\right)^i\left(e^{-\sigma x}\right)^{n-m+i}\right] \cdot\left[\sum_{j=0}^{p}\binom{p}{j}\left(-1\right)^j\left(e^{-\xi x}\right)^{q-p+j}\right]\left(x^{k-1}e^{-rx}\right)\,dx \tag{S67}$$

By the distributive law of multiplication,

$$= \frac{r^k}{\Gamma(k)}\binom{n}{m}\binom{q}{p}\int_0^\infty \sum_{i=0}^{m}\left(\sum_{j=0}^{p}\binom{m}{i}\binom{p}{j}\left(-1\right)^{i+j} \cdot\left(e^{-\sigma x(n-m+i)}\right)\left(e^{-\xi x(q-p+j)}\right)\left(x^{k-1}e^{-rx}\right)\,dx\right) \tag{S68}$$

Again, by Fubini's theorem,

$$= \frac{r^k}{\Gamma(k)} \binom{n}{m} \binom{q}{p} \sum_{i=0}^{m} \left(\sum_{j=0}^{p} \binom{m}{i} \binom{p}{j} \big(-1 \big)^{i+j} \cdot \int_0^\infty \left(e^{-x[\sigma(n-m+i)+\xi(q-p+j)+r]} \right) x^{k-1} \, dx \right) \tag{S69}$$

Using a similar substitution as above, let $v = x[\sigma(n-m+i) - \xi(q-p+j) + r]$, then $x = \frac{v}{\sigma(n-m+i)+\xi(q-p+j)+r}$ and $dx = \frac{dv}{\sigma(n-m+i)+\xi(q-p+j)+r}$.

$$= \frac{r^k}{\Gamma(k)} \binom{n}{m} \binom{q}{p} \sum_{i=0}^{m} \left(\sum_{j=0}^{p} \binom{m}{i} \binom{p}{j} \big(-1 \big)^{i+j} \cdot \int_0^\infty \left(e^{-v} \right) \left(\frac{v}{\sigma(n-m+i) + \xi(q-p+j) + r} \right)^{k-1} \frac{dv}{(\sigma(n-m+i) + \xi(q-p+j) + r)} \right) \tag{S70}$$

Combine the denominators and factor them out of the integral.

$$= \frac{r^k}{\Gamma(k)} \binom{n}{m} \binom{q}{p} \sum_{i=0}^{m} \left(\sum_{j=0}^{p} \binom{m}{i} \binom{p}{j} \big(-1 \big)^{i+j} \frac{1}{(\sigma(n-m+i) + \xi(q-p+j) + r)^k} \int_0^\infty \left(e^{-v} \right) v^{k-1} dv \right) \tag{S71}$$

Then with the definition of the gamma function we have this solution for transmission to members of two groups.

$$p(n, m, q, p) = r^k \binom{n}{m} \binom{q}{p} \sum_{i=0}^{m} \left(\sum_{j=0}^{p} \binom{m}{i} \binom{p}{j} \frac{(-1)^{i+j}}{(\sigma(n-m+i) + \xi(q-p+j) + r)^k} \right) \tag{S72}$$

Three Groups

We now switch to notation involving many subscripts. Groups n_1, n_2, and n_3 with m_1, m_2, and m_3 infections with susceptibilities σ_1, σ_2, and σ_3. The setup follows that of the setup for the two group case, with the sigma mediated survival function as the probability in the binomial PMF for each group, and the gamma

PDF to account for the gamma distributed infectiousness.

$$p(n_1, m_1, ..., n_3, m_3) = \int_0^\infty \left[\prod_{i=1}^{3} \binom{n_i}{m_i} (1 - e^{-x\sigma_i})^{m_i} (e^{-x\sigma_i})^{n_i - m_i}\right] \frac{x^{k-1} e^{-rx} r^k}{\Gamma(k)} dx$$

Factor out the coefficients which are independent of the variable of integration.

