Modeling of Infectious Disease Informed by Investigative Focus and Data Characteristics, From the Microscopic to County Community Level

by

Jess A. Millar

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in the University of Michigan
2022

Doctoral Committee:

Assistant Professor Robert J. Woods, Chair
Professor Betsy Foxman
Associate Professor Peter L. Freddolino
Professor Denise E. Kirschner
Assistant Professor Michael W. Sjoding

Jess A. Millar

jamillar@umich.edu

ORCID iD: 0000-0001-8945-3396

Dedication

This dissertation is dedicated to all doctoral students dealing with long COVID. I hope you all find the skills and support you need to make it through the process and eventually survive on the other side.

Acknowledgements

I would like to thank my research advisor Robert Woods. He has provided me with opportunities to work on incredibly interesting datasets and given feedback on many ideas and helped see my pet project come to fruition. I would also like to acknowledge and thank Chanese Forté for her assistance in editing my thesis and her non-stop support during my most difficult moments in the dissertation process. Her feedback has helped me become a better writer and helped me continue my degree even as my health decided due to long COVID.

I want to thank Denise Kirschner for allowing me to work with her and encouraging me to improve upon established processes in her lab. She really made me aware how much I have to offer and made me feel like my possibilities were endless. Along with Denise, Betsy Foxman provided endless encouragement. I want to thank her for investing in me as a student, giving me experience as a teaching assistant and feedback in many of my endeavors. I also want to thank Arron King for teaching me a wider array of modeling techniques and helping me to become a much better programmer. I would also like to thank others in my dissertation committee: Peter Freddolino and Michael Sjoding. I appreciate their time and effort reviewing my paper and supporting my research and education.

There are many more people that helped me through my journey that I'd like to mention here. 1) Margit Burmeister, Julia Eussen, and Brain Athey in the Department of Bioinformatics, helping me traverse difficult administrative situations and even checking on me periodically with continuing support. 2) Scott Barolo and Kate Hagadone in the Program in Biomedical Sciences,

offering me support during some of the most difficult times during my PhD. 3) Paul Artale and Ethriam Brammer in Rackham, supporting and even encouraging my Diversity, Equity, and Inclusion work and giving me hope for institutional change. 4) Carl Marrs, Dani Koel, and Justin Colacino in the School of Public Health, in various support roles and generally easing the institutional barriers I encountered. 5) Carolina Rojas Ramirez, Finda Williams, Ruchang Diao, and Morgan Lindback in my cohort, for continual support and making sure we all make it through our degrees. Finally, I'd like to thank my parents for being so supportive, especially as everything got so much harder during the pandemic.

This research was supported by R01AI123093 and R01HL110811 awarded in part to Denise Kirschner. Any simulations also use resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. ACI-1053575, the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant ACI-1548562, and National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility operated under Contract No. DE-AC02-05CH11231. Fellowship and partial funding were provided by the National Science Foundation Graduate Research Fellowship Program (DGE-1256260), University of Michigan Rackham Merit Fellowship, University of Michigan Excellence in Basic Science Award, and University of Michigan Benard Maas Fellowship. Outside funding was also provided by the American Society for Microbiology Research Capstone Fellowship, as well as travel funds from University of Washington SISMID Scholarship and University of Michigan Epidemiology Student Travel Grant.

Table of Contents

List of Tables

List of Figures

List of Appendices

Abstract

Infectious disease modeling is a tool to understand disease characteristics of interest, which is often used to make predictions and guide public health policy. These models rely heavily on observational data, which greatly influence what models can be built and their resulting quality. Some model outputs may not have well developed analyses available, obscuring the general approach to answering specific research questions. In this dissertation, I explore the general process of obtaining, handling, and analyzing data, and apply these concepts to several different infectious disease modeling problems. The problems vary in both the type of disease explored and the scale they inhabit, ranging from molecular scale investigations within the host to population level questions.

In chapter 2, I employ host-level modeling to investigate why T cell activation is suppressed in *Mycobacterium tuberculosis* (Mtb) infections. Working at the within-host level, I modify a model of the immune system response to Mtb using an existing agent-based model (ABM). This methodology capitalizes on how ABMs facilitate capturing emergent behavior, in this case the immune system modulated formation of a granuloma to contain the Mtb infection. Our approach allowed us to view interactions of immune cells during granuloma formation and visualize how these interactions affect the ability of T cells to become activated over the span of an infection. I found recruitment of non-specific T cells and granuloma spatial characteristics contributed to crowding out of the few Mtb specific T cells within the granuloma cell, reducing the chances they could interact with and be activated by infected macrophages.

In chapter 3, I explore how factors in hospital active surveillance for vancomycin resistant *Enterococcus* (VRE), such as non-compliance, affect our ability to estimate endemic rates within hospitals. Using electronic health records at an individual based level, I simulated patient infections using an ABM approach to establish a baseline for hospital infection. I modeled different compliance rates and testing strategies using cost-effective analysis to judge which type of surveillance strategy is most effective in identifying cases. Our analysis revealed that increasing the compliance rate of screening under any current active surveillance strategy maximized efficacy of identifying VRE cases.

In chapter 4, I explore risk factors that affected the COVID-19 case fatality rate (CFR) during the first wave of the pandemic. Having only access to county-level data, this population level study used a case-lag adjust, count-based regression approach to explore the relationship between CFR and other county level indicators such as comorbidity rates, healthcare infrastructure capacity, and non-pharmaceutical interventions. As the new disease was not yet endemic at the time of data collection, the case lag adjustment allowed consideration of the case deaths lag to properly estimate the case fatality rate. This study agreed with previous findings, including relationships between increased asthma occurrences and rates of CFR, and contributed new findings on risk factors, including reduced CFR with bans on religious gatherings.

Each infectious disease problem utilized the properties of the data and best-practice methodologies in order to best answer the research question, whether the problems involved exploring disease modeling at a county, institutional, or cell-level scale. These approaches taken to answer scientific questions should be at the forefront of all projects and will be necessary to build large, effective knowledge based from these efforts.

Chapter 1 Introduction

Infectious disease modeling is a tool to understand disease characteristics of interest.[1] These characteristics can be focused at the population level, such as the spread of an epidemic through a community, or focused on an individual, such as disease progression. These tools can be used to predict future outcomes, such as outbreaks, and even to test different policies to curb spread and improve the outcomes of infected individuals. As data types and quantity expand and questions become more complex, policymakers increasingly rely on infectious disease modeling to assess and make decisions on available intervention options.[2–4]

Infectious disease modeling involves mathematical models, which are descriptions of a system or process by using mathematical concepts.[5] These systems or process are related to the elements required to generate phenomena that occur and/or are observed. There are many ways to characterize types of mathematical models, but for the purposes of this dissertation, I will focus on mechanistic and statistical models.[6] Mechanistic models attempt to describe the specific pieces of a process that generate observations and require some sort of insight into parameterizing the underlying process that results in an outcome of interest. An example of mechanistic models are Susceptible Infected Recovered (SIR) models. SIR models rely on the understanding of underlying disease dynamics, using sets of equations to simulate the process of susceptible people becoming infected and recovering. Alternatively, when I don't have insight into the underlying process statistical models are often used. These rely on patterns in observations to approximate outcomes, such as creating a regression to characterize the number

of people currently infected with a disease. The output data of regression would be the distribution of predictions, which can be presented as the point estimate which fits on the regression line, or simulated as a distribution based on the variance of the predictors. All of these mathematical models generate output data, which can be used to make predictions about the real world.

To accurately characterize real-world problems, infectious disease models rely on observational data.[7] These observational data are used to estimate the parameters of the model, such as transmission rate in SIR models or regression coefficients. In some cases, observations that are not currently available were previously used to directly calculate a parameter of interest, such as transmission rate. Since only the estimate is available, this value is directly input into the model. In other cases, parameters are calibrated to minimize the differences between observations and model predictions. The observational data may vary greatly in their characteristics. Some data may be aggregated to a certain community level or only as a categorical response (when it originally existed as a continuous variable), which may reduce signal detection and limit the scale of a mechanistic model. The quality of the observational data also has major impacts, not just in its accuracy, but its consistency, such as frequency of measurement, especially when linking up data from multiple sources.[8–10] Just as important is how representative the data are of the population I would like to generalize and make inferences about. As model parameters rely on observational data, the characteristics of the data greatly influence what models can be built and the resulting quality and generalizability of their predictions.

Once an infectious disease model is built, further analysis is required to gain insight into the problem of interest. All of these models generate output data in the form of predictions, but

the analysis of this data will depend on what type of output data are generated by the model and the specific question being asked. Sometimes these analyses are well established, such as in regression, where the significance of regression coefficients are tested based on the distributions of their fitted parameters. Other types of mathematical models may not have established and agreed upon analysis pipelines, or they may not match up with the type of question being asked, requiring the creation of new analysis methods. Due to the variability in established approaches to different types of data and subsequent analysis, it is useful to characterize the general approach of obtaining, handling, and analyzing data to answer scientific questions. Addressing these problems requires consideration of the type and restraints (such as protected data or something not possible to collect) of data that can be obtained and characteristics and issues found in the investigative focus.

Over the last decade, the term data scientist has been used to describe the role of a researcher that has the skills needed to solve complex problems and the curiosity to explore what problems need to be solved using data. These individuals require skills in mathematics and statistics, computer programing, data engineering, as well as domain knowledge of a given field, or at least enough to collaborate. The process that data scientists take to answering questions involves many moving parts, but can be generally described as data collection, processing, and analysis. In this chapter I cover this process, highlighting examples of the diverse problems tackled during this dissertation.

1.1 Determining goals and data collection

1.1.1 Setting project goals

In general, the first part of any science project involves setting goals, which will eventually be used to data collection. What is the question I want to address? As a data scientist, this may be predefined by a standard operating protocol in the laboratory or can be created ad hoc through an exploratory method. In chapter 2, questions about T-cell activation in *Mycobacterium tuberculosis* (Mtb) infections arise from previous work and exploring relevant literature, while in chapter 3, questions about vancomycin resistant *Enterococcus* (VRE) screening compliance arose out of exploratory analysis of a large existing dataset, uncovering a problem I didn't previously know was there. In chapter 4, the question asking what are risk factors for COVID case fatality, arose from an immediate public health need in the wake of a new emergent disease I had no previous experience with.

1.1.2 Defining variables

After goals are defined, key variables of interest need to be identified. What are the variables you need to answer your question? This may be something that evolves naturally from previous work, or you may have to actively research the subject to get an idea. In the VRE project, variables arose from quirks in data that spurred our question, while in the COVID project started with a previously unknown disease, so variables needed to be extrapolated from similar types of diseases found in the literature, being generous in include a large enough breadth of possible risk factors to explore.

1.1.3 Assessing data collection methods with the analysis in mind

Once you have identified both goals and variables of interest, viable data collection methods need to be assessed and chosen. Consideration should be given to which data will

maximize accuracy and precision and what resources are available. Will you be collecting the data yourself, or be using secondary sources? If you will be collecting your own data, what experimental design will be used? Is it even possible to get the data you need, or will you need to consider a proxy or even a simulation?

There were a number of limitations that helped shaped the choices for data collection in the dissertation projects. In the Mtb project, it wasn't possible to experimentally acquire longitudinal data on numbers of activated T cells in humans, so the main datasets were simulated, however it was calibrated based on other related experimental datasets. The COVID project had a constraint of relying on government datasets as it was certainly not feasible for our team to collect data on a new emerging disease on a national scale, while in the VRE project, data was already collected, but I had the ability to request from a large range of privileged data.

In many cases, as a data scientist you may not have control over the data that are available to you if you are not included in the collection steps, but you can still assess if the data can in fact be used to address the goals of a project. Is the data available on the specific variable of interest, or is it a proxy (which if the proxy differs greatly may require more mechanistic modeling to get at the variable of interest)? Is the data biased in some way that makes it impossible to accurately measure the outcome of interest? In some cases, it may not be possible to answer a question after the fact, so it is important to be a part of the data collection process if possible. Sometimes it's possible and even necessary to adjust questions and goals of your project based on accessible secondary data.

1.2 Data processing

In any project that includes data analysis, after establishing project goals and collecting data, the next step is to process the data. This is a commonly overlooked step, but one of the most significant as a data scientist. Some statisticians claim about 80% of time on a project is used to process the data and about 20% on the actual analysis.[11–13] Processing data generally covers the areas of structuring, quality control, and cleaning.

1.2.1 Structuring data

Structuring data are done to organize them into an appropriate form, either for downstream analysis, or quality checks.[14] This can be done through combining multiple datasets, organizing data into columns and relational tables, as well as transforming existing variables to create new variables of interest. This can also include parsing free text and assigning meaning to codes to facilitate interpretation. In the VRE project, the main variable of interest was compliance with a protocol, was calculated from the data provided (which included tests conducted, previous results, timing of tests, and the location of patients). In the COVID project, deaths tended to lag behind the initial case diagnosis by about three weeks, causing an underestimate of the case fatality rate as the newly emerging disease was not yet endemic. The outcome of interest needed to be transformed, in this case lag adjusted, to account for the lag in case deaths to properly estimate the case fatality rate.

1.2.2 Quality control/assessment

Within quality control are a number of elements, which include consistency, completeness, and accuracy.[14] Consistency considers if formats and meaning of variables are consistent over time. If they are not, these variables may require more restructuring so all variables match the desired format. Completeness considers if I have all the data or if some of them are missing. If data are missing, ways to rectify this through supplemental sources or other means should be pursued. Accuracy considers observational error, which reflects biases that affect data, such as misclassification and recall.[15] Accuracy can also include things such as instrumental error, when an instrument taking measurements is inaccurate due to internal issues such as calibration or faulty use by the operator. Environmental errors are also possible, due to changes in the environment such as temperature, humidity, and other environmental changes. In each of these cases, if there are issues in quality control, that will then lead to data cleaning.

1.2.3 Data Cleaning

Cleaning is often a vague term that can be specific to a number of different aspects. These include making corrections within data or removing low quality data, or can encompass everything in data processing including structuring and quality control.[16,17] In general, some best practices for cleaning data include correcting data at the point of entry and developing a data quality strategy. In this strategy, expectations are set for your data, data quality key performance indicators are created, finding where most data errors occur, and developing a plan to ensure continuing performance of your data (such as when new data streams are added, or formats change over time due to changes in software).[18] All of this connects to the idea of rigor and reproducibility. Rigor in collection, storage and use of data, making it as unbiased as possible and this process should be reproducible, In the COVID project, some variables were missing a

small percentage of values (<2%), so these were interpolated to improve completeness. In the VRE project, date formats in the electronic health records varied over different data pulls and needed to be standardized. This dataset also had a quirk in the data pull, where erroneous dates were added in places that needed to be removed. I knew in this dataset that misclassification did exist in our test outcomes, but since I did not have another means of fixing them, these had to be kept and accounted for later in analyses.

1.3 Data analysis

Once the data have been processed, the data analysis to address the question of the project can begin. The type of analysis that is chosen will depend heavily on the goals previously decided. The different possible approaches to analyzing data are vast and fill many textbooks, which include areas such as reproducibility and transparency, so I will cover only some highlights of types of analysis. In general, analysis will be determined by the project goals. Here, I will break this down based on whether the goal is to describe the data, identify relationships, or make predictions. In many cases, studies will have multiple objectives that result in different outcomes, sometimes encompassing all three areas.

1.3.1 Descriptive statistics

In many studies, descriptive statistics are used to summarize the characteristics of individual variables. This includes summary statistics, such as mean, standard deviations, box plots, histograms, etc. These descriptive statistics can be used to show the outcome of an experiment, or simply to characterize a population that will be studied. They can also be used on other related variables of interest as sanity checks. Part of the VRE project was to show what the

trends in VRE screening compliance were in our hospital historically. For the Mtb project, it was important to see that the distribution of values for immune cells in our simulations matched those done in previous outside collaboration experiments to confirm our simulated model was calibrated correctly.

1.3.2 Relationships

One of the biggest goals found in scientific studies is finding the relationship between two or more variables. The methodologies available are again vast and include methods such as statistical models (T tests, regression, etc) and mechanistic models (agent-based models (ABM), differential equations, etc). The choice in methodology is often influenced by the level of focus on the data generating process, or the process in the real world that generates the data. Does the study rely on understanding the underlying theory of the data generating process (mechanistic), or does the study rely on general observations rather than theory of the underlying data generating process (statistical). Additional methodologies can be stacked on top of these methods to further explore relationships, such as conducting sensitivity analyses on the outputs of ABMs to see what variables most affect certain model outputs. In chapter 4, I used a regression approach to quantify the relationship between possible risk factors and the COVID-19 case fatality rate (CFR). In the Mtb and VRE projects, sensitivity analysis was used to explore how intracellular processes as well as immune cell crowding influences activation of T cells, and look at the influence of VRE screening characteristics on VRE identification, respectively.

1.3.3 Predictions

In some studies, the goal is to predict what will happen in the future, either what will happen if I perturb a system, or let things continue under current conditions. This goal often falls under the term forecasting. In some fields, a distinctive offshoot is called prescriptive modeling, in which the final output includes recommendations in the prediction to optimize the outcome. In the VRE project, I used a cost effectiveness analysis, which can be considered a form of prescriptive modeling. Using many different simulated outcomes, I were able to predict which VRE screening strategies would optimize our output, in this case detecting VRE cases within our hospital.

1.4 Summary

In this dissertation, I walk the reader through three case studies where I have apply these techniques to real-world problems from across the fields of bioinformatics and infectious disease modeling. As the aforementioned examples allude, this dissertation takes a deep dive into vancomycin resistant Enterococcus (VRE) hospital surveillance, Mycobacterium tuberculosis (Mtb) infection, and COVID-19 case fatality rate (CFR).

In chapter 2, I look into why T-cell activation is lower than expected within the lungs during Mtb infections. Working at the host level, this project approaches modeling the immune system responds using an ABM. This takes advantage of how ABMs facilitate capturing emergent behavior, in this case the formation of a granuloma to contain the infection. This allowed us to view the interactions of immune cells in the granuloma and see how this interaction affects the ability of T cells to become activated over the span of an infection. I find that the spatial characteristics of a granuloma contribute to crowding out T cells, reducing the chances they could be activated by an infected macrophage.

In chapter 3, I explore how compliance and other factors of a hospital's active surveillance protocol for VRE affect our ability to estimate the endemic rates within hospitals. Having access to electronic health records at an individual level, allowed me to simulate patient infections using an agent-based modeling (ABM) approach. Different compliance rates and testing strategies are simulated and cost-effective analysis is used to judge the type of surveillance strategy allows for the most effect in identifying cases. This allowed us to sort out that increasing the compliance rate of screening under any surveillance strategy was most effective at identifying cases.

In chapter 4, I explore risk factors that affected the COVID-19 CFR during the first wave of the pandemic. Having only access to data at the county level, this population level study used a case lag adjusted count-based regression approach to explore the relationship between CFR and other county level indicators, ranging from comorbidity rates, to healthcare infrastructure capacity, to non-pharmaceutical interventions. As the disease was newly emerging and not yet endemic, the case lag adjustment allowed us to account for the lag in case deaths to properly estimate the case fatality rate. This study agreed with previous findings such as asthma rates increasing CFR and contributed new risk factors, such as banning indoor religious gatherings lowering CFR.