$$= \frac{r^k}{\Gamma(k)} \left[\prod_{i=1}^{3} \binom{n_i}{m_i}\right] \int_0^\infty \prod_{i=1}^{3} \left[(1 - e^{-x\sigma_i})^{m_i} (e^{-x\sigma_i})^{n_i - m_i}\right] x^{k-1} e^{-rx}\, dx \tag{S73}$$

Use the binomial theorem in the product in the integral.

$$= \frac{r^k}{\Gamma(k)} \left[\prod_{i=1}^{3} \binom{n_i}{m_i}\right] \int_0^\infty \left[\prod_{i=1}^{3} \left(\sum_{j_i=0}^{m_i} \binom{m_i}{j_i} (1)^{m_i - j_i} \Big(- e^{-x\sigma_i}\Big)^{j_i}\right) e^{-x\sigma_i (n_i - m_i)}\right] x^{k-1} e^{-rx}\, dx \tag{S74}$$

The power of one rule simplifies things a bit.

$$= \frac{r^k}{\Gamma(k)} \left[\prod_{i=1}^{3} \binom{n_i}{m_i}\right] \int_0^\infty \left[\prod_{i=1}^{3} \left(\sum_{j_i=0}^{m_i} \binom{m_i}{j_i} \Big(- e^{(-x\sigma_i)}\Big)^{j_i}\right) \left(e^{-x\sigma_i}\right)^{n_i - m_i}\right] x^{k-1} e^{-rx}\, dx \tag{S75}$$

Separate the product in the integral into two.

$$= \frac{r^k}{\Gamma(k)} \left[\prod_{i=1}^{3} \binom{n_i}{m_i}\right] \int_0^\infty \left[\prod_{i=1}^{3} \left(\sum_{j_i=0}^{m_i} \binom{m_i}{j_i} (-e^{(-x\sigma_i)})^{j_i}\right)\right] \left[\prod_{i=1}^{3} \left(e^{-x\sigma_i}\right)^{n_i - m_i}\right] x^{k-1} e^{-rx}\, dx \tag{S76}$$

Factor -1 out of the exponential term in the summation in the integral.

$$= \frac{r^k}{\Gamma(k)} \left[\prod_{i=1}^{3} \binom{n_i}{m_i}\right] \int_0^\infty \left(\prod_{i=1}^{3} \sum_{j_i=0}^{m_i} \binom{m_i}{j_i} (-1)^{j_i} (e^{-x\sigma_i j_i})\right) \left[\prod_{i=1}^{3} \left(e^{-x\sigma_i}\right)^{n_i - m_i}\right] x^{k-1} e^{-rx}\, dx \tag{S77}$$

Expand the first product in the integral.

$$\begin{aligned} = \frac{r^k}{\Gamma(k)} \left[\prod_{i=1}^{3} \binom{n_i}{m_i}\right] \int_0^\infty & \left(\sum_{j_1=0}^{m_1} \binom{m_1}{j_1} (-1)^{j_1} \left(e^{-x\sigma_1 j_1}\right)\right) \left(\sum_{j_2=0}^{m_2} \binom{m_2}{j_2} (-1)^{j_2} \left(e^{-x\sigma_2 j_2}\right)\right) \\ & \cdot \left(\sum_{j_3=0}^{m_3} \binom{m_3}{j_3} (-1)^{j_3} \left(e^{-x\sigma_3 j_3}\right)\right) \left[\prod_{i=1}^{3} \left(e^{-x\sigma_i}\right)^{n_i - m_i}\right] x^{k-1} e^{-rx}\, dx \end{aligned} \tag{S78}$$

Combine the summations.

$$= \frac{r^k}{\Gamma(k)} \left[\prod_{i=1}^{3} \binom{n_i}{m_i}\right] \int_0^\infty \left(\sum_{j_1=0}^{m_1} \sum_{j_2=0}^{m_2} \sum_{j_3=0}^{m_3} \binom{m_1}{j_1}\binom{m_2}{j_2}\binom{m_3}{j_3} (-1)^{j_1+j_2+j_3} \cdot \left(e^{-x[\sigma_1 j_1+\sigma_2 j_2+\sigma_3 j_3]}\right)\right) \left[\prod_{i=1}^{3} \left(e^{-x\sigma_i}\right)^{n_i-m_i}\right] x^{k-1}e^{-rx}\, dx \tag{S79}$$