Whether the problem at hand was exploratory, investigating effectiveness, or modeling the inner workings of a process, each infectious disease problem utilizes the properties of data collection, processing and analysis to best answer the problem.

Chapter 2 Spatial Organization and Recruitment of Non-Specific T Cells May Limit T Cell-Macrophage Interactions Within *Mycobacterium tuberculosis* Granulomas

This chapter is a published work:

Millar JA, Butler JR, Evans S, Matilla J, Linderman JJ, Flynn JL, Kirschner DE. (2021). Spatial organization and recruitment of non-specific T cells may limit T cell-macrophage interactions within Mycobacterium tuberculosis granulomas. Front Immunol. 11:613638.

2.1 Abstract

Tuberculosis (TB) is a worldwide health problem; successful interventions such as vaccines and treatment require a 2better understanding of the immune response to infection with *Mycobacterium tuberculosis* (Mtb). In many infectious diseases, pathogen-specific T cells that are recruited to infection sites are highly responsive and clear infection. Yet in the case of infection with Mtb, most individuals are unable to clear infection leading to either an asymptomatically controlled latent infection (the majority) or active disease (roughly 5%–10% of infections). The hallmark of Mtb infection is the recruitment of immune cells to lungs leading to development of multiple lung granulomas. Non-human primate models of TB indicate that on average <10% of T cells within granulomas are Mtb-responsive in terms of cytokine production. The reason for this reduced responsiveness is unknown and it may be at the core of why humans typically are unable to clear Mtb infection. There are a number of hypotheses as to why this reduced responsiveness may occur, including T cell exhaustion, direct downregulation of antigen

presentation by Mtb within infected macrophages, the spatial organization of the granuloma itself, and/or recruitment of non-Mtb-specific T cells to lungs. We use a systems biology approach pairing data and modeling to dissect three of these hypotheses. We find that the structural organization of granulomas as well as recruitment of non-specific T cells likely contribute to reduced responsiveness.

2.2 Introduction

Tuberculosis (TB) is caused by infection with *Mycobacterium tuberculosis* (Mtb). It is one of the leading causes of death due to infectious disease, killing 1.7 million people per year.[19] The pathologic hallmark of this infection is the formation of lung granulomas, which are collections of host immune cells (e.g. macrophages & T lymphocytes) that organize in an attempt to contain and eliminate the infection.[20–22] Although bacterial infection preferentially occurs within macrophages, T cells are key players in the proper functioning of granulomas, and are necessary for macrophage activation.[20,23–25]

T cells play a central role in the host adaptive immune response. CD4+ T cells are activated by binding MHC class II (MHCII) complexes on the surface of antigen presenting cells like macrophages. CD4+ T cells provide help for CD8+ T cells and once activated, both CD4+ and CD8+ T cells serve a number of immune roles such as cytotoxic function, regulatory function, and cytokine production, (e.g. interferon-gamma (IFN-γ) and TNF) that recruit other immune cells and activate macrophages.[26–29] Activated macrophages kill Mtb and also produce cytokines and chemokines that recruit other immune cells.[20,30,31] Mtb-specific T cells play an important role in controlling Mtb infection by influencing the initiation and maintenance of the adaptive immune response, leading to formation of lung granulomas.[32,33] T cells have been

shown to be necessary for control of Mtb infection in studies in non-human primates (NHPs) and mice,[34–38] and also from studies from humans who are co-infected with HIV-1 and do much worse. Since granulomas are the infection sites within lungs and provide the potential for frequent interactions between Mtb and host immune cells, we expect them to be enriched in Mtb-responsive T cells (i.e. producing cytokines in response to Mtb). Surprisingly, it has been observed that in granulomas from non-human primates, on average <10% of T cells are producing canonical T cell cytokines (IFN-γ, TNF, IL-2, IL-17, or IL-10) throughout the course of Mtb infection.[39] This low level of cytokine-producing T cells could be one explanation for how granulomas balance excessive inflammation with bacterial control. Regardless, since 2 billion people in the world are infected with TB, it is useful to understand this delicate balance of T-cell responsiveness and why the frequencies of cytokine-producing T cells in granulomas are lower than expected.

There are a few lines of thinking that have been explored to date to explain these observed low levels of Mtb-responsive T cells observed during infection. One hypothesis is that T cells may become exhausted during Mtb infection, as exhausted T cells have been described in other chronic infectious diseases.[40–43] However, we have shown through both experimental and computational work that T cell exhaustion is limited in most NHP TB granulomas.[44] A second hypothesis is that T cells are down-regulated directly by the action of Mtb. Mtb's role in regulating parts of the immune system has been established in studies involving Mtb-derived glycolipids inhibiting pathways in antigen presentation.[45–49] Downstream, this would lead to reduced stimulation of T cells. A third hypothesis is that the spatial organization of granulomas affects the ability of T cells to reach macrophages and thus be activated *via* antigen presentation.[50–52] The structural organization of granulomas tends toward a typical pattern: Mtb

are mostly found within the caseous necrotic core or in epithelioid macrophages adjacent to the core of granulomas, which is then surrounded by layers of macrophages and lymphocytes.[53] We provided evidence that T cells had a higher likelihood of exhaustion after penetrating deeper into the granuloma where they could encounter Mtb antigen, but this penetration of T cells occurs infrequently in established granulomas.[44] Compounding T cell-macrophage interaction dynamics is the recruitment of T cells into granulomas. T cells localize to and are rapidly recruited into mycobacterial granulomas in the absence of antigen recognition.[54–56] If the majority of T cells recruited are Mtb non-specific, Mtb-specific T cells would be less likely to find macrophages and become fully activated due to crowding. Thus, a fourth hypothesis is that non-specific T cells are recruited to granulomas. During chronic infections, there are ongoing signals that can recruit non-specific T cells into lungs due to both the inflammatory nature of granulomas and also the highly vascularized lung environment. In addition, it has also been shown that 90% of non-Mtb-specific T cells are lung tissue resident memory T cells.[57,58] Here we test hypotheses to determine the potential contribution of Mtb modulation, granuloma spatial organization, and T cell recruitment. Our goal is to determine how these factors contribute, either alone or together, to the relatively low levels of observed cytokine-producing T cells established within granulomas during Mtb-infected NHPs.

To address these studies, we need an approach that can explore and compare these hypotheses. The spatial organization of granulomas is crucial to outcomes, as has been suggested in NHPs, mice and rabbit studies.[54,59,60] In addition, temporal dynamics are important, tracking discrete cells and bacteria as they evolve over the course of granuloma formation and maintenance. Finally, events that participate in the immune response to Mtb occur over biological scales ranging from molecules to cells to tissue. Thus, our approach must

accommodate all of these features. To this end, we use a systems biology approach, pairing computational multi-scale modeling with experimental studies in NHPs. Our lab has previously created a multi-scale (intracellular through tissue scales) agent-based model (ABM), *GranSim*, that tracks bacteria and individual immune cells as agents.[44,50,51,61–63] This model captures the host response to Mtb and allows spatial tracking of granuloma formation and function. It also tracks bacterial heterogeneity in terms of growth and division by following each individual bacterium within its micro-environments (intracellular, extracellular and trapped within caseum) over time. Using an agent-based model has the additional advantage that it can capture emergent behavior (in this case, the formation of the granuloma) through rules governing immune cell interaction. Herein, we modify *GranSim* to include an additional sub-model that tracks intracellular-level dynamics of macrophage antigen presentation to examine the impact of Mtb on antigen presentation and thus to T cell outcomes within granulomas. To do this, we integrate our previously published model of Mtb-mediated down-regulation of MHCII presentation of peptides[64,65] within each macrophage in *GranSim*. This will allow us to explore mechanisms of Mtb downregulation of antigen presentation on T-cell responsiveness. At the same time, this multi-scale model can aid understanding of how granuloma structure impacts macrophage and T cell dynamics and also how recruitment to lung granulomas balances T cell specificity/non-specificity. We pair our modeling studies with datasets from NHP granulomas to calibrate and validate our models and predictions.

2.3 Methods

2.3.1 Immunohistochemistry and imaging

Four randomly selected, formalin fixed paraffin embedded (FFPE) granulomas were derived from 3 cynomolgus macaques (Macaca fascicularis), necropsied at approx. 10-11 weeks post infection (Figures 2.1A–D), and were deparaffinized and antigen retrieval was performed as previously indicated.[53] Granulomas were stained with cocktails of antibodies including polyclonal rabbit anti-CD3 (Agilent Technologies, Santa Clara, CA), IgG2a mouse anti-CD11c (clone 5D11; Leica Microsystems, Buffalo Grove, IL). Primary antibodies were labeled with fluorochrome-labeled secondaries including anti-isotype (IgG2a) specific antibodies (Jackson ImmunoResearch, West Grove, PA). Coverslips were mounted with Prolong Gold with DAPI (ThermoFisher Scientific) and the sections were imaged on an Olympus FluoView confocal microscope (Center Valley, PA) or Nikon e1000 epifluorescence microscope (Nikon Instruments, Melville, NY) with Nikon NIS Elements (Nikon Instruments).

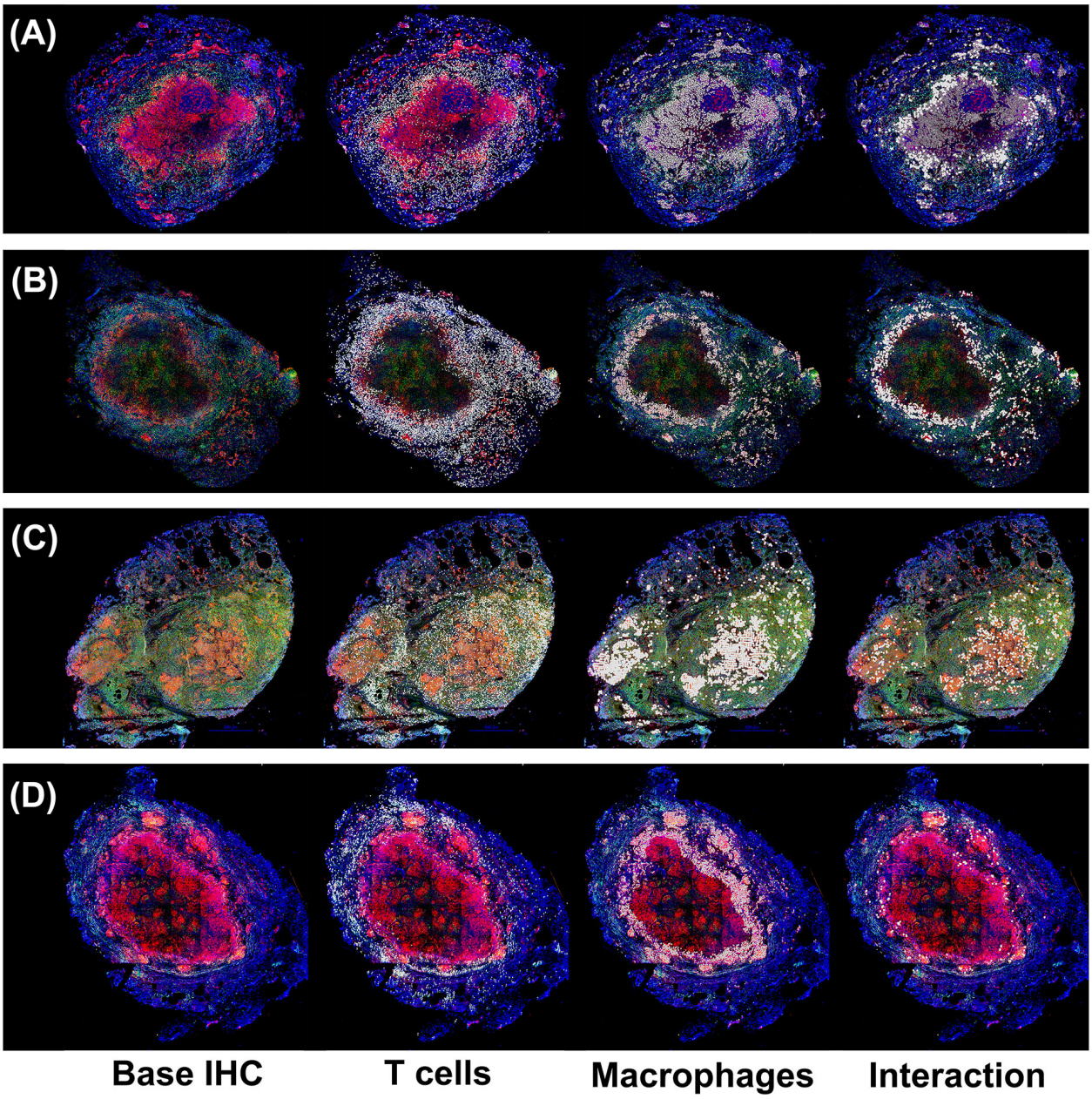

Figure 2.1 Immunohistochemistry analysis of four non-human primates (NHP) granulomas. [Shown in Panels (A-D)] examining spatial distributions of both T cells and macrophages, and also where they intersect. Four distinct, randomly chosen granuloma images with extracted cell distributions. Column 1 shows the immunohistochemically stained preparation for CD3 (green), CD11c+ macrophages (red), and nuclei (dark blue). White points represent Geographical Information Systems Technology (GIS) analyses of these images revealing cell locations for T cells (Column 2), macrophages (Column 3), and their intersections (Column 4), as follows. Rows represent four distinct granulomas. The data for the cell numbers in these granulomas are given in Table 2.2. On average, about 9.75% (median 8.6%, StDev is 4.5%) of T cells interacted with macrophages.

2.3.2 Geographical information systems to extract T cell-macrophage interactions in a granulomas

Granulomas were obtained and stained as described above. DAPI stained images were provided together with the IHC. We applied an unsupervised classification, iso-cluster, image-classification process to four randomly selected, original immunohistochemically, stained NHP granuloma digital microscopy images. This initial classification technique generated between 28 and 45 classifications. Classes correspond to different cell types, even portions of cells like cell borders, caseum, cellular debris, sample background, and co-expression. These initial classifications were collapsed to the single cell types of interest. For accuracy assessment, the classified image was superimposed onto the DAPI image to ensure that cell location and size were correct. In addition, this process removed cellular debris versus true cells. These classified raster images, where the objects in the image are defined by individual pixels instead of vectors, were then converted to vector-based or polygons and the polygons were then assessed for classification errors. The two different polygon images were subjected to a join technique that recorded the locations where different polygons intersected. From the classification process we created a raster, classified image. This raster image was converted to a vector (polygons) image (ArcPro 2.6). We extracted individual cell distributions by cell type (T cells and macrophages only). We then performed a spatial join (ArcPro 2.6) between the T cell/macrophage distributions based on cell-cell interactions to determine overlap and/or border interactions. The cells were marked on the images, and the numbers of each type together with the interactions were quantified.

2.3.3 Multi-scale model overview

To test our three hypotheses and address how bacterial factors, granuloma spatial organization and T cell recruitment lead to reduced T-cell responsiveness, we create a next-generation computational model. Briefly, the main model (mesoscale) operates at the cellular/tissue scales, tracking host immune cells and individual Mtb in the immune system environment, leading to granuloma formation. *GranSim* is an agent-based model (ABM) drawing on well-described cellular and pro- and anti- inflammatory cytokine interactions that is continuously updated and curated with the latest data. These dynamics are all captured between immune cells and individual Mtb using stochastic simulations, operating in two dimensions (2D) [with versions working in three dimensions, but not used here[66]]. We now link an intracellular scale sub-model, capturing MHCII processing and presentation by macrophages using a system of ordinary differential equations (ODEs) previously described.[64] The cellular/tissue scales and intracellular scale sub-model are linked through the processes of IFN-γ receptor ligand binding, Mtb antigens, and MHCII Mtb-peptide complexes on macrophage surfaces to activate T cells (Figure 2.2).

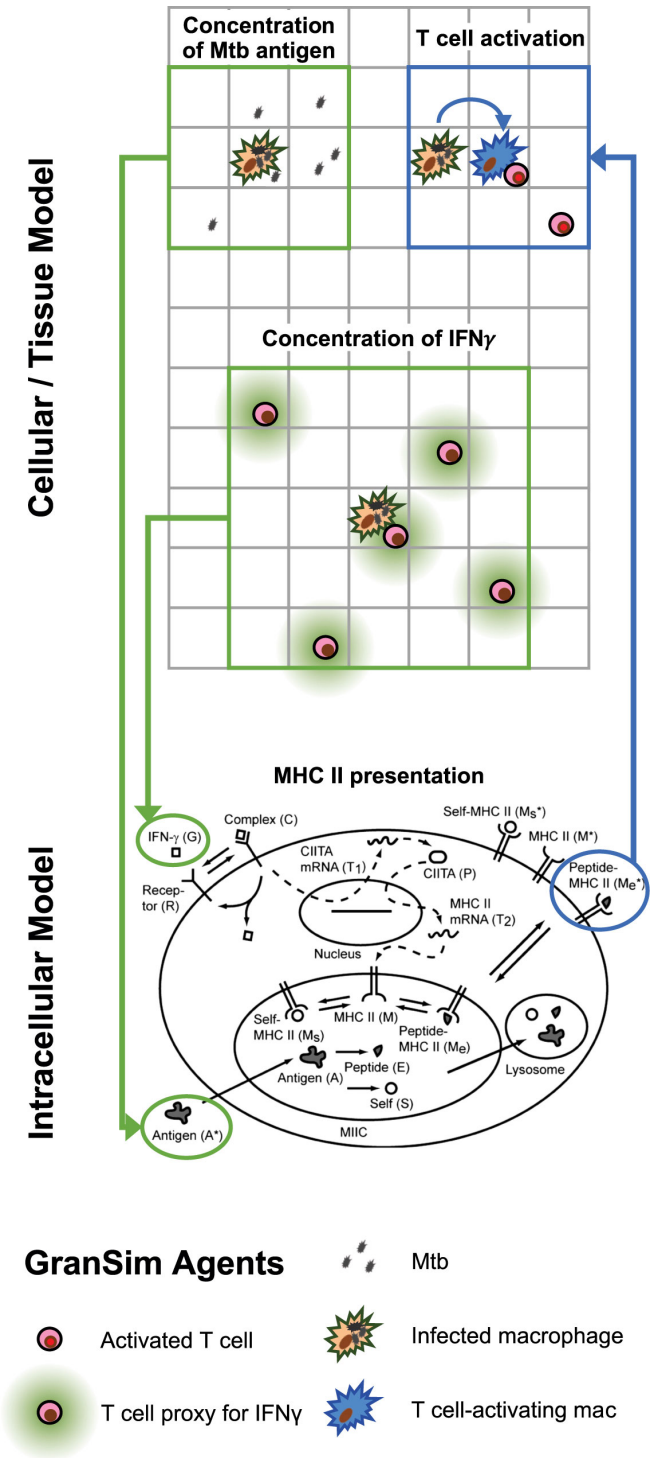

Figure 2.2 Multi-scale model schematic showing the integration of the intracellular scale model into the mesoscale cellular/tissue model. Within our combined multi-scale model, the cellular/tissue scale is modeled with GranSim, an agent-based model capturing dynamics of immune cells, effector molecules and mycobacteria within lungs leading to formation of granulomas. Full cellular and molecule dynamics are not shown for GranSim, only the places where the intracellular model links with GranSim. For the intracellular model, 14 non-linear

ordinary differential equations (ODEs) are represented by this schematic for 14 variables (published previously and listed in Table A.1 for reference here). GranSim linking is accomplished via inputs to ODEs (green arrows) including the concentration of free IFN-γ (calculated based on T cell numbers) and concentration of free Mtb antigen (calculated using the number of Mtb in a macrophage's one Moore neighborhood). The ODE output of MHCII-Mtb complexes on a macrophage surface is linked back into GranSim (blue arrow). Once these macrophages reach a threshold of surface MHCII-Mtb complexes, they are able to activate T cells within their neighborhood.