Expand the last product in the integral.

$$= \frac{r^k}{\Gamma(k)} \left[\prod_{i=1}^{3} \binom{n_i}{m_i}\right] \int_0^\infty \left(\sum_{j_1=0}^{m_1} \sum_{j_2=0}^{m_2} \sum_{j_3=0}^{m_3} \binom{m_1}{j_1}\binom{m_2}{j_2}\binom{m_3}{j_3} \big(-1\big)^{j_1+j_2+j_3} \cdot \left(e^{-x[\sigma_1 j_1+\sigma_2 j_2+\sigma_3 j_3]}\right)\right) \left[e^{-x[\sigma_1(n_1-m_1)+\sigma_2(n_2-m_2)+\sigma_3(n_3-m_3)]}\right] x^{k-1}e^{-rx}\, dx \tag{S80}$$

Multiply the exponentials from the integral on the last line.

$$= \frac{r^k}{\Gamma(k)} \left[\prod_{i=1}^{3} \binom{n_i}{m_i}\right] \int_0^\infty \left[\sum_{j_1=0}^{m_1} \sum_{j_2=0}^{m_2} \sum_{j_3=0}^{m_3} \binom{m_1}{j_1}\binom{m_2}{j_2}\binom{m_3}{j_3} (-1)^{j_1+j_2+j_3} \cdot \left(e^{-x[\sigma_1 j_1+\sigma_2 j_2+\sigma_3 j_3]}\right)\right] \left[e^{-x[\sigma_1(n_1-m_1)+\sigma_2(n_2-m_2)+\sigma_3(n_3-m_3)+r]}\right] x^{k-1}\, dx \tag{S81}$$

Multiply the last line of the equation into the summations in the integral.

$$= \frac{r^k}{\Gamma(k)} \left[\prod_{i=1}^{3} \binom{n_i}{m_i}\right] \int_0^\infty \left[\sum_{j_1=0}^{m_1} \sum_{j_2=0}^{m_2} \sum_{j_3=0}^{m_3} \binom{m_1}{j_1}\binom{m_2}{j_2}\binom{m_3}{j_3} (-1)^{j_1+j_2+j_3} \cdot \left(e^{-x[\sigma_1(n_1-m_1+j_1)+\sigma_2(n_2-m_2+j_2)+\sigma_3(n_3-m_3+j_3)+r]}\right) x^{k-1}\right] dx \tag{S82}$$

By Fubini's theorem we pull the summations out of the integral.

$$= \frac{r^k}{\Gamma(k)} \left[\prod_{i=1}^{3} \binom{n_i}{m_i}\right] \left[\sum_{j_1=0}^{m_1} \sum_{j_2=0}^{m_2} \sum_{j_3=0}^{m_3} \binom{m_1}{j_1}\binom{m_2}{j_2}\binom{m_3}{j_3} (-1)^{j_1+j_2+j_3} \cdot \left(\int_0^\infty \left(e^{-x[\sigma_1(n_1-m_1+j_1)+\sigma_2(n_2-m_2+j_2)+\sigma_3(n_3-m_3+j_3)+r]}\right) x^{k-1}\, dx\right)\right] \tag{S83}$$

Then using a substitution as above, and the definition of the gamma function,

$$= \frac{r^k}{\Gamma(k)} \left[\prod_{i=1}^{3} \binom{n_i}{m_i}\right] \left[\sum_{j_1=0}^{m_1} \sum_{j_2=0}^{m_2} \sum_{j_3=0}^{m_3} \binom{m_1}{j_1}\binom{m_2}{j_2}\binom{m_3}{j_3} (-1)^{j_1+j_2+j_3} \cdot \left(\frac{\Gamma(k)}{[\sigma_1(n_1-m_1+j_1)+\sigma_2(n_2-m_2+j_2)+\sigma_3(n_3-m_3+j_3)+r]^k}\right)\right] \tag{S84}$$

We then cancel out the gamma functions, and have a workable solution for the four-person household. It is a bit of a mess to write down, but straightforward to code. In small epidemics, like the four-person household most of the values will be quite small, and the summations collapse fairly quickly.

$$p(n_1, m_1, ..., n_3, m_3) = r^k \left[\prod_{i=1}^{3} \binom{n_i}{m_i}\right] \left[\sum_{j_1=0}^{m_1}\sum_{j_2=0}^{m_2}\sum_{j_3=0}^{m_3} \binom{m_1}{j_1}\binom{m_2}{j_2}\binom{m_3}{j_3} \cdot \left(\frac{(-1)^{j_1+j_2+j_3}}{[\sigma_1(n_1-m_1+j_1)+\sigma_2(n_2-m_2+j_2)+\sigma_3(n_3-m_3+j_3)+r]^k}\right)\right] \tag{S85}$$

Induction for Any Number of Groups

We are able to identify a pattern, and can repeat this procedure for q groups n_1, n_2...n_q, with their respective number of infected individuals m_1, m_2...m_q, and their relative susceptibilities σ_1, σ_2...σ_q.

$$p(n_1, m_1, ..., n_q, m_q) = \int_0^\infty \left[\prod_{i=1}^{q} \binom{n_i}{m_i}\left((1-e^{-x\sigma_i})^{m_i}(e^{-x\sigma_i})^{n_i-m_i}\right)\right] \frac{x^{k-1}e^{-rx}r^k}{\Gamma(k)}\, dx \tag{S86}$$

The solution for q groups is,

$$p(n_1, m_1, ..., n_q, m_q) = r^k \left[\prod_{i=1}^{q} \binom{n_i}{m_i}\right] \cdot \left[\sum_{j_1=0}^{m_1}\cdots\sum_{j_q=0}^{m_q} \left[\prod_{i=1}^{q}\binom{m_i}{j_i}\right]\left(\frac{(-1)^{(\sum_{i=1}^{q} j_i)}}{[\sum_{i=1}^{q}\sigma_i(n_i-m_i+j_i)+r]^k}\right)\right]$$

We have demonstrated the base case fairly exhaustively, so we will now turn our attention to the $q+1$ case. Since $p(n_1, m_1, ..., n_q, m_q)$ holds for $q = 1, 2, 3$, assume for some fixed but arbitrary $q > 1$, $p(n_1, m_1, ..., n_q, m_q)$ is true. So we consider $p(n_1, m_1, ..., n_q, m_q, n_{q+1}, m_{q+1})$.

$$p(n_1, m_1, ..., n_{q+1}, m_{q+1}) = \int_0^\infty \prod_{i=1}^{q+1}\binom{n_i}{m_i} \prod_{i=1}^{q+1}\left((1-e^{-x\sigma_i})^{m_i}(e^{-x\sigma_i})^{n_i-m_i}\right)\frac{x^{k-1}e^{-rx}r^k}{\Gamma(k)}\, dx \tag{S87}$$

Factor out the coefficients.

$$= \frac{r^k}{\Gamma(k)}\left[\prod_{i=1}^{q+1}\binom{n_i}{m_i}\right]\int_0^\infty \prod_{i=1}^{q+1}\left((1-e^{-x\sigma_i})^{m_i}(e^{-x\sigma_i})^{n_i-m_i}\right)x^{k-1}e^{-rx}\, dx \tag{S88}$$

Separate the two terms in the product in the integral.

$$= \frac{r^k}{\Gamma(k)} \left[\prod_{i=1}^{q+1} \binom{n_i}{m_i} \right] \int_0^\infty \prod_{i=1}^{q+1} \left((1 - e^{-x\sigma_i})^{m_i} \right) \prod_{i=1}^{q+1} \left(e^{-x\sigma_i (n_i - m_i)} \right) x^{k-1} e^{-rx} \, dx \tag{S89}$$

Leverage the binomial theorem one final time.