2.3.4 Cellular/tissue scale model

Hybrid multi-scale model (*GranSim*)

In this work we build an antigen presentation model into our existing hybrid multi-scale agent-based model (ABM) of granuloma formation, *GranSim*. *GranSim* has been curated and used for testing hypotheses in TB since 2004. The model has been developed in conjunction with extensive experimental datasets regarding the immune response to *M. tuberculosis* within the lungs of non-human primates (NHP), leading to the formation of granulomas[50,51] (see our detailed website http://malthus.micro.med.umich.edu/GranSim for full model details, all published manuscripts using this model and an executable program). *GranSim* tracks the cellular immune response in lungs following infection with Mtb that ultimately leads to emergence of a granuloma (if the initial infection is not cleared). *GranSim* is an agent-based (individual-based) model that is comprised of five features:

Agents

Immune cells that are individually tracked as follows: four macrophage states (resting, activated, infected, and chronically infected- see below for more details), three T cell classes (cytotoxic, IFNγ producing, and regulatory), and cytokines and chemokines IFN-γ, TGF, IL10, TNF, CCL2, CXCL9, and CCL5. In addition, individual bacterium are each tracked and are in

one of three environments leading to different growth states (intracellular, extracellular and non-replicating trapped in caseum).

The environment

The model environment represents a section of lung tissue that is 4 mm x 4 mm in size, allowing for granulomas to grow to a size that is average of what is observed *in vivo*. The grid is 2D (although we have a 3D version available- see http://malthus.micro.med.umich.edu/3D-GranSim/) and the model grid is subdivided into 20 micron x 20 micron microcompartments. 20 microns is the average size of our largest immune cell class, macrophages. The lung grid also is populated with blood vessels that are placed on the grade based on NHP studies of healthy lungs. These portals are where cells, chemokines, or cytokines can enter the lung space.

Rules

Rules are based on probabilistic interactions between cells and the lung environment, derived and validated on extensive datasets of observed interactions of NHP immune cells and molecules. The list of rules is extensive and it is housed on our *GranSim* website.

Parameters

GranSim is parameterized by dozens of parameters that have been estimated on datasets over the past 15 years. Further, we study their values and impacts using both uncertainty and sensitivity analyses. The last piece of an ABM is to define the time steps of the fastest process occurring on the grid. Here, that process is molecular diffusion, which is on a time scale of 6 s. When simulated, *GranSim* computes rules and agents at the cell and molecular scale and leads to emergent behavior of a granuloma that reads out at the tissue scale. For an executable file and detailed model rules, please see our website which is continuously curated on a regular basis http://malthus.micro.med.umich.edu/GranSim.

2.3.5 Intracellular MHCII presentation model

Antigen presentation occurs when an antigen presenting cell presents foreign antigenic peptides to T cells to activate an adaptive immune response. This process of antigen presentation occurs within lymph nodes on a continual basis and in other locations during infection. Within granulomas, activation of CD4+ T helper cells depends on presentation of Mtb derived peptide–MHC class II complexes (pMHCII) presented on the surface of macrophages.[31,67] T cell activation is required for granulomas to control infection, as it induces IFN-γ secretion, which, in conjunction with other factors, activates macrophages to kill Mtb.[31] This is the key step in Mtb cell-mediated immunity and may help determine the outcome of infection.[67] In baseline *GranSim*, T cells are recruited to the site of the infection as already activated, with their specificity based on parameter probabilities. This tended to overshoot the proportion of activated T cells as we identified previously in Figure A.1.[39]

Previously, we created a model that describes MHC class II-mediated antigen presentation by antigen-presenting cells.[64] This model is developed as an intracellular scale model representing a single antigen presenting cell (e.g. a macrophage). The model comprises 14 ODEs and was created in the context of capturing datasets from multiple in vitro studies. The model includes all of the intracellular events occurring during the process of antigen presentation: INFγ-receptor ligand binding, leading to MHC class II transcription through CIITA and uptake of Mtb antigens and the creation of host "self"-peptides, both leading to MHCII peptide loading and expression on macrophage surface [see Table A.1 for a full list of variables and ODE equations from Chang et al.[64]]. This model was simulated over short time scales as the process of antigen presentation, and the *in vitro* studies that were used to develop this model,

occurred on a time scale of less than 100 h. This fast time-scale model is linked within macrophages in our longer time-scale model, GranSim, that represents approximately 1-year post-infection. This creates a hybrid model that crosses space and time scales ranging from intracellular to tissue and from minutes to months. The inclusion of these intracellular dynamics calibrates the proportion of activated T cells in GranSim to match T cell levels that we identified previously in Figure A.1.[39] Below we describe how we linked these two model frameworks.

2.3.6 Linking models

There is currently no standard way to link different models, particularly models that are created with different formulations (ODEs, ABMs, PDEs, etc.).[68] We connect the intracellular antigen presentation scale model to the cellular/tissue scale model, *GranSim*, in three ways. Figure 2.2 shows how the two models are linked, and Table 2.1 shows two linking equations and corresponding parameters for those processes.

(1) MHCII transcription depends initially on IFN-γ- derived from CD4+ helper T cells binding to macrophages.[68] Levels of binding are thus controlled by the presence of T cells near a macrophage. To link the ODEs and ABM (intracellular to cellular), we determine this concentration by calculating the number of T cells present in the neighborhood of a macrophage at any given time. A neighborhood is defined as a Moore neighborhood (nine grid squares) or the two-Moore shell *neighborhood* (16 grid squares). Macrophages can sample antigen from at least a two-Moore neighborhood (see Table 2.1 for values and equations and calculations below).

(2) The activation of T cells (CD4+ T helper cells) depends on the presentation of Mtb peptide–MHCII complexes on macrophage surface. The level of complexes seen on

macrophage surface depends on the concentration of Mtb antigens present in the surrounding medium.[68] Mtb produces a variety of protein and glycolipid antigens. Glycopeptidolipid antigen are some of the most persistent, with only 1%–3% degradation after four days.[69] To simulate Mtb antigens, we use dead Mtb as a proxy (dead Mtb are generated directly by macrophages killing them or indirectly by cytotoxic T cells killing infected macrophages). We calculate this concentration directly in *GranSim* by calculating how many dead Mtb are present in the neighborhood of a macrophage (see Table 2.1, and calculations below).

(3) To analyze activation levels of T cells, we define how macrophages interact with T cells at given time and space points. A threshold number of Mtb peptide–MHCII on a macrophage surface are needed to activate T cells, and we set a binding threshold of 120.[70,71] Only macrophages that meet this threshold and are not chronically infected have an ability to activate a T cell (this includes currently infected macrophages and previously infected that cleared their bacterial load). We connect the output of an ODE representing the number of Mtb-MHCII complexes on the surface of a macrophage (Table A.1, variable M_S^*) to each individual macrophage within *GranSim*. For a macrophage to stimulate a T cell, it must be within a one-Moore neighborhood of a macrophage that has a number of Mtb-MHCII complexes that surpasses the binding threshold.

Linking Equation description	Equation
IFN-γ Receptor Ligand Binding	
Molar concentration of IFN-γ in a Moore neighborhood of radius 2 around the macrophage (G)	$(\#.\text{ of } \gamma \text{ T cells in 1 Moore Neighborhood}) \cdot l_{\text{IFN}-\gamma-\text{MN1}} + (\#.\text{ of } \gamma \text{ T cells in Moore Shell Radius 2}) \cdot l_{\text{IFN}-\gamma-\text{MN2}}$
Mtb Antigens	
Molar concentration of Mtb lipid antigen in a Moore neighborhood of radius 1 around the macrophage (A*)	$(\#.\text{ of bacteria in 1 Moore Neighborhood}) \cdot l_{\text{Mtb}-\text{MN1}}$

Linking Parameter	Parameter description	Value
$l_{\text{IFN}-\gamma-\text{MN1}}$	Molar concentration of IFN-γ in a 1 Moore neighborhood around a macrophage	$2.0 \cdot 10^{-9}\text{M}$
$l_{\text{IFN}-\gamma-\text{MN2}}$	Molar concentration of IFN-γ in a Moore shell of radius 2 around a macrophage	$5.7 \cdot 10^{-10}\text{M}$
$l_{\text{Mtb}-\text{MN1}}$	Molar concentration of Mtb lipid antigen in a 1 Moore neighborhood around a macrophage	$4.5 \cdot 10^{-8}\text{M}$
$l_{\text{MHC}-\text{bind}}$	Number of surface MHCII Mtb-peptide complexes required for binding T cells	1.2×10^{2}

Table 2.1 Equations and parameters for the new linking that are needed to combine the intracellular-scale model and *GranSim*.

2.3.7 Linking equation calculations

First, we calculate the *production of free, extracellular IFN-γ*. We use the following translations to perform the calculation in Table 2.1. IFN-γ produced per T cell: 0.0001 U;[72] IFN-γ molecular weight (mature dimer, biologically active): 34 kDa;[73] Volume of 1 grid cell: 8.0×10^{-12} L; Number of compartments in 2-Moore shell of radius: 16; IFN-γ U to μg: $2.x10^4$ U = 1 μg.[74] Therefore,

(1) IFN-γ produced/T cell: $1.5 \cdot 10^{19}$ *mol*

(2) IFN-γ produced/T cell in one compartment: $1.8 \cdot 10^{-8} M$

Thus, the Molar concentration of IFN-γ in a 2-Moore neighborhood (*G*), (where Tg is an IFN-γ -producing T cell), is:

(3) $G = (\#.\ of\ Tg) \cdot 2.0 \cdot 10^{-9} M + (\#.\ of\ Tg)\ 5.7 \cdot 10^{-10} M$

Similarly, we calculate *production of extracellular Mtb antigens*: Mtb Antigen mature weight (approx., range 10–40kDa): ~20 kDa;[75,76] Volume of 1 grid cell: 8.0×10^{-12} L; Number of compartments in 1 Moore neighborhood: 9; Mtb biomass: 1.96×10^{-13} g;[70,77] and Fraction of Mtb that are lipids: 0.33.[77]

(4) Approx. lipid antigens produced/dead bacteria: $3.2 \cdot 10^{-18} mol$

(5) Approx. lipid antigens produced/dead bacteria in one compartment: $4.0 \cdot 10^{-7} M$

Therefore, the Molar concentration of Mtb lipid antigen in a 1-Moore neighborhood is:

(6) $A^* = (\#.\ of\ bacteria) \cdot 4.5 \cdot 10^{-8} M$

2.3.8 Parameter estimation: Literature and uncertainty and sensitivity analysis

GranSim parameter values were estimated from literature [described in detail elsewhere[51,61,62,78–80]]. For the intracellular model, ODE parameter and initial condition values were also estimated from the literature [Chang et al.[64]. and shown again in Tables A.2 and A.3]. Linking parameters are calculated as shown above. If data were not available, we implemented uncertainty analysis using a Latin hypercube sampling scheme (LHS) [reviewed elsewhere[81,82]]. We use LHS to sample and find parameters for the ODEs that represent the dynamics within the macrophages (intracellular) and also to calibrate *GranSim* to experimental datasets. Extensive data on numbers of macrophages, Mtb, and T cells were provided by the Flynn lab as previously described.[39,51,61,67,83] To narrow down parameter ranges and mechanisms of interest, we identify critical parameters that map to specific model mechanisms that impact model outputs. To do this, we take a two-step process: we pair LHS with Partial rank correlations (PRC) analysis (sensitivity), which allows us to quantify the correlation of model outputs with parameters, including those with non-linear relationships.[81] We do this by calculating partial rank correlation

coefficients (PRCCs) that are between -1 and 1 and indicate the strength of the correlations. These are nonlinear correlations, so that is why they are ranked. PRCCs p-values were corrected for multiple testing using Bonferroni.[84]

2.3.9 Mtb-mediated inhibition of antigen presentation

Mtb may inhibit MHCII Mtb-antigen presentation within macrophages by interfering with MHCII mRNA transcription, antigen processing, MHCII maturation, and/or MHCII peptide loading.[48,49,64] To determine if these fast-time and short physiological scale events manifested at the granuloma scale, we tested the effects of inhibiting these four processes on MHCII antigen-presentation on downstream T cell activation. To do this, we examined a range of rates of down regulation of antigen processing (Table A.4) to be used with Michaelis–Menten dynamics (Table A.5), as done previously in.[64]

2.3.10 T cell spatial characteristics

Based on previous work, we have identified that the spatial organization of granulomas can be a determinant in granuloma outcomes.[52] Thus, we separately explored the spatial organization of granulomas and the role that may play in reducing the number of Mtb-responsive T cells. To do this, we use the combined ODE-ABM multi-scale model with no inhibition of antigen-presentation processes. From these 500 simulations, 37 scenarios were removed as the bacterial infection did not occur or did not generate T cells within the first 50 days, leaving 463 simulated granulomas to be analyzed. (Results using similar runs gave similar results.) We chose a replicate run of *GranSim* that fit the median characteristics of the 463 *GranSim* simulations

over the time scale of 200 days post infection (Figure A.2). These characteristics include: numbers of macrophages, Mtb counts, proportion of activated T cells, and average presentation of antigenic peptide–MHCII complexes on macrophage surface (see example median runs chosen based on these characteristics in Figure 2.3). We calculated minimum, average, and maximum distance of immune cell populations based on their relation to the granuloma center of mass [as in Renardy et al.[85]].

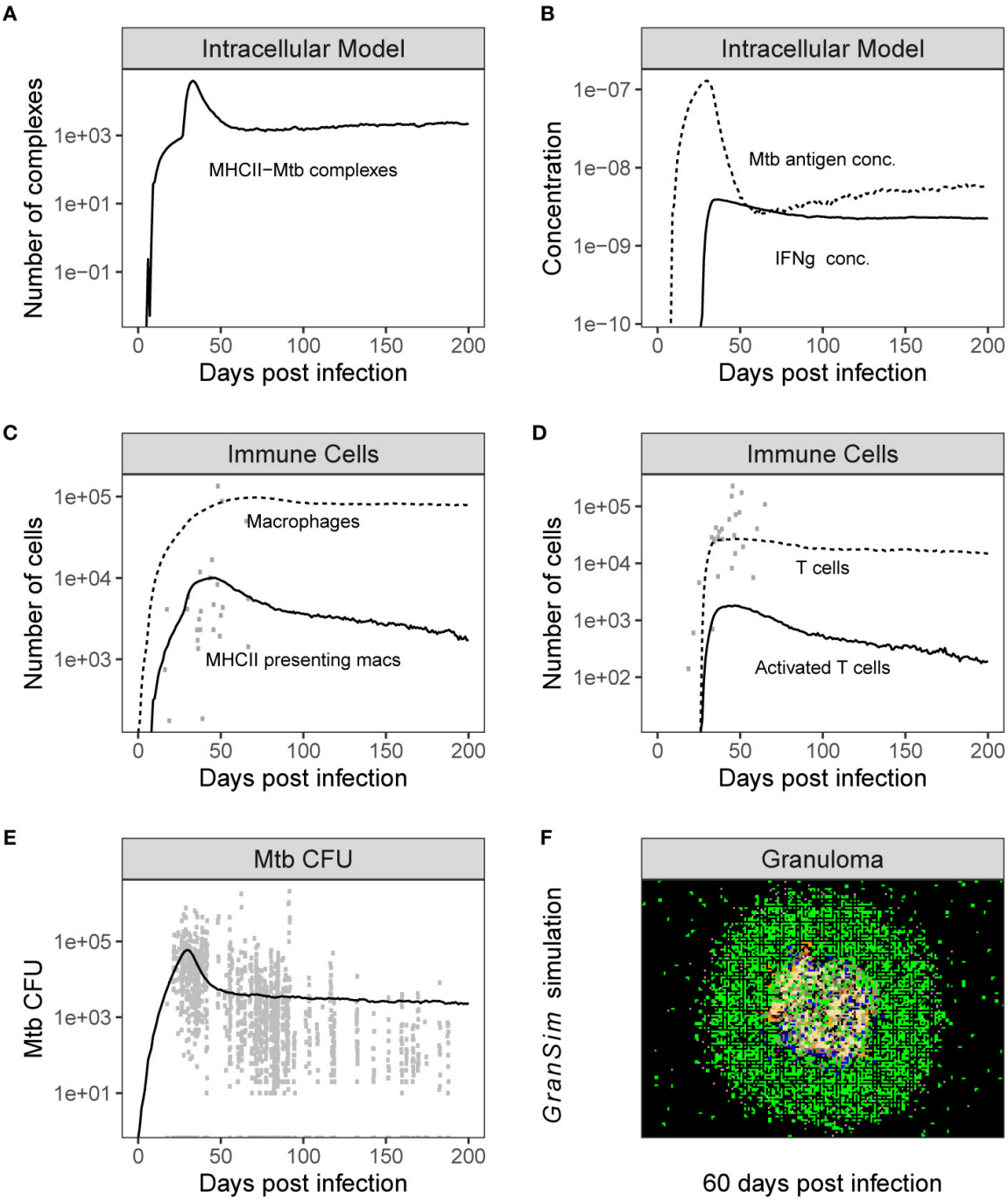

Figure 2.3 Simulation *GranSim* output is consistent with NHP data. Time series plots showing the dynamics of both individual model outputs from the combined Multi-scale model. Panels (A, B) shows the variables in the intracellular model over time, and Panel (C, D) show the time series of total populations of immune cells [shown together with non-human primate (NHP) datasets from Wessler et al.[86]]. Panel (E) shows the CFU [shown together with NHP data sets from Wessler et al.[66]] and Panel (F) is a time series snapshot at day 60 of the Multi-scale GranSim granuloma model, 2x2 mm scale. Cell types in Panel (F): Macrophages: resting (green),

active (blue), infected (orange), chronically infected (red); T cells: Mtb-specific (dark pink), non Mtb-specific (light pink); caseation (tan); extracellular bacteria (dark yellow).

2.3.11 Modeling recruitment of Mtb-specific versus non-specific T cells

All T cells in the model (three functional classes: IFN-γ -producing T cells (Tgammas), cytotoxic T cells (Tcyts), and regulatory T cells (Tregs)) are recruited into *GranSim* as either specific or non-specific T cells.[80,87] At each time step, at each vascular source from where cells are recruited, T cell classes are recruited as determined by the chemokine concentrations at each vascular source. Three parameters (Tgam.probCognate, TCyt.probCognate and Treg.probCognate) determine the ratio of Mtb-specific to non-specific T cells. Patankar et al. showed in mice granulomas that 5%–20% of T cells are Mtb-specific.[88] Here, we varied the frequency of each Mtb-specific T cell class from 1%-25% to capture a potential larger range occurring within primates. Both specific and non-specific T cells enter the grid in a T_h0 state that requires further stimulation to fully differentiate and perform effector function. Non-specific T cells remain in a T_h0 state in the granuloma throughout the simulation and do not have the ability to kill Mtb. Both Mtb-specific and non-specific T cells have the ability to move on the grid and die from old age, or *via* TNF induced apoptosis. Macrophages with sufficiently bound surface MHCII receptors (larger than the MHCII binding threshold) can activate Mtb-specific T cells in their neighborhood.

2.3.12 Computer simulations and visualization

The 14 equation ODE model describing intracellular antigen presentation dynamics is solved within each macrophage within *GranSim* along with the new equation terms linking

models for each model time step. If the number of surface bound MHC II receptors of a macrophage is at or above a threshold (parameter MHCII Binding Threshold), then any specific T_h0 cell in the 1-neighborhood of a macrophage's transitions from the T_h0 state to an active state. *GranSim* was implemented in C++ with Boost and FFTw libraries. Partial differential equations describing diffusion are solved using Alternating Direction Explicit method. MHCII dynamic ODEs are solved using Runga-Kutta 4 method. Simulations for parameter sweeps were run without graphical visualization. The graphics visualization version was then used to load saved simulation states and generate graphics images to visually track granuloma formation. Computational model simulations were performed on XSEDE's Comet cluster and NERSC's Cori and Edison systems. For details on the system we use see https://www.sdsc.edu/support/user_guides/comet.html.

2.4 Results

Our goal is to study three key hypotheses explaining the relatively low frequency of T cells producing cytokines in TB granulomas. We test each hypothesis individually using both temporal and spatial modeling and appropriate control studies. In some instances, we have data that have been derived herein to provide support for our predictions.

2.4.1 Hypothesis 1: T cells are down-regulated directly by the action of Mtb

To explore this hypothesis, we include the role of intracellular dynamics of MHCII-mediated antigen presentation within macrophages within our multi-scale model of granuloma formation, *GranSim*. We then test four different types of down-regulation by Mtb on antigen

presentation individually and combine to observe the effects on downstream T cell activation. As a control, we implement no downregulation (as in the *GranSim* model without the submodel in place).

We performed 500 simulations using a wide range of biologically relevant parameters values (generated by LHS, see Methods). Figure 2.3 shows outputs for both the intracellular scale model and cellular/tissue scale model for different variables of interest. Model dynamics agree with datasets derived from NHP studies on granulomas and other in vitro studies for intracellular dynamics.[89–91] We also compared our model predictions to GranSim without MHCII presentation dynamics to confirm our model behaved accurately (positive control) (with values in Table A.5 set to 1, Figure A.3). Lastly, we calculated the numbers of Mtb-responsive T cells for 500 granulomas and compared it with data derived from 50 granulomas from an NHP study (Figure 2.4).[39]

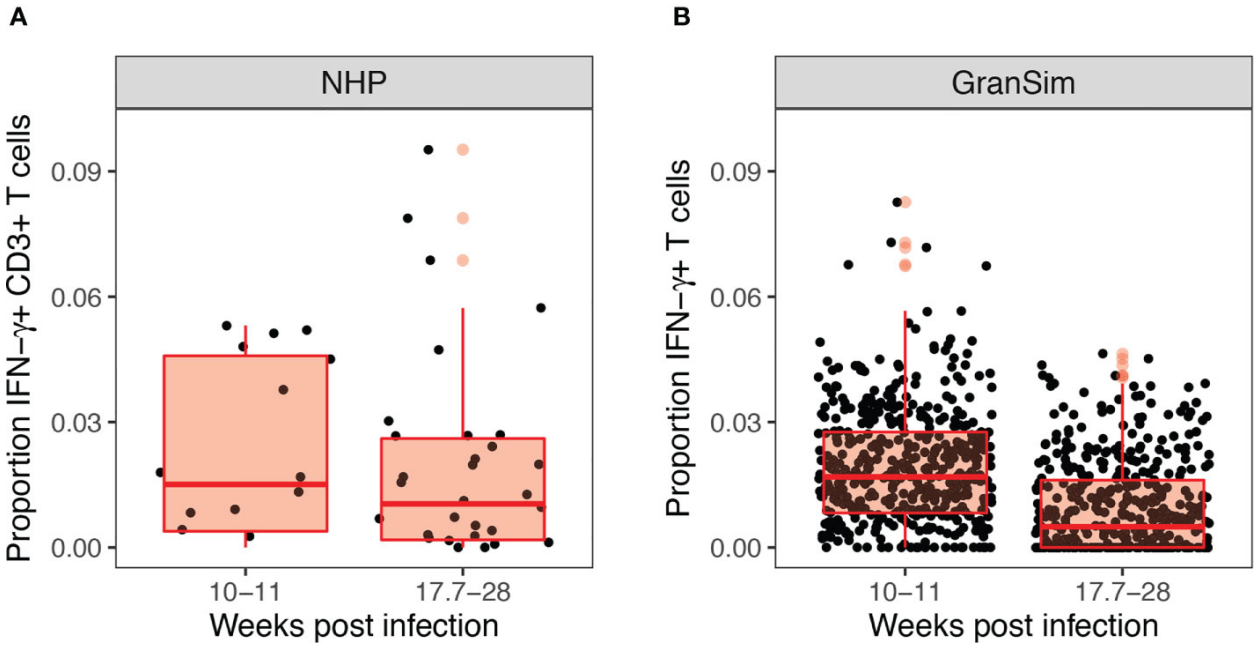

Figure 2.4 Experimental and computational models both reveal low levels of Mtb-responsive cells producing interferon-gamma (IFN-γ) within the granuloma. Panel (A) shows the experimental proportion of T cells exhibiting an IFN-γ response in non-human primates from Gideon et al.,[39] from 50 granulomas derived from 12 non-human primates (NHPs). Panel (B) shows our simulated prediction of the proportion of IFN-γ producing T cells in 500 simulated granulomas over the 28 weeks course of infection grouped to match the NHP dataset.

MHC self-derived peptides increase T cell activation when MHCII peptide loading is inhibited

We performed a sensitivity analysis on the intracellular-scale model (without Mtb assisted downregulation of MHCII presentation) and identified several parameters correlated with MHCII-mediated antigen presentation (Figure A.4). We found that increasing Mtb antigen processing rates by macrophages leads to increased levels of MHCII-Mtb presentation, but that these effects waned by 75 days post infection (Figure A.4). This matches general trends predicted in.[64] We also found that increases in Mtb antigen degradation or increases in MHCII-Mtb antigen dissociation leads to decreased MHCII-Mtb presentation on the surface of macrophages, also waning by 75 days post infection (Figure A.4). At the 75 days post in infection, Mtb levels within granulomas have leveled off, likely leading to diminishing returns for macrophage MHCII-Mtb presentation (Figure 2.3A). However, we did not identify correlations of antigen presentation parameters with a critical downstream effect, namely, levels of activated T cells within simulated granulomas. Previously a number of labs had identified different pathways in the antigen processing and presentation of MHCII-peptides by macrophages that were inhibited by Mtb.[48,49] Chang used the single cell model to study how Mtb affects certain processes of MHCII presentation and specifically studied Mtb interference of MHCII transcription, MHCII maturation (CIITA translation rate), antigen processing, and MHCII peptide loading.[64] We explored these same four processes within the linked multi-scale

model, where we tested a range for the maximum rate of Mtb down-regulation of each of the processes, at saturating bacterial levels, ranging from 25% to 100%.

Of the four processes that we studied for inhibition by Mtb, we found that reducing MHCII peptide loading was the only process that had a significant effect on T-cell activation levels (Figure A.5). When Mtb acts to down regulate MHC II peptide loading, as the degradation rate of peptide-MHC complexes increases, all T cell classes showed increased levels of activation, but only after 100 days post infection (not shown, $p < 0.0005$). At this point in the infection as Mtb levels plateau (see Figure 2.3E, day 75), degrading peptide-MHC complexes may help remove MHCII complexes loaded with host "self"-peptide. Continually degrading these complexes may help cycle through peptides quicker, making it more likely that MHCII are loaded with Mtb-peptides. This cycling through of peptides may be the only way to maintain a certain threshold of MHCII complexes presenting Mtb when peptide loading is greatly inhibited and Mtb levels have fallen. However, this contribution is small when observed as a proportion of activated T cells (Figure A.5). As a control, we also performed a sensitivity analysis on Mtb-induced downregulation of MHCII presentation for each of the processes, varying down regulation from 1% to 100% (Figure A.6). Any significant correlations with T-cell activation were small and transient, suggesting that these processes contribute very little.

Although previous wetlab and modeling studies showed that mycobacteria significantly inhibit antigen presentation processes, the focus of these *in vitro* studies was on less than 100 h.[48,49,64] However, our results suggest that bacterial mechanisms alone do not account for the observed low T-cell responsiveness levels of cytokine production observed in NHPs at a granuloma scale.[39]

2.4.2 Hypothesis 2: Spatial organization of granulomas affects the ability of T cells to reach macrophages and thus be activated via antigen presentation

We test a second hypothesis, namely that the spatial arrangement of cells within granulomas may create insufficient numbers of interaction opportunities between macrophages and T cells. This would imply that even if Mtb downregulation of processes is important, the chances for impact are few. In other words, granuloma spatial characteristics may contribute to low T cell responsiveness.[44]

To explore this idea, we used a two-pronged approach analyzing both experimental and simulated granulomas to better understand the spatial arrangements of immune cells. First, we randomly selected four experimental immunohistochemistry (IHC) images derived from four distinct NHP granulomas to directly identify and quantify the spatial organization between T cells and macrophages. We applied a novel approach using Geographical Information Systems Technology (GIS) similar to what we have done previously to analyze cell composition of granulomas [see Methods and Pienaar et al.[92]]. Here we not only identify the T cells and macrophage populations, but we additionally quantified the interaction overlap between T cells and macrophages, defined where these two cell boundaries intersect on the IHC image. We found that T cell-macrophage interactions occurred for, on average, only about 9.75% of the T cells identified (median: 8.6%, StDev: 4.5%), for at least the four granuloma that we examined (See Table 2.2).

Granuloma	CD3	CD11c	CD3/CD11c	Ratio contacts to T cells
9714_30	4,969	9,876	491	0.098
17613_37	15,448	3,943	1,127	0.073
17613_51	8,612	4,496	1,397	0.162
20612_29	3,284	5,535	197	0.060

Table 2.2 Geographical Information Systems Technology (GIS) analysis identifies numbers of immune cells. Includes cell types and numbers of contacts from four immunohistochemistry (IHC) granulomas.

Next, we used simulated granulomas and performed the same analysis to predict spatial locations of interactions between T cells and activated macrophages. For this, we simulated GranSim and controlled for bacterial inhibition as a factor by removing all possible Mtb-mediated down-regulation of MHCII presentation processes. We performed 500 simulations of granulomas and ran them out for 200 days post virtual infection. For each day during the virtual infection, we calculated median counts of T cells, new T-cell activation events, and numbers of distinct T cells that interacted with at least one macrophage in each of the 1-day intervals. Similar to the NHP granuloma T cell- macrophage interactions (Figure 2.1), we observed a similar order of magnitude difference between numbers of T cells and T cell-macrophage interactions, as well as new T cell activation events (Figure 2.5). This last feature is something we can uniquely track in *GranSim*. Distinct T cells that interacted with at least one macrophage occurred for only about 10% (with a range of 0%–22%, StDev: 4.4%) of all T cells identified for a given time point at 11-weeks post virtual infection and slowly declined to about 5% at 25-weeks post virtual infection (Figure A.7). As a control, we compared model predictions of T cells activation to *GranSim* with varying 'flexibility of T-cell density' as follows. We allowed the maximum number of T cells that can fit within one grid compartment to vary from 2 to 5, 2 being the default negative control (Figure A.8). Increasing the maximal allowable T cell density within a grid space did increase the proportion of activated T cells, but resulted in values that did not capture most of the data observed in the NHP study.[39] We also performed a sensitivity analysis on T cell density over the same range (Figure A.9). Increasing the density of T cells is correlated with increasing T cell activation. However, increasing T cell density did not result in

higher Mtb clearance. As T cell density increased, we saw a stabilizing effect on Mtb CFU after day 50 (Figure A.10). This is likely due to increased crowding on the grid, where T cells can slip by, but larger macrophages become stuck, making it more difficult to find and kill Mtb.

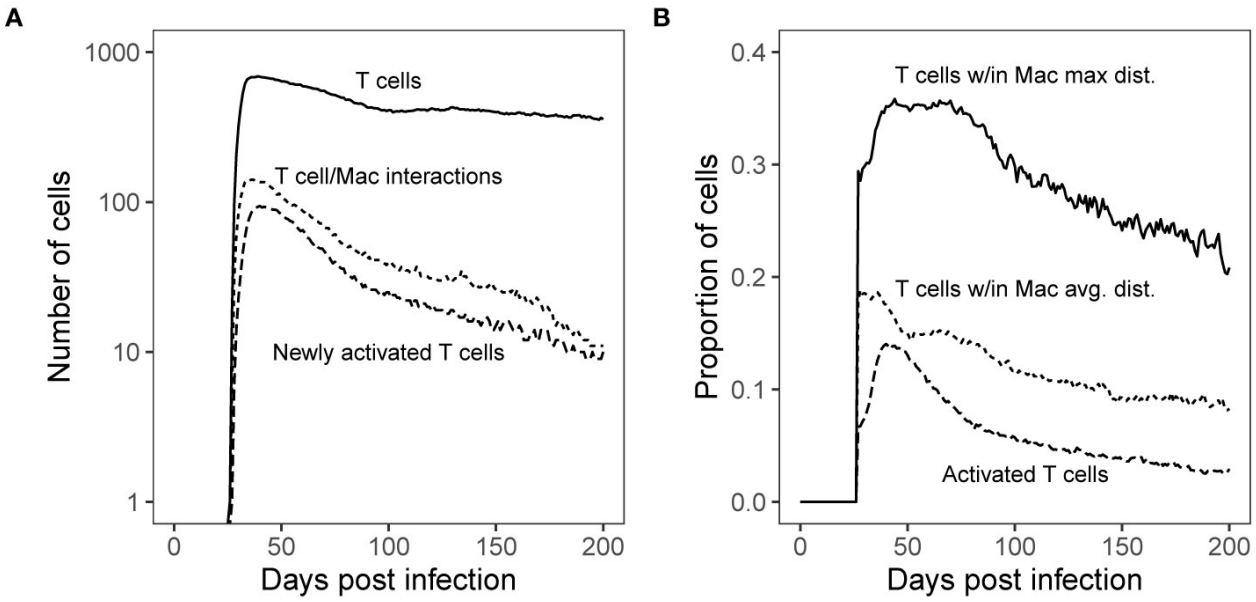

Figure 2.5 Simulated T cell/Macrophage Interactions in time Shows Spatial Analysis of Simulated Granulomas. Panel (A): Median count T cells over the course of Mtb infection compared with the median count of distinct T cells that interact with macrophages and new T cell activations per time step. Panel (B): Proportion of T cells found within the maximum boundary and average distance of antigen-presenting macrophages (measured from the granuloma center of mass) as compared with the proportion of total activated T cells (all activated T cells are Mtb-specific). These distances are shown spatially in Figure 2.6.

The majority of T cells are not being stimulated

Secondly, we test that the majority of T cells are not being stimulated. Both wetlab and computational studies to support the idea that within granulomas, T cells are often not sufficiently close to macrophages to become activated.[50,51] We can use our simulations to determine how far each cell type is from the center of a granuloma. For each day of the virtual infection, we used *GranSim* to calculate median proportions of T cells found within both the

average and maximum distance of macrophages from the granuloma center of mass [as in Renardy et al.[85]]. These numbers are compared with the proportion of activated T cells (Figure 2.5B). Within our simulations, roughly a third of T cells travel deep enough within the granuloma to have the possibility of reaching activated macrophages. Of these, about half made it past the average distance of T cell stimulating macrophages from the granuloma center of mass, increasing their chance of encountering a T cell stimulating macrophage (Figure 2.5B).

To get a more detailed look at the simulated granulomas and spatial distributions of cells, we extracted the coordinates of all macrophage and T cell agents in our simulated granuloma (see Methods). In Figure 2.6, the distribution of macrophages (Figure 2.6A) and T cells (Figure 2.6B) are drawn in relation to the granuloma center of mass. The area shaded gray is the distribution of activated T cells. At the height of T cell activation (occurring about 7-weeks post infection), almost all activated T cells (Figure 2.6A, shaded gray) can be found within the spatial region of macrophages that are able to stimulate T cells. That is, very few activated T cells are found near activated macrophages. This distribution of activated T cells can be divided further, with two thirds residing within the average distance of activated macrophages from the granuloma center of mass. About 90% of the activated T cells are found within the average Mtb-specific T cell distance from the granuloma center of mass (Figure 2.6B). In general, the distribution of activated T cells closely follows the distribution of activating macrophages. Low T cell stimulation, taken together with limited T cell access to macrophages and an observed increase in T cell activation by increasing T cell density suggest that spatial mechanisms play a major role in the observed low T-cell responsiveness levels of cytokine production observed in NHPs at a granuloma scale.

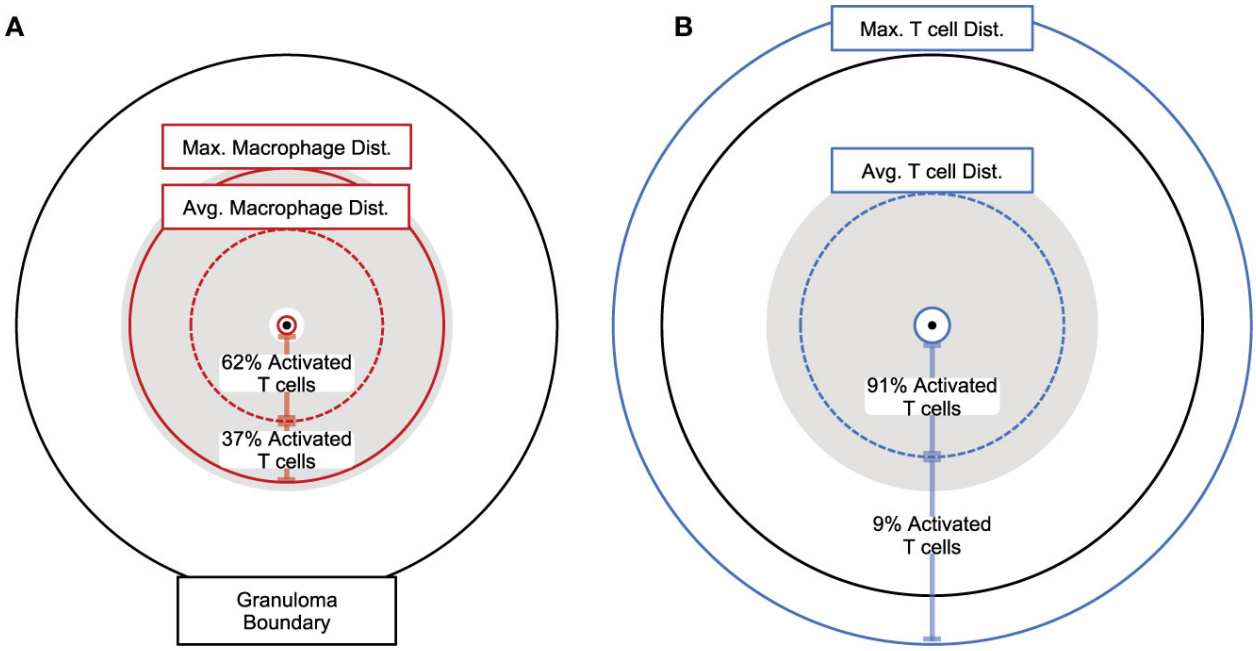

Figure 2.6 Spatial analysis of granulomas showing distances between macrophages and T cells. Proportional distances of average maximum, and minimum (A) activated macrophages presenting antigen or (B) Mtb-specific T cells from the granuloma center of mass (7-weeks post infection). Maximum and minimum distances shown as solid circles, average distances show as dotted circles. Gray shaded areas encompass all activated Mtb-specific T cells present. Percent of all activated T cells is shown between minimum and average, as well as average and maximum distances for both (A) activated macrophages presenting antigen and (B) Mtb-specific T cells.