$$= \frac{r^k}{\Gamma(k)} \left[\prod_{i=1}^{q+1} \binom{n_i}{m_i} \right] \int_0^\infty \prod_{i=1}^{q+1} \left(\sum_{j_i=0}^{m_i} \binom{m_i}{j_i} 1^{m_i - j_i} (-e^{-x\sigma_i})^{j_i} \right) \prod_{i=1}^{q+1} \left(e^{-x\sigma_i (n_i - m_i)} \right) x^{k-1} e^{-rx} \, dx \tag{S90}$$

Simplify given the power of one.

$$= \frac{r^k}{\Gamma(k)} \left[\prod_{i=1}^{q+1} \binom{n_i}{m_i} \right] \int_0^\infty \prod_{i=1}^{q+1} \left(\sum_{j_i=0}^{m_i} \binom{m_i}{j_i} \left(- e^{(-x\sigma_i)} \right)^{j_i} \right) \prod_{i=1}^{q+1} \left(e^{-x\sigma_i (n_i - m_i)} \right) x^{k-1} e^{-rx} \, dx \tag{S91}$$

Factor the -1 out of the exponential in the summation in the integral.

$$= \frac{r^k}{\Gamma(k)} \left[\prod_{i=1}^{q+1} \binom{n_i}{m_i} \right] \int_0^\infty \prod_{i=1}^{q+1} \left(\sum_{j_i=0}^{m_i} \binom{m_i}{j_i} (-1)^{j_i} (e^{-x\sigma_i j_i}) \right) \left[\prod_{i=1}^{q+1} e^{-x\sigma_i (n_i - m_i)} \right] x^{k-1} e^{-rx} \, dx \tag{S92}$$

Expand the first product in the integral.

$$\begin{aligned} &= \frac{r^k}{\Gamma(k)} \left[\prod_{i=1}^{q+1} \binom{n_i}{m_i} \right] \\ &\quad \cdot \int_0^\infty \left(\sum_{j_1=0}^{m_1} \binom{m_1}{j_1} (-1)^{j_1} \left(e^{-x\sigma_1 j_1} \right) \right) \cdots \left(\sum_{j_{q+1}=0}^{m_{q+1}} \binom{m_{q+1}}{j_{q+1}} (-1)^{j_{q+1}} \left(e^{-x\sigma_{q+1} j_{q+1}} \right) \right) \\ &\quad \cdot \left[\prod_{i=1}^{q+1} \left(e^{-x\sigma_i (n_i - m_i)} \right) \right] x^{k-1} e^{-rx} \, dx \end{aligned} \tag{S93}$$

Combine the summations.

$$\begin{aligned} = \frac{r^k}{\Gamma(k)} \left[\prod_{i=1}^{q+1} \binom{n_i}{m_i} \right] \int_0^\infty \left[\sum_{j_1=0}^{m_1} \cdots \sum_{j_{q+1}=0}^{m_{q+1}} \left(\prod_{i=1}^{q+1} \binom{m_i}{j_i} (-1)^{j_i} \left(e^{-x[\sigma_i j_i]} \right) \right) \right] \\ \cdot \left[\prod_{i=1}^{q+1} \left(e^{-x\sigma_i (n_i - m_i)} \right) \right] x^{k-1} e^{-rx} \, dx \end{aligned} \tag{S94}$$

Separate the terms in the first product in the integral.

$$\begin{aligned} = \frac{r^k}{\Gamma(k)} \left[\prod_{i=1}^{q+1} \binom{n_i}{m_i} \right] \int_0^\infty \left[\sum_{j_1=0}^{m_1} \cdots \sum_{j_{q+1}=0}^{m_{q+1}} \left(\prod_{i=1}^{q+1} \binom{m_i}{j_i} \right) \left(\prod_{i=1}^{q+1} (-1)^{j_i} \right) \left(\prod_{i=1}^{q+1} (e^{-x[\sigma_i j_i]}) \right) \right] \\ \cdot \left[\prod_{i=1}^{q+1} \left(e^{-x\sigma_i (n_i - m_i)} \right) \right] x^{k-1} e^{-rx} \, dx \end{aligned} \tag{S95}$$

Use the exponent power rule on the several products in the integral.

$$= \frac{r^k}{\Gamma(k)} \left[\prod_{i=1}^{q+1} \binom{n_i}{m_i}\right] \int_0^\infty \left[\sum_{j_1=0}^{m_1} \cdots \sum_{j_{q+1}=0}^{m_{q+1}} \left(\prod_{i=1}^{q+1} \binom{m_i}{j_i}\right)\right)\left((-1)^{\sum_{i=1}^{q+1} j_i}\right)\left(e^{-x[\sum_{i=1}^{q+1} \sigma_i j_i]}\right)\right] \cdot \left[\left(e^{-x[\sum_{i=1}^{q+1} \sigma_i (n_i - m_i)+r]}\right)\right] x^{k-1}\, dx \tag{S96}$$

Multiply the terms in the integral not in the summation into the summation.

$$= \frac{r^k}{\Gamma(k)} \left[\prod_{i=1}^{q+1} \binom{n_i}{m_i}\right] \cdot \int_0^\infty \left[\sum_{j_1=0}^{m_1} \cdots \sum_{j_{q+1}=0}^{m_{q+1}} \left(\prod_{i=1}^{q+1} \binom{m_i}{j_i}\right)\right)\left((-1)^{\sum_{i=1}^{q+1} j_i}\right)\left(e^{-x[\sum_{i=1}^{q+1} \sigma_i (n_i - m_i + j_i)+r]}\right) x^{k-1}\right] dx \tag{S97}$$

Use Fubini's theorem one final time.

$$= \frac{r^k}{\Gamma(k)} \left[\prod_{i=1}^{q+1} \binom{n_i}{m_i}\right] \cdot \left[\sum_{j_1=0}^{m_1} \cdots \sum_{j_{q+1}=0}^{m_{q+1}} \left(\left(\prod_{i=1}^{q+1} \binom{m_i}{j_i}\right)\right)\left((-1)^{\sum_{i=1}^{q+1} j_i}\right) \int_0^\infty \left(e^{-x[\sum_{i=1}^{q+1} \sigma_i (n_i - m_i + j_i)+r]}\right) x^{k-1}\, dx\right)\right] \tag{S98}$$

Perform one final substitution.
Let $v = x(\sum_{i=1}^{q+1} \sigma_i (n_i - m_i + j_i) + r) \rightarrow x = \frac{v}{\sum_{i=1}^{q+1} \sigma_i (n_i - m_i + j_i) + r} \rightarrow dx = \frac{dv}{\sum_{i=1}^{q+1} \sigma_i (n_i - m_i + j_i) + r}$.

$$= \frac{r^k}{\Gamma(k)} \left[\prod_{i=1}^{q+1} \binom{n_i}{m_i}\right] \left[\sum_{j_1=0}^{m_1} \cdots \sum_{j_{q+1}=0}^{m_{q+1}} \left(\left(\prod_{i=1}^{q+1} \binom{m_i}{j_i}\right)\right)\left((-1)^{\sum_{i=1}^{q+1} j_i}\right) \int_0^\infty \left(e^{-v}\right) \left(\frac{v}{\sum_{i=1}^{q+1} \sigma_i (n_i - m_i + j_i) + r}\right)^{k-1} \frac{dv}{\sum_{i=1}^{q+1} \sigma_i (n_i - m_i + j_i) + r}\right] \tag{S99}$$

Combine denominators.

$$= \frac{r^k}{\Gamma(k)} \left[\prod_{i=1}^{q+1} \binom{n_i}{m_i}\right] \left[\sum_{j_1=0}^{m_1} \cdots \sum_{j_{q+1}=0}^{m_{q+1}} \left(\left(\prod_{i=1}^{q+1} \binom{m_i}{j_i}\right)\right)\left((-1)^{\sum_{i=1}^{q+1} j_i}\right) \int_0^\infty \left(e^{-v}\right) (v)^{k-1} \frac{dv}{[\sum_{i=1}^{q+1} \sigma_i (n_i - m_i + j_i) + r]^k}\right)\right] \tag{S100}$$

Factor the coefficient out of the integral.