2.4.3 Hypothesis 3: The majority of T cells within granulomas are non-Mtb specific

To test the hypothesis that the majority of T cells within granulomas are non-Mtb Specific, we focus on the spatial distribution of Mtb-specific T cells (Figure 2.6). Given previous studies showing that T cells are recruited to macrophages indiscriminately,[54–56] we expanded upon our study of the composition of Mtb-specific T cells within simulated granulomas. We use our multi-scale model to determine the proportion of Mtb-specific T cells within granulomas over the course of infection as compared to non-specific T cells present by comparing which frequencies match the dataset derived from NHP. To do this, we varied the frequency of Mtb-

specific T cell classes from 1%-25% to capture a potential larger range occurring within primates.[88] For each day of virtual infection, we used GranSim to calculate median, numbers, and proportions of Mtb-specific T cells versus non-specific T cells found within simulated granulomas (Figure 2.7). Within our simulations, non-Mtb-specific T cells greatly outnumber Mtb-specific T cells (Figure 2.7A). The proportion of Mtb-specific T cells peaks at around day 40 at about 20% (Figure 2.7A) and declines to about 10% by 200 days post virtual infection.

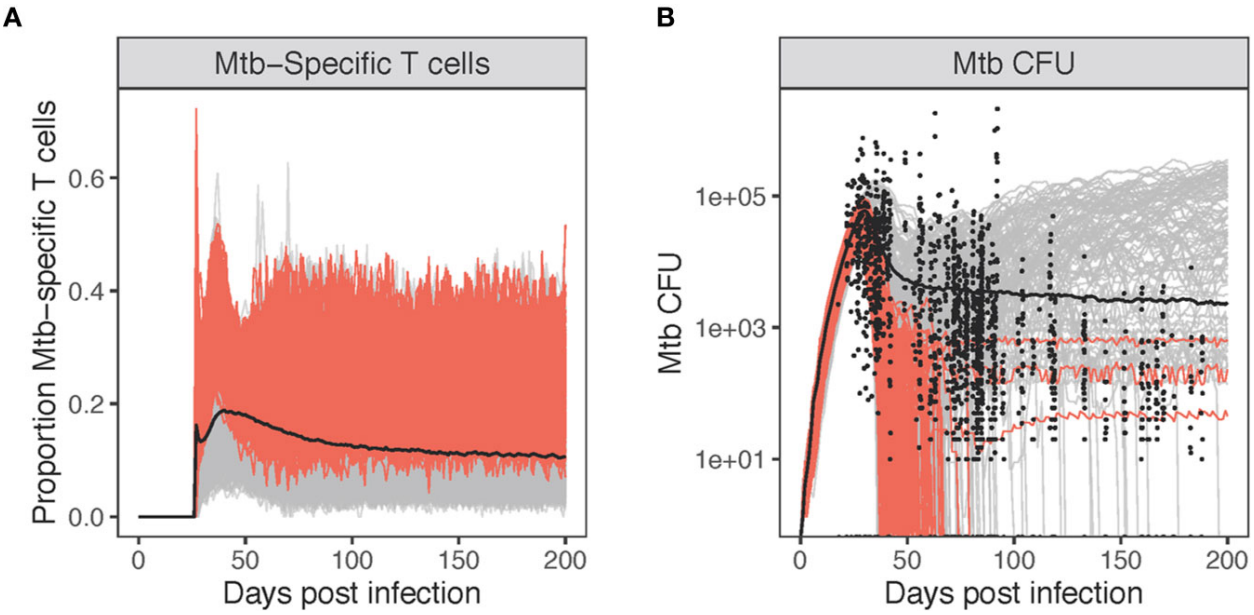

Figure 2.7 Varying levels of Mtb-specificity in granuloma T cells to match non-human primate (NHP) studies. Panel (A): The proportion of Mtb-specific T cells over the course of virtual infection. Panel (B) shows the CFU [shown together with 1,994 NHP granulomas (black circles)]. Highlighted in red are simulations where the probability of Mtb-specific T-cell classes recruited are allowed to exceed 10% (up to 25%). All other simulations are gray (1%–10%).

Given the low number of Mtb-specific T cells within granulomas, we looked at how these simulations matched with NHP Mtb CFU data (Figure 2.7B). Highlighted in red are simulations where the probability of Mtb-specific T-cell classes recruited is allowed to exceed 10% (all other simulations are gray). Allowing recruitment of a great portion of Mtb-specific T cells results in

the majority of granulomas sterilizing by day 100 (226/229 = 0.987). The red simulations (Figure 2.7B) also overlap with relatively few of the NHP CFU data points. This suggests that low levels of Mtb-specific T cells are what likely is present within granuloma, leading to the observed NHP Mtb CFU datasets (Figure 2.7B). Any further increases in levels of Mtb-specific T cells within granulomas leads to sterilization of 99% of all granulomas, which is not a typical outcome; more typically around 50%.[60] As a control, we also performed a sensitivity analysis on Mtb-specific T cell proportions for each class of T cells represented in *GranSim*, varying specificity from 1% to 25% (Figure A.9). The probability of IFN-γ producing T cells being Mtb-specific had the greatest effect on T cell activation and Mtb clearance. When T cells were first recruited, they had a strong, positive effect on the number and proportion of activated T cells. By day 50, this shifted to a negative correlation, due to the small numbers of activated T cells seen at this stage of infection. Increasing Mtb-specific IFN-γ producing T cells also increased the clearance of Mtb, matching the sterilization observed in Figure 2.7B.

2.4.4 Combining the 3 hypotheses

We performed a sensitivity analysis, varying parameters for all three hypotheses at the same time to search for combined effects, using the previously specified parameter ranges (Figure A.11). In general, correlation patterns for these input parameters in combination were similar to those seen when simulated separately (Figures A.6 and A.9). One difference we observed is that parameters controlling the proportion of Mtb-specific T cells continued to have a positive effect on the number and proportion of activated T cell past day 50. Together, these results suggest that the proportion of Mtb-specific T cells influences the observed low T-cell

responsiveness levels of cytokine production observed in NHPs at a granuloma scale and the other hypotheses may help extend these effects over time.

2.5 Discussion

Over 90% of Mtb infections in humans are well controlled and asymptomatic, known as Latent TB infection (LTBI), indicating that the immune response to Mtb, which is characterized by granuloma formation, is relatively successful at containing infection. Cynomolgus macaques also present with active or latent TB. As the majority of granulomas in both latent and even active NHPs eventually sterilize,[60] this means that granulomas (on an individual basis) have the ability to clear infection. It is surprising that low T-cell activation levels through measuring IFN-γ and other cytokine responses have been observed within non-human primate granulomas.[39] There are a number of hypotheses as to why low numbers of responsive T cells might be present in TB granulomas, including T cell exhaustion, Mtb-mediated downregulation of antigen presentation by macrophages, the spatial organization of cells within granulomas and the presence of non-Mtb specific T cells. We previously explored exhaustion; however, our results indicate that it cannot explain the observed low levels.[44] Herein we explored the other three hypotheses as to why the numbers of Mtb-specific T cells are low. First, we focused on macrophages and their role in this outcome: We asked whether Mtb was down-regulating MHCII presentation of Mtb antigens and whether that reduced T cell activation. At the scale of the entire granuloma, we did not see significant differences in MHCII presentation of Mtb antigens by macrophages, with or without Mtb down-regulating of MHCII presentation (Figure A.5). Previous work suggesting that inhibition by Mtb was a key player in reducing MHCII presentation were based on studies that spanned time scales of 1-100 h, while granulomas

survive for months to years.[48,49,64] The dynamics observed on the scale of a few hours may be washed out given the extended lifespan of macrophage and Mtb- dynamics with granulomas. Further, even if large scale reductions in antigen presentation are occurring, our further studies indicate that there are insufficient interactions occurring between T cells and macrophages for that to manifest as a key factor.

Although we did not see a significant decline in MHCII presentation of Mtb antigens on the surface of macrophages at any given time, we did observe that only 5-10% of macrophages in our simulations were capable of activating T cells. In *GranSim*, only macrophages that have contact with both IFN-γ (*via* Stat 1) and Mtb or TNF (*via* NFkB) can present Mtb antigens via MHCII. Macrophages must continually receive those stimuli until they surpass an MHCII surface level threshold required to activate T cells.[71] These activated macrophages were spatially located mostly within the center of our simulated granulomas, where they would have access to Mtb and Mtb antigens. Since many macrophages did not receive stimuli necessary for MHCII Mtb-antigen presentation, it is not surprising that the inhibition of antigen presentation by Mtb was minimal at the tissue scale. It is also possible that since our model does not account for the effect of chemokines attracting T cells to antigen presenting cells, and Mtb antigens do not include secreted antigens, antigen concentration and T cell numbers responsive to infection may be underestimated. However, in previous work, we have examined this idea of APCs secreting chemokines to attract T cells, and have shown that it leads to tremendous crowding around APCs, limiting stimulation.[93]

Since direct inhibition of antigen-presentation by Mtb was insufficient to reduce T cell responsiveness, we explore the second hypothesis of how granuloma spatial structure may affect T cell activation. Our analysis of four HIC images from NHP granulomas suggest that there are

limited interactions between T cells and CD11c+ macrophages within granulomas, and further analysis with additional granulomas is warranted. Previous studies have also shown that within granulomas, T cells are highly motile but restricted by space, with movement occurring mostly at the borders of the granuloma.[50,51,55] In addition, the typical structure for a granuloma is a lymphocytic cuff surrounding macrophages and other cells including bacteria and more centralized necrosis.[94,95] All activated macrophages within our simulations are located near the center of granulomas and most T cells are unable to reach them. In fact, only about 5% of all T cells in *GranSim* interact with macrophages at any given time. Activated T-cell life spans are short (on average 3 days), so large numbers are unlikely sustainable. It should be noted that T cells can have functions other than cytokine production, and our study used only data on T cell production of IFN-γ. Assessing other T cell effector functions, such as other cytokines and cytotoxic potential, could result in an increase in numbers of T cells that are responsive to the infection. Within this study, low T cell stimulation, taken together with limited T cell access to macrophages and an observed increase in T cell activation by increasing T cell density suggest that the structural organization of the granuloma seems to impact the T cells in a significant manner. However, that does not rule out that other factors are playing a simultaneous role. What we have shown is that the spatial effects are necessary condition for this reduction in T cell responsiveness; however, it could certainly hold true that these other factors are also playing a role in augmenting those dynamics, albeit less significant. Our combined analysis of all three hypotheses simultaneously confirmed this.

Continuing with the idea of limited T cell-macrophage interaction dynamics, we explore the issue of recruitment of Mtb-specific T cells into granulomas. Previous studies have shown that T cells localize to and are rapidly recruited into mycobacterial granulomas in the absence of

antigen recognition.[54–56] One study found around 5-20% of CD4+ T cells recognize Mtb-infected macrophages by 19- to 22-weeks post infection.[88] Comparatively in *GranSim*, 10%–20% T cells are Mtb-specific T cells 7- to 28-weeks post infection, with only Mtb-specific T cells given activation capabilities. When these percentages increase in *GranSim*, T cell activation goes up and the vast majority of granulomas are sterilized. Given that such a small percentage of T cells can recognize Mtb antigen, along with the limited migration of T cells, these two factors combine to make T-cell activation a rarer occurrence than one would expect. As most granulomas can sterilize or greatly reduce bacterial numbers, this level of T cell activation may generally be effective in conjunction with other help from the immune response. However, if specificity and location could be affected in a direct way, the numbers of activated T cells would increase and infection would likely be cleared within all granulomas. An appropriate vaccine could lead to this outcome.

Chapter 3 Assessing Strategies for Improving Detection of Colonization with Vancomycin Resistant *Enterococcus* in Hospitals

3.1 Abstract

Vancomycin resistant Enterococci (VRE) are a leading cause of nosocomial infections, with the major site of transmission being acute and long-term health care facilities. To combat VRE's continued persistence, both the CDC and Society for Healthcare Epidemiology of America created recommendations for prevention and control. Due to the asymptomatic nature of VRE, these recommendations rely on the use of surveillance with the goal of identifying VRE colonized patients prior to spread. It is unknown how well various testing strategies identify VRE colonization status under different levels of transmission. We use an agent-based model of VRE transmission to simulate VRE spread using 6 years of hospital patient location and movement data to better understand the performance of different active surveillance protocols as well as the effect of improving compliance with these screening protocols. In these simulations, we varied parameters for VRE spread (the transmission rate and the proportion of patients colonized on admission) as well as compliance rates for different testing strategies and assessed the proportion of VRE positive patient days correctly identified. Cost effectiveness analysis was used to identify screening elements that increase the efficiency of identifying VRE cases. We identify that the current surveillance protocol that test patients on admission and weekly thereafter is consistent with simulations that miss 20-60% (depending on the compliance rate) of the patient-days of VRE colonization across the 9 hospital units that undergo active surveillance.

Maximizing compliance with the existing protocol improved the efficiency in detecting VRE more than increasing the frequency of testing. Improving compliance with existing surveillance protocols may be a simple and effective way to improve detection of VRE colonization.

3.2 Introduction

Vancomycin resistant *Enterococcus* (VRE), caused by *Enterococcus faecium* and *Enterococcus faecalis*, is a leading cause of nosocomial infections, with higher morbidity and mortality than vancomycin-susceptible Enterococci.[96,97] It is a hospital-acquired pathogen, and the major reservoirs for transmission are asymptomatically colonized patients in acute and long-term health care facilities.[98,99] The risk of disease from infection has been linked with medical co-morbidities such as organ transplantation, intensive chemotherapy, and severity of illness.[100–102] VRE has also evolved resistance to other antibiotics, including daptomycin and linezolid, leading to worse patient outcomes.[100,101,103,104] Since first identified in 1988, VRE spread rapidly, increasing in hospitals in the US between 1989-1993.[105] VRE has become endemic in many hospitals in the US and in other countries.[106,107] Currently VRE comprises >25% of all enterococcal bloodstream infections in the US.[108,109] The increase in VRE in hospitals is thought to be attributed to its continued persistence within hosts, transmission between hosts, and its capacity to evolve resistance.[110–112] In general, VRE responds poorly to aggressive infection prevention measures.[113,114] Without effective intervention, hospital endemic VRE continues to persist.[115,116]

Following the emergence of VRE, regulatory bodies in the US created recommendations to improve VRE prevention and control. In November 1994, the CDC's Healthcare Infection Control Practices Advisory Committee (HICPAC) ratified recommendations for prevention and

control of vancomycin resistance.[106] These recommendations include a variety of interventions to reduce doctor-to-patient and patient-to-patient transmission, as well as active surveillance through rectal cultures.[117–119] The protocol for conducting active surveillance was less clear, with a note that, "The frequency and intensity of surveillance should be based on the size of the population at risk and the specific hospital unit(s) involved."[106] In 2003, the Society for Healthcare Epidemiology of America (SHEA) expanded upon these guidelines, emphasizing that the asymptomatic nature of VRE lends itself to silent transmission and much of the VRE reservoir would be unrecognized without active surveillance cultures.[120–124] Due to this, SHEA recommended the use of active surveillance with the goal to specifically identify the reservoir of VRE for spread.[116,125–134] SHEA gave a number of different suggestions on how to perform active surveillance, based on the literature available at the time. These recommendations included who should be tested (patients at high risk for transmitting or acquiring VRE) and when/with what frequency. As for when patients should be screened, SHEA suggested screening on admission for patients with high risk for carriage of VRE and weekly for those at risk of acquisition.[116,126,130,133,135] SHEA encouraged that active surveillance should be used throughout the healthcare system but tailored based on prevalence of VRE and risk factors for colonization.

Protocols for VRE active surveillance in hospitals vary greatly, and in some hospitals is not practiced at all. The percentage of hospitals that participate in active screening for VRE in the US is unknown, but VRE screening is known to be variable in other countries, such as Canada (76%) and Switzerland (44%).[136,137] Many types of screening protocols have been implemented, drawing upon recommendations from the CDC and SHEA. One of the most common screening protocols focuses on specific units deemed high risk for acquisition (such as ICU, transplant, or hematology units). Other variations include targeting patients prescribed

vancomycin, week or longer stay, previous hospital stay, and other patient characteristics that put patient at high risk (such as underlying disease or injection drug use).[138,139] Regarding timing/frequency of screening, protocols often include a screen +/- 24-48 hours upon admission to the hospital or upon transfer to a high-risk unit. Periodic testing is also often carried out for patients deemed at high risk of acquisition usually 7-9 days after a previous test or on specific testing day of the week, but other variations include every 3-4 days to biweekly.[113,140,141] Some also screen at discharge from a high-risk unit.

How well hospitals adhere to their screening protocols is not well characterized in the literature. In practice, compliance is not often included in published studies. When it is reported, observed rates have varied wildly between 30-90%, but often 60-70% is seen.[116,142–147] With variable rates of compliance and often little information on its characteristics (low compliance for certain timing of tests, certain units, etc.) it becomes difficult to judge the efficacy of any complimentary intervention meant to reduce VRE spread (contact precautions, hand hygiene, reducing antibiotic use, etc.). Before testing the efficacy of downstream interventions on transmission reduction, there is need to understand how well we are detecting cases, as undetected cases may mitigate efforts to reduce spread.

Finding the optimal testing strategy to identify VRE cases has been elusive. Compared to the variety of protocols used, few studies have looked at optimal testing strategies. Most of the studies that look at active surveillance couple screening with downstream interventions such as contact precautions, focusing on the reduction of VRE as an outcome.[113,141,145–147] Some studies have looked at how many more cases were captured with different testing strategies, but these often overlook compliance rate, or assume it is very high (90-100%%).[148,149]

To better understand the performance of different active surveillance protocols and the effect of compliance, we use a mathematical model of VRE transmission to simulate spread. The model utilizes existing hospital patient location and movement data. Model parameters are varied for VRE spread (transmission rate and proportion of patients infected on admission) as well as compliance rate for different testing strategies and assessed for the proportion of VRE positive patient days correctly identified. We also consider differences in identifying patients colonized with VRE prior to admission and those whom acquired it during their stay. We also provide estimates from a hospital, and calibrate transmission and compliance based on data from our active surveillance program.

3.3 Methods

3.3.1 Study location

University of Michigan Healthcare system is an approximately 1000 inpatient bed hospital in Ann Arbor, MI. The system serves as a tertiary referral hospital for southeastern Michigan. This is a retrospective open cohort investigation of the effect of active screening compliance on VRE detection through June 1, 2013 through December 2019.

3.3.2 Screening

Active surveillance for VRE is conducted in 9 units where patients are deemed at higher risk for infection, including intensive care units, the hematology and oncology ward, and the bone marrow transplant ward. Patients are screened for VRE on admission to one of these units, and weekly afterwards if previous tests were negative. Patients are eligible for an admission test in a screening unit (+/- 24 hours within admission to unit) if they have not previously been tested

for VRE during their stay or their last test was negative. If an admission swab is negative (or missed), patients are eligible for periodic screening each week if their previous test was negative. If all tests during stay in screening unit are negative, patients are eligible for a discharge test (+/- 24 hours within discharge from unit). Detection for VRE was conducted with rectal swab samples applied to Bio-Rad VRESelect chromogenic medium.

3.3.3 Patient location and VRE testing data

Location data for patients was pulled from the electronic health record (EHR) system to include all hospital admissions from January 1, 2012 through December 31, 2019. This totaled 317,792 patients, 602,622 admissions, and 819,699 additional transfers between hospital units.

Data for rectal swabs testing data positive or negative for Vancomycin resistant *Enterococcus faecium* and *faecalis* were pulled from the EHR to include all tests conducted from June 1, 2013 through December 31, 2019. 88,320 total tests occurred (7,738 positive), corresponding to 43,541 patients. Blood clinical isolate testing data positive for Vancomycin resistant *Enterococcus faecium* and *faecalis* were pulled from the EHR to include all tests conducted from June 1, 2013 through December 31, 2019. 517 total tests occurred, corresponding to 318 patients. Non-blood clinical isolate testing data for positive for Vancomycin resistant *Enterococcus faecium* and *faecalis* were pulled from the EHR to include all tests conducted from June 1, 2013 through December 31, 2019. 1,830 total tests occurred, corresponding to 1,317 patients.

3.3.4 Patient VRE status

A patient was considered to have VRE if they had at least 1 positive test (surveillance or clinical) during their hospital admission, which included 5,012 patients (with 7,080 admissions). A total of 10,085 tests were positive (7,738 surveillance, 318 blood, and 2,029 non-blood clinical). A patient was considered to have hospital acquired VRE if they had an initial negative swab sample within the first 72 hours of admission to the hospital, followed by a positive sample after 72 hours. There were 811 patients (with 830 admissions) who fit this definition.

3.3.5 Hospital variable calculations

VRE prevalence within the hospital was calculated as the total number of patient-days a patient is known to be VRE positive per 1000 patient-days. The proportion of patients positive for VRE on initial hospital admission was calculated as the proportion of patients who tested positive for VRE that were administered a test within the first 48 hours of admission to the hospital.

Active surveillance compliance rates were calculated based on whether a patient was entering a unit (admission screening), was eligible for additional screening during their stay (periodic screening), or was leaving a unit (discharge screening). As mentioned previously, patients lost eligibility for further screening if their previous test was positive.

Admission screening compliance was calculated as the number of eligible patients with a valid admission test (+/- 1 day from entering a screening unit) divided by the number of all patients that were eligible for screening. Periodic screening compliance was calculated as a ratio of patient-days in compliance with screening to the patient-days eligible for screening. If a patient had an admission test, they were considered compliant for 7 days after admission. If the patient did not have an admission test, they were considered compliant for 1 day after admission

as a grace period for screening. If the admission compliance time covered the rest of the patient's stay, or the patient tested positive, the patient was considered compliant during the entire unit stay. If the admission compliance time did not cover the rest of the patients stay and they tested negative, the patient was considered compliant if they had a periodic or discharge screening before the previous compliance time elapsed. A patient with a periodic or discharge screen was considered compliant for 7 days after administering the test. If there were gaps in time between a compliance window from one test and a subsequent test, the time between would be considered non-compliant. (For example, administering an admission screening, followed by a periodic test 9 days later, the patient would have 2 days where they were non-compliant.) Discharge screening compliance was calculated as the proportion of eligible patients with valid discharge test (+/- 1 day from leaving a screening unit) from all patients that were eligible for screening.

3.3.6 Modeling pathogen spread and testing

To better understand the performance of different active surveillance protocols and the effect of compliance, we use a simulation of VRE transmission to simulate spread based existing hospital patient location and trajectory data. Briefly, the main model operates on a between-hosts scale, tracking patient infection status as patients move through the hospital. In this agent-based model, pathogens are considered agents and the presence of patients within hospital units allow for patient to patient transmission of VRE. These dynamics are all captured between patients based on known patient location trajectories and stochastic interactions of patients within units. Inputs to this model are β, the effective transmission rate, γ, the recovery rate, p_0, the proportion of patients infected with VRE on admission to the hospital, and a file containing patient movements and times.

Transmission events occur based on a Poisson process, $\text{Pois}(\beta Iu)$, where β is the effective transmission rate, I is the number of infected patients within a unit, and u is the time the patient spends within that unit. βIu is the number of new infected patients that we would expect to occur over the time period, u, given the number of current infections and infection rate. Recovery events occur based on first calculating the number of patients that could recover for a given infection duration and the number of patients currently infected at that time step. To model recovery events, a Binomial process is used, $\text{B}(n, 1 - e^{-\gamma d})$, where n is the number of patients with current VRE infections, γ is the recovery rate, and d is amount of time since recovery was previously assessed (at least 6 hours). In this instance, $1 - e^{-\gamma d}$ is the per-capita probability of recovery for an infected patient. The number of patients that recover is then calculated from the Binomial distribution, with one draw per infected individual. The resulting number is used to choose the number of infected patients, at random without replacement, that will recover at that time point.

The model input for p_0 was calculated as the overall hospital average over the middle three years, 2015-2017, with a value of 0.06 that's within the range observed in US academic centers.[150] Recovery rate, γ, was calculated using the catalytic model, $\gamma(t) = -\ln\big(1 - F(t)\big)$, where $F(t)$ is the proportion of patients who converted from positive to negative since time t. $\gamma(t)$ was then averaged over all t, resulting in a value of 0.004 per day, within an order of magnitude of values found in the literature.[150,151] β was estimated to be 0.05 per day through varying β from 0.01-0.20 and calibrating based on the overall number of tests conducted per unit and the proportion of positive tests observed (Figures B.1-B.2).

Patient screening was modeled based on movement within screening units. Different periodic screening protocols were considered from once daily to every 14 days, as well as

including an initial admission screening to a unit. For each regimen, admission and periodic tests occurred based on unit specific compliance rates for each, either as the known average of that unit, or over a range of 10 to 100%. The result of the test would be decided based on the true infection status and the random assignment of false positives (range 0.05-5%), and false negatives (range 5-20%).[152–155]

3.3.7 Sensitivity and cost effectiveness analysis

To narrow parameters of interest, a partial rank correlation coefficient sensitivity analysis was conducted, using the R package sensitivity (v1.27.0). Parameters of interest were varied to characterize their uncertainty (Table B.1). The sensitivity analysis then characterizes how much relative influence a parameter has on the model output. Cost effectiveness analysis was conducted using the R package dampack (v1.0.1). This analysis helps to examine the cost and outcome of different interventions. In this analysis, cost was limited to the relative cost to the hospital for each screening conducted (1 screening equaling a cost of 1, 2 screening equaling a cost of 2, etc.). Interventions included different screening frequencies, different compliance rates, and combinations of both. All calculations, plots, and simulation of patient screening was conducted in R (v4.1.3). Simulation of pathogen transmission was implemented in Python (v3.8.8). Code is available in our GitHub repository.

3.4 Results

In general, the data from this hospital show that patients who test positive for VRE comprise about 1% of our patient population (Table 3.1). About 15% of VRE positive patients have an initial negative sample followed by a positive sample at least 72 hours later, meeting our

conservative definition of hospital acquired VRE. Overall, the majority of patients who test positive for VRE in our hospital visit at least one screening unit during their stay (~80%). Most of the VRE cases were detected though active surveillance swabs (~80%).

	All patients	All VRE positive	Hospital Acquired
Admissions			
All Hospital	460,703	6,751	1,005
Visit 1+ Screen Unit	63,036	5,303	996
Swab Tests			
Positive	7,060	7,060	1,281
Negative	71,392	2,586	1,990
Positive Clinical Test			
Blood	528	528	165
Other	2,303	2,303	273

Table 3.1 Totals of UM hospital admissions and tests for patients without positive VRE test result. Encompasses patients admitted between 2012-2019 at UM over the entire hospital.

During the study period, the overall VRE incidence trended downward (Figure 3.1A). This pattern matches a similar trend observed in the proportion of patients positive for VRE on admission (Figure 3.1B). However, overall screening compliance also decreased about 20% during this time (Figure 3.1C).

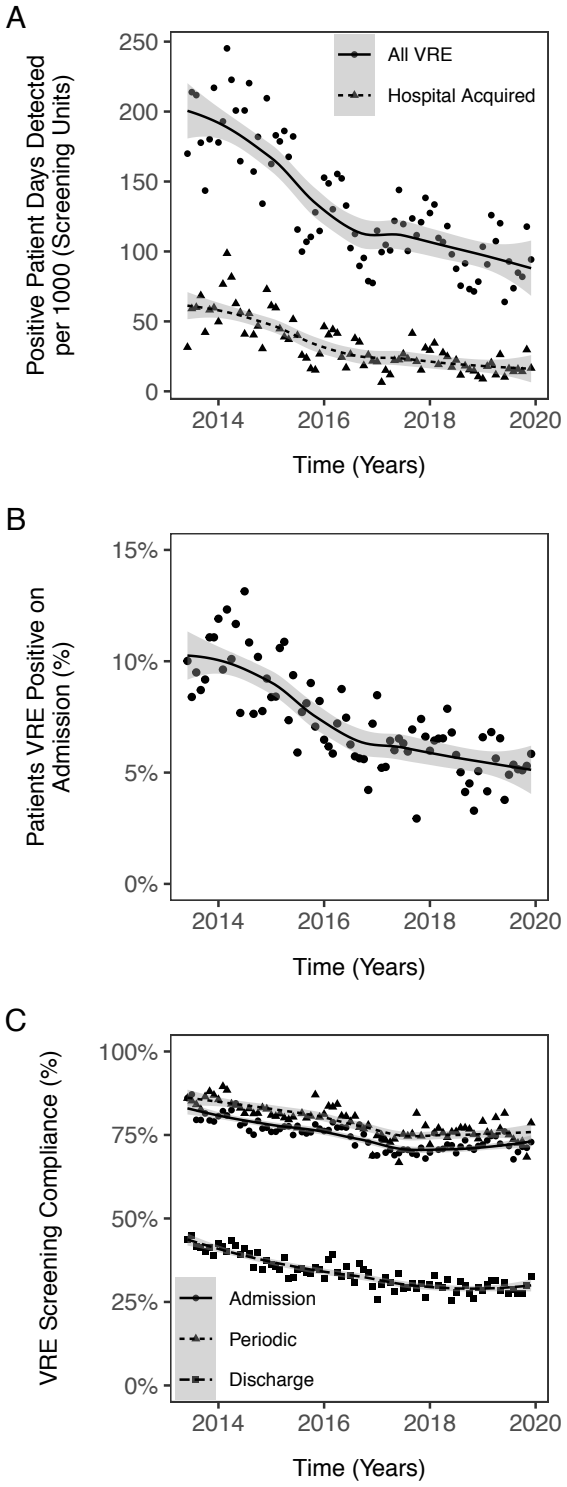

Figure 3.1 Trends in UM hospital VRE cases and screening compliance 2013-2019. Panel A: Levels of VRE observed historically in screening units over time by month. Panel B: The proportion of patients tested for VRE during the first 48 hours of hospital stay that are VRE

positive. Panel C: VRE screening compliance rates observed in the hospital over time by month. Loess smooth shown as black lines for all panels.

Screening compliance rates varied greatly between units and over time. While numbers differed between the different types of compliance, temporal trends were fairly consistent between them (Figures 3.2, B.3, B.4). Several units saw downward trends in compliance during all or the latter half of this observed time period, with several dipping below 50% in periodic screening compliance. Periodic, weekly follow-up compliance rates were generally higher than admission, while discharge compliance was almost always below 50%.

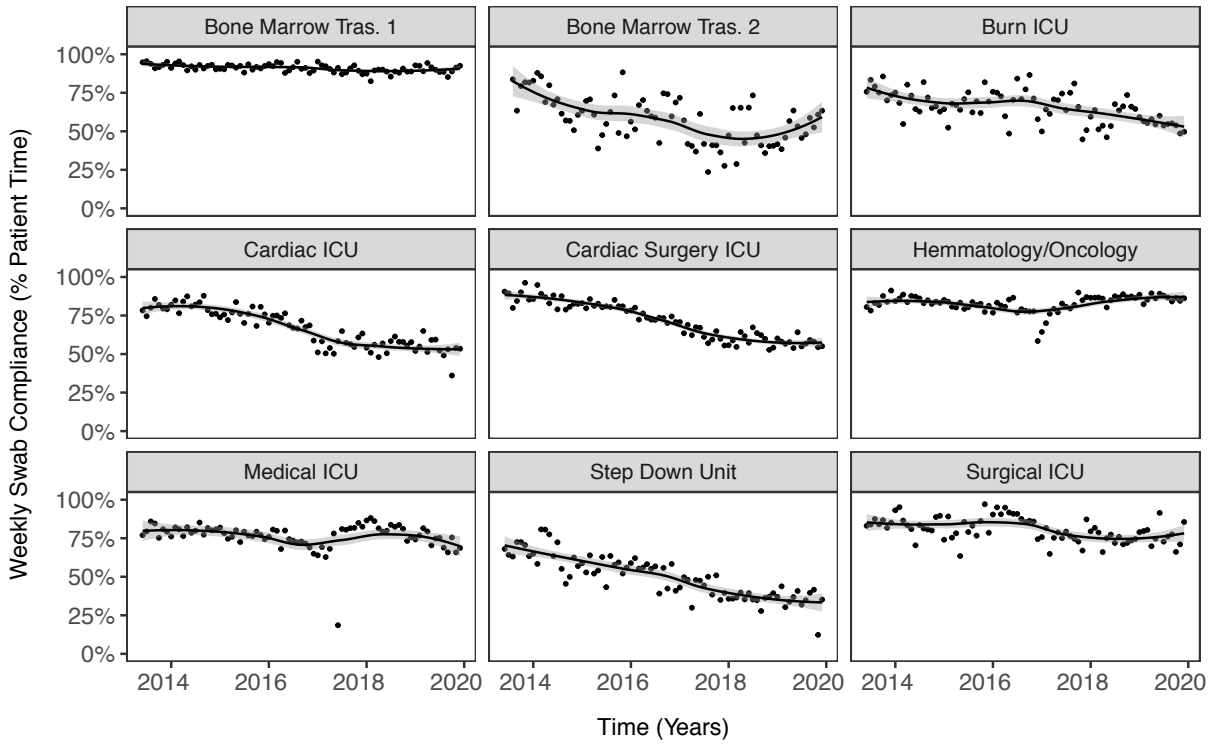

Figure 3.2 Comparison of patient time in compliance with weekly swab tests performed in screening units by month. Loess smooth shown as black lines.

3.4.1 Compliance rate and testing frequency have the greatest effect on detecting VRE

To estimate the VRE incidence captured through active surveillance we simulated pathogen transmission over our known patient's movement trajectories over ranges of transmission rates, screening compliance rates, sensitivity and specificity of the screens, and proportion of patients positive for VRE on admission. Screening protocols were then applied to this simulated transmission patterns that varied the frequency of periodic testing and the inclusion of a unit admission screening.

We performed a sensitivity analysis on our VRE simulations and identified compliance rate and testing frequency had the greatest impact on the percent of positive patient days detected (Figure 3.3). This relationship can be seen with the analysis expanded to individual screening units in Figure 3.4. Simulating the model and assuming perfect compliance, in most cases no more than 20% of positive patient days are missed even with periodic screening spaced out to once every two weeks (Figure 3.4). Simulating compliance dropping to 75% indicates that testing would be required every 2-5 days to capture the same amount VRE positive patient days, and we predict that testing every two days would be required once compliance hits 50%. The relationship between screening frequency and other parameters (transmission rate, screening compliance rates, and sensitivity and specificity of the screens) on the percent of positive patient days missed were generally less pronounced (see Figures B.5-B.13).

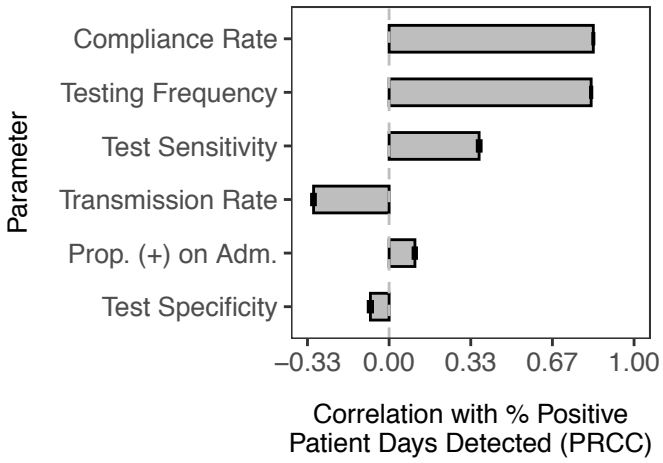

Figure 3.3 Results of sensitivity analysis. Partial rank correlation coefficients are shown for six model parameters inputs with their effects on the percent of VRE positive patient days detected as the output variable. 95% bootstrapped confidence intervals shown as error bars.

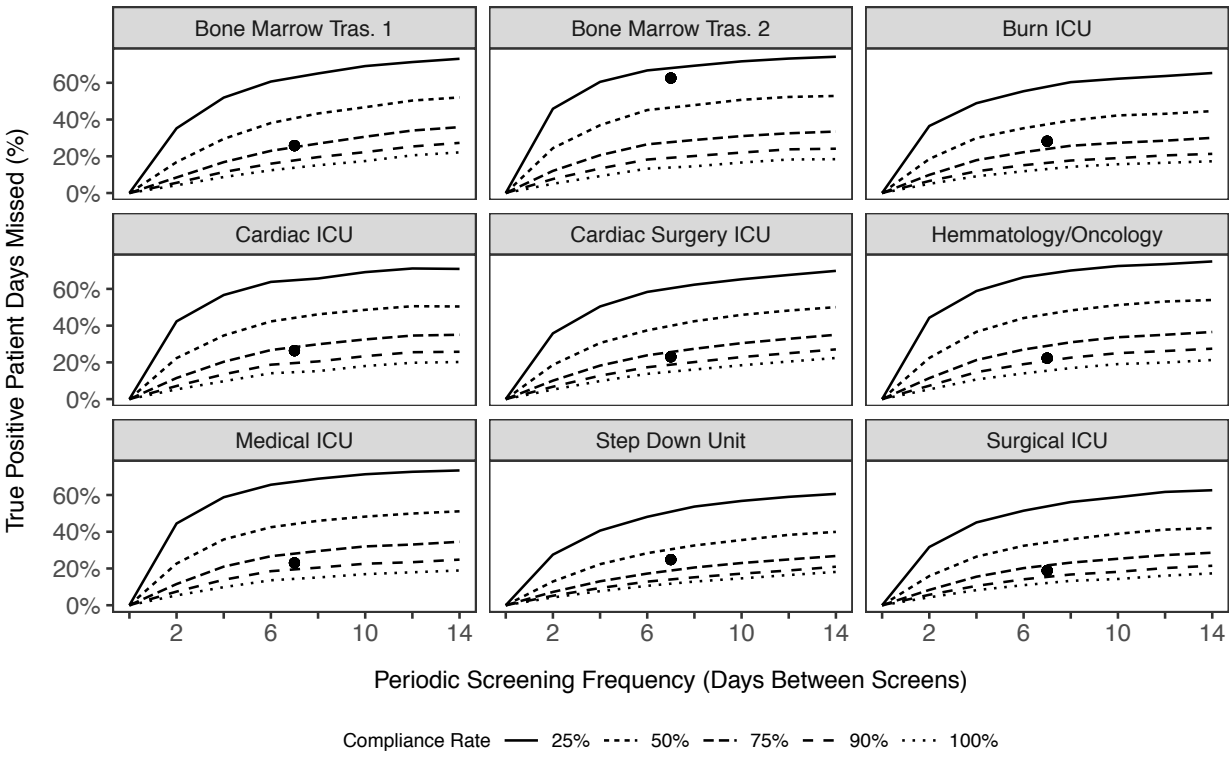

Figure 3.4 Comparison of varying compliance rates and screening frequency on the percent of VRE positive patient days missed in all screening units. Black dots indicate the weekly screening regime used with current unit compliance rates for comparison.