$$= \frac{r^k}{\Gamma(k)} \left[\prod_{i=1}^{q+1} \binom{n_i}{m_i}\right] \left[\sum_{j_1=0}^{m_1} \cdots \sum_{j_{q+1}=0}^{m_{q+1}} \left(\left(\prod_{i=1}^{q+1} \binom{m_i}{j_i}\right) \right.\right. \\ \left.\left. \frac{\left((-1)^{\sum_{i=1}^{q+1} j_i}\right)}{[\sum_{i=1}^{q+1} \sigma_i(n_i - m_i + j_i) + r]^k} \int_0^\infty \left(e^{-v}\right) (v)^{k-1} \, dv \right)\right] \tag{S101}$$

By the definition of the gamma function,

$$= \frac{r^k}{\Gamma(k)} \left[\prod_{i=1}^{q+1} \binom{n_i}{m_i}\right] \left[\sum_{j_1=0}^{m_1} \cdots \sum_{j_{q+1}=0}^{m_{q+1}} \left(\left(\prod_{i=1}^{q+1} \binom{m_i}{j_i}\right) \frac{\left((-1)^{\sum_{i=1}^{q+1} j_i}\right)\Gamma(k)}{[\sum_{i=1}^{q+1} \sigma_i(n_i - m_i + j_i) + r]^k} \right)\right] \tag{S102}$$

Cancel out the gammas.

$$= r^k \left[\prod_{i=1}^{q+1} \binom{n_i}{m_i}\right] \left[\sum_{j_1=0}^{m_1} \cdots \sum_{j_{q+1}=0}^{m_{q+1}} \left(\left(\prod_{i=1}^{q+1} \binom{m_i}{j_i}\right) \frac{(-1)^{\sum_{i=1}^{q+1} j_i}}{[\sum_{i=1}^{q+1} \sigma_i(n_i - m_i + j_i) + r]^k} \right)\right] \tag{S103}$$

This matches the form from above, so by the principle of mathematic induction we have shown

$$p(n_1, m_1, ..., n_q, m_q) = r^k \left[\prod_{i=1}^{q} \binom{n_i}{m_i}\right] \\ \cdot \left[\sum_{j_1=0}^{m_1} \cdots \sum_{j_q=0}^{m_q} \left[\prod_{i=1}^{q} \binom{m_i}{j_i}\right] \left(\frac{(-1)^{\sum_{i=1}^{q} j_i}}{[\sum_{i=1}^{q} \sigma_i(n_i - m_i + j_i) + r]^k} \right)\right]$$

Chapter 5

Conclusion

Mathematical modeling is valuable tool by which we can test hypotheses and design experiments which would be difficult to practicably examine in a traditional double-blind, randomized experiment. The unending challenge in epidemiology is often just that, developing bodies of work comprehensive enough to push the needle of scientific consensus enough that we might turn the correlational patterns observed into causative explanations. On the math side of things, we struggle with striking a balance to achieve the proper level of granularity while maintaining the assumptions and abstractions upon which the calculations are predicated. As our world becomes ever more data rich, I aim to develop tools to help us utilize said data. I was repeatedly faced with the reality the data collected are seldom precisely what one might want but learned that careful planning and hypothesis testing can give us a hope to answer what are often the most pressing questions. As I developed this dissertation, I was incredibly lucky to have collaborators who trusted me with hard questions and the responsible use of their data.

In this dissertation I have presented three projects which contribute to the collective understanding of infectious disease transmission. In the forecasting chapter, I successfully developed forecasts of COVID-19 hospitalizations three weeks into the future. This approach provides utility in many ways: it will inform public health approaches to forecasting moving forward, I will use it as the foundation for a structured methodology to be refined with access to more data and expanded to other diseases, as I aim to build even better forecasts. In the vaccination and household transmission chapter I built a household simulation model to answer a relatively simple question about the effect of

vaccination on chains of transmission. That chapter inspired deeper questions about the underlying system being explored. The analytic solution for small epidemics provides a tractable approach to estimating final size distributions and secondary transmissions of infection for small epidemic and opens worlds of possibilities for efficiently computing results for multi-compartment models and estimating transmission parameters.