3.4.2 Increasing screening compliance rate maximizes the effect of detecting VRE

In order to identify an optimal testing strategy, we performed a cost effectiveness analysis, where cost was the number of tests conducted in a biweekly period (costs associated with the tests were considered to be additive) and effect was the percent of positive patient days detected. Both screening frequency and compliance were considered, where lower compliance rates would result in reduced cost due to missed screenings. Maximal efficiency in detecting VRE was characterized by screening regimes that had 100% compliance (Figure 3.5A). In our hospital, where compliance is around 75%, by increasing compliance to 100%, we would be able to identify 12 patient-days per 1,000 more than we currently do without making any other changes to current screening protocols (Figure 3.5B).

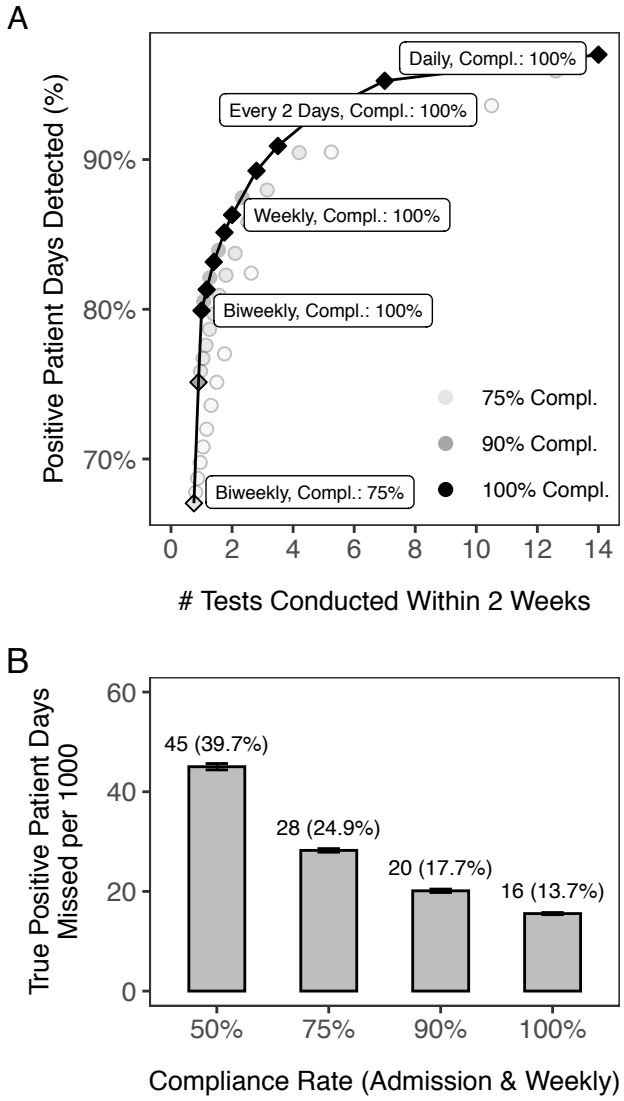

Figure 3.5 Testing strategies and hospital metrics of interest. Panel A: Cost effectiveness analysis comparing the relative costs related to testing with the percent of VRE positive patient days detected. Most incrementally efficient strategies are shown as diamond markers connected by a black line. All other strategies are shown as circle markers. Panel B: Incremental reduction in missed VRE patient days per 1,000 as compliance increases in screening units. Standard deviations shown as error bars.

3.5 Discussion

Using models and parameter estimates based on VRE screening data from a single university tertiary and acute care hospital, we project that testing patients upon unit admission

and weekly thereafter with perfect compliance would capture 86% of positive patient days in screening units. In UM hospital, this would reduce the proportion of missed positive patient days by 43%. In order to capture around 90% of cases, testing would have to increase to every four days, almost doubling the weekly rate. In general, these finding agree with the guidelines outlined by CDC and SHEA recommending admission and weekly screening for VRE. Our finding of identifying percent of VRE positive patient days detected also match a recent modeling study estimating the percentage of VRE cases captured through active surveillance with this same screening strategy (87%).[156]

In general, it has been difficult to tease out the efficacy of VRE interventions, such as isolating VRE positive patients, on reducing VRE transmission. The studies that have been conducted vary greatly in their efficacy. This may be due in part to low screening compliance rates, which are often between 60-70% but rarely reported. Of the studies available, we were unable to find any patient isolation intervention studies outside of model simulations where a screening compliance rate 90% or above was reported. However, in models where compliance was high, a reduction in VRE cases was observed where active surveillance was coupled with patient isolation procedures.[148,149] Slip ups in active screening have already been reported to cause silent outbreaks that can be difficult to contain once in motion.[157] Work that incorporates surveillance compliance rates is needed to better understand their effect on downstream patient isolation interventions.

In addition to exploring differences in screening compliance, this modeling study had a number of strengths. First, instead of simulating patients, we used actual patient location data from our hospital over the span of 8 years. This allowed us to use complex trajectories and patient-to-patient interactions, including ICU stays greater than two weeks that were not included

in previous modeling studies. In this hospital, patients with hospital acquired VRE on average were admitted for over a month, highlighting the need to include longer-stay patients to properly capture these cases. We also looked at each screening unit separately instead of combining into a single unit/entire hospital. For simplicity of execution and interpretation, many modeling studies simulate VRE spread within a single hospital ward.[148,149,156] By keeping the screening units separate, unit-based differences in length of patient stay and compliance highlight various magnitudes of change in capturing VRE cases.

Despite these strengths, our study had several limitations. Our model included a number of simplifications, including that the risk of a patient transmitting VRE was the same for all units and did not vary with time. Unit based differences can be seen in our calibration for both the cardiac surgery ICU as well as the medical ICU (Figures B.1-B.2). These differences likely can be attributed to differences in the underlying patient characteristics of these units. The simulation also did not vary risk for acquiring VRE based on other factors Patient, such as the unit they occupied or antibiotic use. Future models should consider how these factors vary based on the underlying patient populations or individuals at risk. We also focused our analysis on screening units instead of the entire hospital, which limits these findings to patients deemed at risk for VRE, instead of the hospital as a whole. Finally, our pathogen spread simulations did not consider the addition of isolation procedures once a patient was identified to be infected, as the effectiveness of these procedures is still a debated topic in the field. It is possible that use of downstream patient isolation may have lowered the overall number of VRE cases through reduced spread. Since admission plus weekly screening identified more than 90% of community acquired VRE cases (Figure B.14), it is possible that if isolation procedures are affective, we

could see a substantial reduction in cases and increase in identified cases without having to resort to greatly increasing testing frequency.

Chapter 4 Risk Factors for Increased COVID-19 Case-Fatality in the United States: A County-Level Analysis During the First Wave

This chapter is a published work:

Millar JA, Dao HDN, Stefopulos M, Estevam CG, Fagan-Garcia K, Taft DH, Park C, Alruwaily A, Desai AN, Majumder M. (2021). Risk factors for increased COVID-19 case-fatality in the United States: A county-level analysis during the first wave. PLoS One. 16(10):e0258308.

4.1 Abstract

The ongoing COVID-19 pandemic is causing significant morbidity and mortality across the US. In this ecological study, we identified county-level variables associated with the COVID-19 case-fatality rate (CFR) using publicly available datasets and a negative binomial generalized linear model. Variables associated with decreased CFR included a greater number of hospitals per 10,000 people, banning religious gatherings, a higher percentage of people living in mobile homes, and a higher percentage of uninsured people. Variables associated with increased CFR included a higher percentage of the population over age 65, a higher percentage of Black or African Americans, a higher asthma prevalence, and a greater number of hospitals in a county. By identifying factors that are associated with COVID-19 CFR in US counties, we hope to help officials target public health interventions and healthcare resources to locations that are at increased risk of COVID-19 fatalities.

4.2 Introduction

The Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) originated in Wuhan, China in November 2019 and has since spread to 210 countries worldwide.[158] By the last day considered for inclusion in the present work, June 12th, 2020, SARS-CoV-2 had caused over 2 million Coronavirus Disease 2019 (COVID-19) cases and 114,753 deaths in the United States (US).[159,160] The excess mortality from COVID-19 is likely to be underestimated, and recent work estimates 912,345 deaths from COVID-19 between March 2020 and May 2021, compared to the officially reported 578,555 deaths.[161] Early work on COVID-19 has highlighted patient characteristics that increase an individual's risk of death,[162,163] however it is unclear to which individual risk factors are best suited to understanding which populations are most at risk of high fatality rates from COVID-19. The distribution of infected cases and fatalities in the US has been heterogeneous across counties,[164] and identification of sub-populations at risk of increased morbidity and mortality remains crucial to effective response efforts by federal, state, and local governments.[165] Counties where governing officials are aware that their populations are at a higher risk of COVID-19 mortality, meaning the population experiences a higher case-fatality rate, may opt to tailor state policies or take earlier action to curtail the spread of SARS-CoV-2. Additionally, the federal government may opt to target vaccine resources to counties experiencing higher COVID-19 mortality rates.

The case-fatality rate (CFR) is defined as the number of deaths divided by the total number of confirmed cases from a given disease.[166] When a disease is non-endemic, the CFR fluctuates over time. During the beginning of an epidemic, there is often a lag when counting the number of deaths compared to cases and hospitalizations, leading to an underestimation of the CFR. Furthermore, CFR will fluctuate rapidly early in an epidemic when each additional case or

death has an excessive impact on calculating CFR. It is important to not only account for the lag between cases and deaths (i.e., lag-adjusted CFR), but also to ensure that the CFR is no longer fluctuating.

In this study, our objective was to use a lag-adjusted CFR to conduct a county-level mortality risk factor analysis of demographic, socioeconomic, and health-related variables in the US during the first wave of the COVID-19 pandemic (March 28, 2020 to June 12, 2020). This will provide critical information on what population characteristics are most informative to identify counties at high risk of experiencing high COVID-19 mortality rates. We expand upon prior work by considering possible risk factors of an increased CFR from multiple categories (e.g., non-pharmaceutical interventions such as shelter-in-place orders,[167] prevalence of pre-existing conditions such as cardiovascular disease,[168] and socio-economic circumstances such as hospital accessibility[169]) in a single model. This is also the first paper to focus on this range of risk factors during the first wave of the pandemic, so that results from this can be used for targeted intervention at the county level at the beginning of a pandemic.

4.3 Methods

All code for our work can be found on our GitHub repository.[170]

4.3.1 Study population

We conducted a cross-sectional ecological study to assess risk factors associated with an increased COVID-19 lag-adjusted CFR in US counties. Our study population included 3,004 counties or county-equivalents with Federal Information Processing Standards (FIPS), a unique

code for US federal identification (Figure C.1). Only publicly available aggregate data were used; therefore, no IRB approval was required.

4.3.2 County-level variables

We identified potential risk factors across several different categories: demographic, socioeconomic, healthcare accessibility, comorbidity prevalence, and non-pharmaceutical interventions. Each category-targeted risk factor relevant to the risk of COVID-19 mortality by conducting a comprehensive review of existing literature by March 28, 2020 supplemented with variables relevant to other respiratory epidemics.[171–173] Table C.1 provides detailed justifications for the inclusion of each risk factor. Only variables with publicly available data sources at the county- or state-level were included. Table C.2 lists data sources, variable descriptions, and manipulations (if applicable). We directly imported and cleaned the datasets using R (v3.6.3).

We included five demographic variables: total population, population density, the percentage of the population over age 65, the percentage of population 17 or younger, and race/ethnicity. All demographic variable data were from the 2018 American Community Survey 5-Year Data from the US Census annual survey, except for race/ethnicity data from the U.S. Census Populations with Bridged Race Categories.[174]

We included 13 socioeconomic variables, with their data primarily from the 2018 American Community Survey 5-Year Data.[175] In addition to the commonly used socioeconomic variables, we included certain variables contributing to the composite Social Vulnerability Index (SVI). The SVI was created by the Centers for Disease Control and Prevention (CDC) to describe US geographic areas by their social vulnerability and has been validated by multiple studies within and outside of the CDC[176–181] Social vulnerability is defined as "the characteristics

of a person or community that affect their capacity to anticipate, confront, repair, and recover from the effects of a disaster."[176] We included individual SVI variables on socioeconomic status, household composition and disability, minority status and language, and housing and transportation. We preferred to use the individual variables rather than overall SVI or by theme because we were most interested in understanding which components of social vulnerability contributed to increased CFR.

We included 5 healthcare-related variables: number of hospitals per capita, number of ICU beds per capita, number of primary care physicians per capita, percentage of residents without health insurance, and percentage of Medicaid eligible residents. Variable data was from the Kaiser Health News,[182] the Heart Disease and Stroke Atlas,[183] and the 2018 American Community Survey 5-Year Data.[175]

We included 18 comorbidity variables: diagnosed diabetes prevalence; diagnosed obesity prevalence; hypertension hospitalization and death prevalence, cardiovascular disease (CVD), chronic obstructive pulmonary disease (COPD), asthma, and cancer; Medicare beneficiaries with heart disease percentage, current smokers prevalence, and stroke-related hospitalization and mortality prevalence. Variable data was from the US Diabetes Surveillance System,[184] the Heart Disease and Stroke Atlas,[183] the Behavioral Risk Factor Surveillance System,[185] and the State Cancer Profiles by the National Cancer Institute.[186]

Non-pharmaceutical intervention data (including information on closing of public venues such as restaurants, gathering size limits, complete lockdown of non-essential activity in the county, if religious gatherings were included in gathering size limits, shelter-in-place orders, and social distancing mandates) were extracted from the COVID-19-intervention GitHub page, an open source data-sharing platform and compiled by Keystone Strategy.[187] However, this resource

does not cover all counties, thus missing data was supplemented from a variety of governmental executive orders and news articles detailed in the supplementary code. Variables with dates were transformed to how many days the event occurred after the first case in a county. States where an intervention never occurred were given a zero. Since 47% of all counties did not ban religious gatherings, data on when religious gatherings were banned in a county was transformed into an indicator variable (1 if the ban occurred, 0 if not).

4.3.3 Lag adjusted case-fatality rate (CFR) data and calculation

To calculate CFR during the first COVID-19 wave in the US, we obtained open access county-level COVID-19 data from the New York Times through June 12, 2020, the date the CDC released guidance for easing restrictions as states began to reopen.[159,188] Only data that contained FIPS county codes to identify case and death locations were included. County-level data for New York City, NY was accessed from the New York City Department of Health and Mental Hygiene.[189] To calculate lag-adjusted CFR (laCFR), we used Nishiura et al.'s method, expanded upon by Russell et al., to account for the delay between COVID-19 diagnoses and deaths.[190,191] We updated this approach by using time-from-hospitalization-to-death from the US population.[191,192] The final dataset included 1,779 counties with 1,968,739 cases and 106,279 deaths, comprising 96.8% of national cases and 96.8% of national deaths as of June 12, 2020.

During the first wave of the pandemic, SARS-CoV-2 was non-endemic, leading the case-fatality rate (CFR) to fluctuate over time. This is due to a lag when counting the number of deaths compared to cases and hospitalizations, leading to an underestimation of the CFR. The CFR continues to fluctuate rapidly early in an epidemic when each additional case or death has

an excessive impact on calculating CFR. It is important to not only account for the lag between cases and deaths (i.e., lag-adjusted CFR), but also to ensure that the CFR is no longer fluctuating.

To do this, we use a method developed by Nishiura et al. and expanded upon by Russell et al., where case and death incidence data are used to estimate the number of cases with known outcomes, i.e. cases where the resolution, death or recovery, is known to have occurred [33, 34]:[190,191]

$$u_t = \frac{\sum_{i=0}^{t} \sum_{j=0}^{\infty} c_{i-j} f_i}{\sum_{i=0}^{t} c_j}$$

where c_t is the daily case incidence at time t, (with time measured in calendar days), f_t is the proportion of cases with delay t between onset or hospitalization and death; u_t represents the underestimation of the known outcomes and is used to scale the value of the cumulative number of cases in the denominator in the calculation of the laCFR. Russell et al. used the estimated distribution in Linton et al., based on data from China up until the end of January 2020. For this study, we instead used United States centric data from Lewnard et al., which estimates the distribution of time from hospitalization to death based on data from Washington and California.[192]

Lewnard et al., fits the distribution conditionally on age resulting in a Weibull distribution for each age group.[192] The overall distribution was obtained empirically by weighting the densities at time t across all age groups. Because of this, the overall distribution doesn't have its own shape/scale parameters. However, we were able to estimate what these parameters would be by fitting a Weibull distribution that captures the 2.5, 25, 50, 75, and 97.5 percentiles (1.6, 7.3, 12.7, 19.8, 37.4), as well as the average (14.5).

Use of the laCFR assumes the measure has stabilized.[190] Counties where the laCFR is still rapidly changing cannot be used in the study as these are not unbiased estimates of the true

CFR. laCFRs were calculated incrementally for each day and assessed whether they changed on average less than 1% a week for the last two weeks of available data. The final calculation based on all data available was used as the laCFR in our model.

4.3.4 Statistical analysis

To reduce multicollinearity, we eliminated linear combinations and variables with correlations >0.5 using the R package caret (v6.0.86). Remaining variables were screened for missingness and missing values were imputed using five imputations in the R package mice (v3.8.0).[193] Data were randomly split into training (1,186 counties) and testing sets (593 counties) to assess generalizability (a table of the characteristics can be seen in Table C.3). A negative binomial linear model with an offset for the number of COVID19 cases per county was chosen based on Kolmogorov-Smirnov and dispersion tests found in the R package DHARMa (v0.3.1). Variable selection was conducted using purposeful selection, an iterative process in which covariates are removed from the model if they are neither significant nor confounders.[194,195] With clinical risk factors, purposeful selection outperforms other variable selection procedures and tests for the presence of confounders.[194] Removing highly correlated variables beforehand reduces the chance of multicollinearity between non-significant variables that may have been retained in purposeful selection due to confounding effects. Per Bursac et al., we used the 0.1 α-level for initial selection using bivariate models and a change of >20% in any remaining model coefficients compared with the full multivariate model for confounding evaluation.[194] All variables in the final model were significant at the 0.05 α-level, and no statistical confounders were included in the final model.

4.3.5 Model fit

We observed the fit of the model using a half-normal plot (Fig 4.1A). The simulated envelope for the deviance residuals in the half-normal plot serves as a guide of what to expect under a well-fitted model, with most of our model's deviance residuals lying within.[196] We compared the mean and variance seen within our model predictions to the theoretical mean and variance expected in a Poisson and negative binomial model. After grouping the fitted predictions into 20 quantiles and calculating their means and variances, we saw the negative binomial model captures our data variance well.[197] Loess smooth was used for the empirical mean (Fig 4.1A). As an additional check, we calculated the ratio of Pearson residuals to degrees of freedom, which was 1.04, indicating we accounted for most of the over-dispersion in laCFR using the negative binomial model. The Cox and Snell Pseudo R2 for our model was 0.86, which accounts for the majority of the variance present in our outcome variable. All variables had a variance inflation factor of less than 2, indicating collinearity was not an issue with our variables (Table C.4).

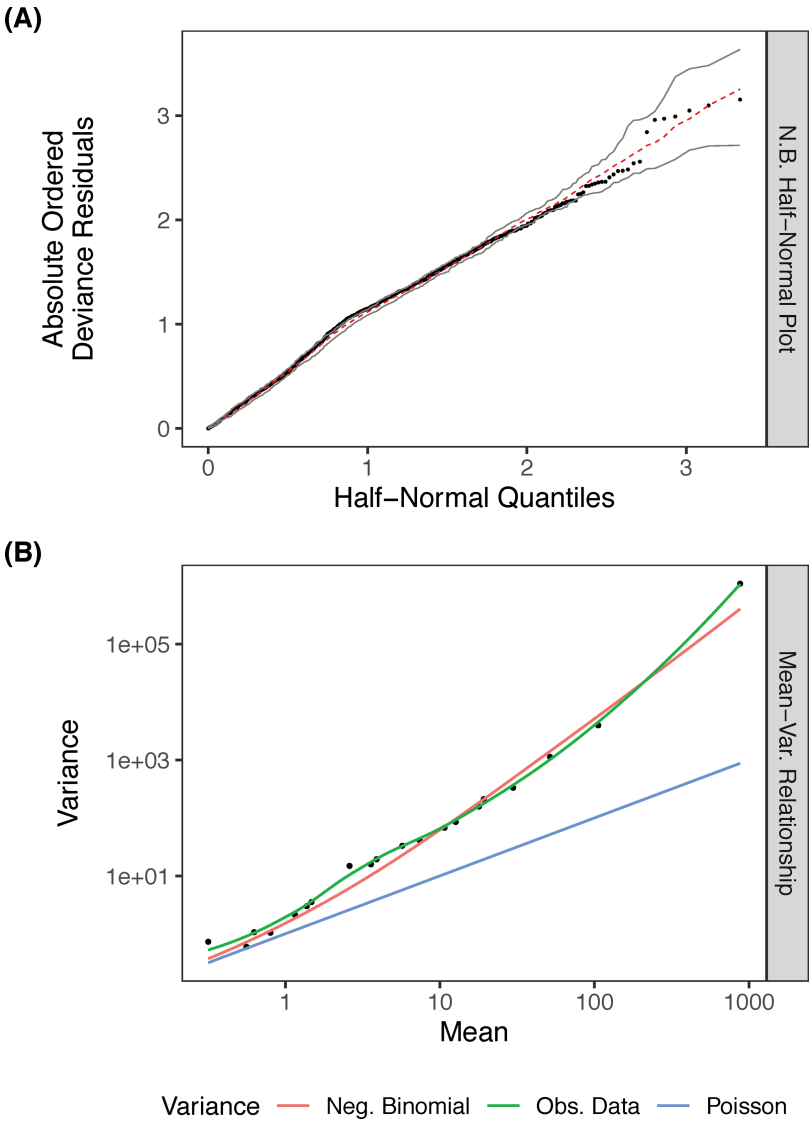

Figure 4.1 Assessing model fit. Plots showing (A) half-normal residuals and (B) mean-variance relationship of the observed county-level COVID-19 laCFRs.

We checked the coverage, which is the probability that our model outcomes are found within our prediction interval. To estimate our predictive coverage (empirical coverage), we simulated a prediction interval. The coverage was 0.9730 for the training data and 0.9713 for testing data (Figure 4.2A). Similar to ROC, a gain curve plot measures how well the model score sorts the data compared to the true outcome value.[198] When the predictions sort in exactly the

same order, the relative Gini coefficient is 1. When the model sorts poorly, the relative Gini coefficient is close to zero, or even negative. The relative Gini scores were high for both our training set and testing set. (0.9840 and 0.9829, respectively, Figure 4.2B).

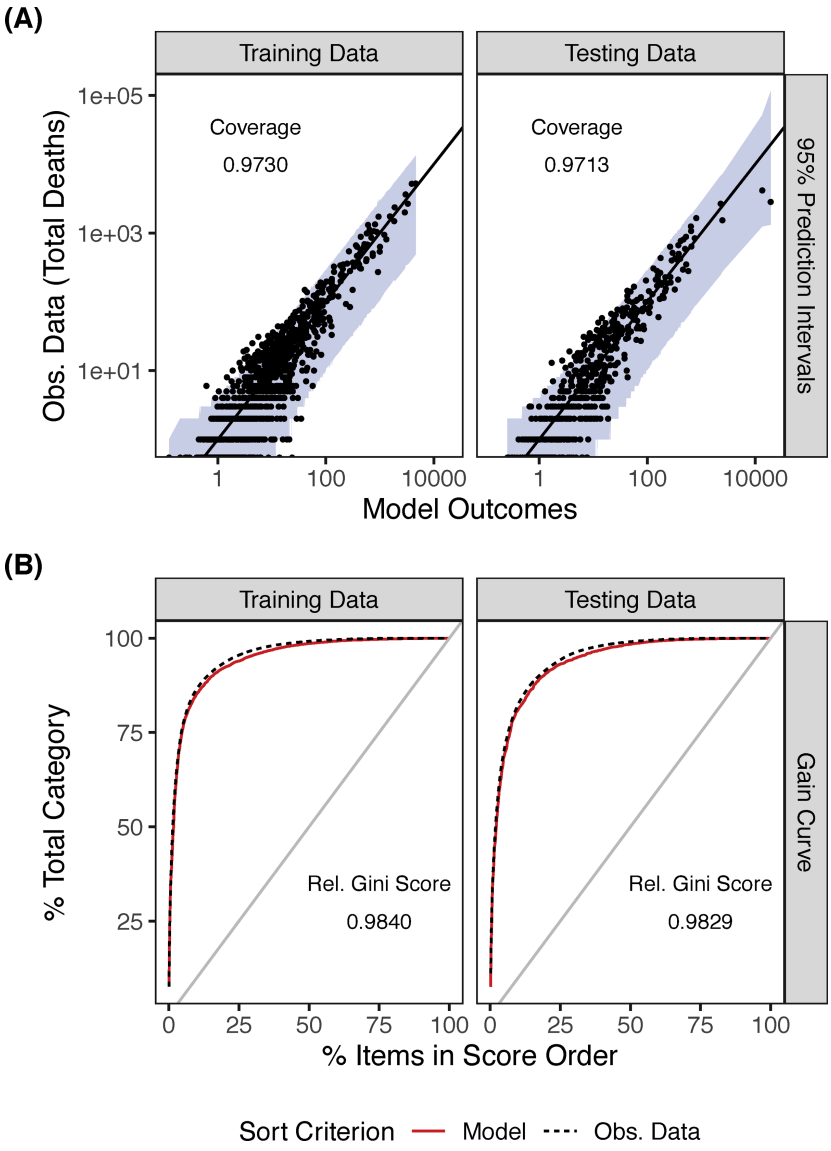

Figure 4.2 Assessing model generality. Plots showing (A) model outcomes found within the prediction intervals for training data and testing data for the county-level COVID-19 laCFRs and (B) gain curves for training data and testing data for the county-level COVID-19 laCFRs.

4.4 Results

Of the 64 variables collected, 22 were retained for analysis after minimizing correlation (Tables C.1-C.4). Multiple imputation was used to correct for missingness (less than 2%) in two of the retained variables, neither of which appeared in the final model. Fifteen variables were significant in bivariate models in the first step of purposeful selection, and were included in the initial multivariate model. Eight variables were significant in the initial multivariate model and were retained in the final model. Including variables that were non-significant in the bivariate models with these eight variables did not significantly change the performance of the model, as determined by the Likelihood Ratio Test. No potential confounders were identified among the correlation minimized variables that were previously discarded due to non-significance in the models.

The final model is shown in Table 4.1. The negative binomial model appears to be a good fit, capturing the mean-variance relationship observed in the data and displaying expected residuals (Figure 4.1A and 4.1B). The model was well-calibrated, with the training and testing model having comparable coverage and relative Gini score (Figure 4.2A and 4.2B). Since we used a negative binomial model with an offset, the exponentiated coefficients represent the change in laCFR observed for a one-unit increase in each continuous variable, assuming all other variables in the model are held constant. Four variables were inversely associated with laCFR: number of hospitals per 10,000 people (-32% laCFR per additional hospital per 10,000), banning religious gatherings during the initial state or county shutdown (-13% laCFR if religious gatherings were banned), percentage of housing units that were mobile homes (-0.79% laCFR per 1% increase in the proportion of mobile homes), and percentage of population without health insurance (-1.5% laCFR per 1% increase in percentage uninsured). Four variables were directly

associated with laCFR: percentage over age 65 (+4.5% laCFR per 1% increase in population over age 65), percentage Black or African American (BAA) (+0.97% laCFR per 1% increase in BAA population), percentage with asthma (+9.5% laCFR per 1% increase in asthma prevalence), and number of hospitals (+3.2% laCFR per one additional hospital). Figure 4.3 demonstrates the relationship between each variable and the laCFR over a range of values. We have stratified these variables further for comparison, and results can be found in Figures C.2-C.8.

Variable	Coefficient	95% CI	p-value
Intercept	-4.5	(-5.1, -3.9)	<0.001
Percentage population aged 65+	0.044	(0.030, 0.059)	<0.001
Percentage population Black or African American	0.0097	(0.0063, 0.013)	<0.001
Hospitals per 10,000 persons	-0.39	(-0.59, -0.19)	<0.001
Asthma prevalence	0.091	(0.039, 0.14)	<0.001
Total number of hospitals	0.031	(0.0099, 0.054)	0.0017
Ban on religious gatherings indicator	-0.13	(-0.24, -0.030)	0.011
Percentage housing stock that were mobile homes	-0.0079	(-0.015, -0.0011)	0.024
Percentage population without health insurance	-0.015	(-0.029, -0.00021)	0.052

Table 4.1 Parameter estimates for the final multivariate model of laCFR.

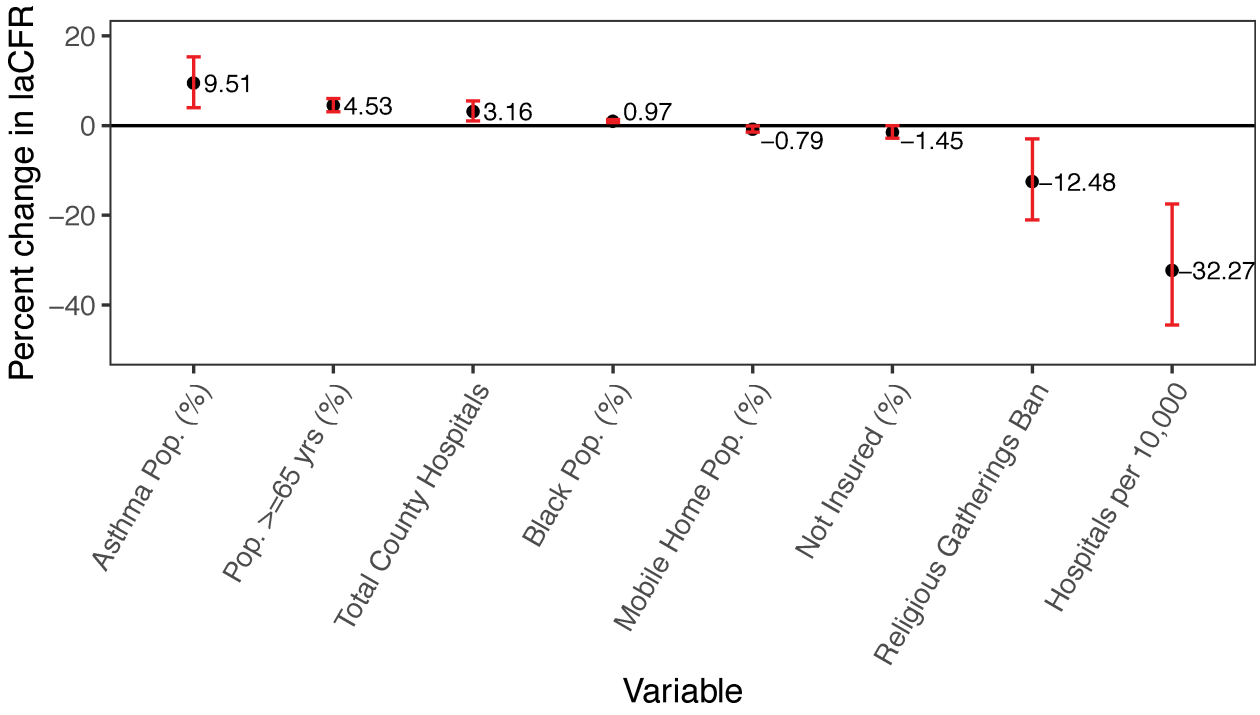

Figure 4.3 Percentage change in COVID-19 laCFR given a 1 unit increase in the variable for each individual variable. Estimate shown as black dots. 95% confidence interval shown in red using training data.

4.5 Discussion

In this ecological study of mortality due to SARS-CoV-2 infection during the first wave of COVID-19 in the US, we found that county-level laCFR was significantly associated with eight variables. Four variables–banning religious gathering, proportion of mobile homes, hospitals per 10,000 persons, and proportion of uninsured individuals in a county–were associated with decreased laCFR. Four variables–percentage of population over age 65, total number of hospitals per county, prevalence of asthma, and percentage of population BAA–were associated with increased laCFR. Each variable provided unique insights into factors that may be worth considering for county-level COVID-19 response efforts.

4.5.1 Inverse association with case-fatality rate

Our model indicated a 13% reduction in the average laCFR for counties that banned religious gatherings compared to counties that did not. Gatherings often involve dense mixing of people in a confined space, sometimes over long periods of time,[199] which drives COVID-19 transmission.[200] Interventions targeting increased physical distancing and limiting contact were introduced in some countries, including the closure of schools, places of worship, malls, and offices.[199] Our model suggests that specifically exempting religious gatherings from bans may increase the laCFR, consistent with the combination of findings that [1] religious gatherings across the globe were linked to COVID-19 superspreader events[201] and [2] older Americans (who are more likely to attend religious services than younger Americans[202]) are at increased risk of death due to COVID-19.

The percentage of the population living in mobile homes was also associated with a decrease in laCFR. A 1% increase in mobile home living was associated with a 0.79% decrease in laCFR. While a small difference at first glance, it becomes more meaningful when considering the large variation in mobile home living across counties. Between counties at the 25th percentile of percentage living in mobile homes (4%) and 75th percentile (18%), the difference in the percentage of mobile home living correlated with an 11% decrease in laCFR. This might represent a built-environment effect, given that mobile homes have separate plumbing and ventilation unlike apartments and other multi-family residences. Recent work suggests that fecal aerosol transmission of SARS-CoV2 can occur.[203] Ventilation patterns in apartment complexes represent additional opportunities for transmission.[204] The benefit of separate units such as mobile homes may be especially important to low-income workers who are both more at risk of

death from COVID19 due to increased chance of having a co-morbid condition and more likely to live in multi-family housing with maintenance issues.[205]

The number of hospitals per 10,000 was also inversely associated with laCFR. We found that for each additional hospital per 10,000 inhabitants, the laCFR decreased by 32%, despite the exclusion of other healthcare-related variables due to non-significance (e.g., ICU bed availability). Prior work demonstrated that the percentage of ICU and non-ICU beds occupied by COVID-19 patients directly correlated with COVID-19 deaths[206] and a county with more hospitals per 10,000 inhabitants may be able to cope with more COVID-19 cases before reaching the same percentage of hospitals beds occupied as a county with fewer hospitals per 10,000 inhabitants. Furthermore, because adding beds requires fewer resources than adding hospitals, the number of hospitals per 10,000 persons in a county might represent a greater ability to expand capacity. As a result, using hospitals per 10,000 may be a better indicator of healthcare capacity than the number of ICU beds early on in the pandemic. Because healthcare resources in the US correlate with community wealth,[207] the rate of hospitals per 10,000 may also reflect increased community wealth and the protective effect of higher socio-economic status on health. More hospitals per 10,000 persons may also represent increased competition for patients, which is associated with decreased mortality from community-acquired pneumonia.[208]

Unexpectedly, the percentage uninsured was inversely associated with laCFR. We found a 1.5% reduction in laCFR for every 1% increase in uninsured inhabitants. Prior studies found longer travel times to COVID-19 testing facilities were directly associated with percentage uninsured.[209,210] Because uninsured persons may be unable to readily access testing, this finding may relate to incomplete reporting, such that only individuals who survive long enough are

tested for COVID-19, leading to a potential undercount of deaths attributable to SARS-CoV-2 infection.

4.5.2 Direct association with case-fatality rate

In our model, a 1% increase in the population over 65 years old was associated with a 4.5% increase in average laCFR. This is consistent with recent epidemiological studies demonstrating an association between the severity of COVID-19 infection and age. According to provisional death data from the National Center of Health Statistics, people aged 65 and older have a 90- to 630-fold higher risk of mortality due to COVID-19 than 18-29-year olds.[211]

Also, directly associated with laCFR was the total number of hospitals per county, with an observed increase of 3.2% in average laCFR per additional hospital. This variable was strongly correlated with total population ($r = 0.92$). Given that the number of hospitals per 10,000 was associated with decreased laCFR, this correlation suggests that total hospitals might be a proxy indicator for total population. Previous work assessed population density as a risk factor for increased laCFR, but not total population [43]. Since our analysis focused on the first wave of COVID-19, this variable could reflect overwhelmed healthcare systems in highly populated counties where most of the COVID-19 cases initially occurred.[212]

Asthma prevalence was also directly associated with laCFR. A 1% increase in asthma prevalence was associated with a 9.5% increase in laCFR. Evidence regarding asthma as a risk factor in COVID-19 is mixed. Although the US CDC has determined that patients with moderate to severe asthma belong to a high-risk group,[213] the Chinese CDC indicated that asthma was not a risk factor for severe COVID-19.[214] One study showed that COVID-19 patients with asthma were of older age and had an increased prevalence of multiple comorbidities compared to those

without asthma,[215] but that the presence of asthma alone was not a risk factor for increased mortality.[215] Thus, despite our findings, it is unclear whether asthma has a direct impact on COVID-19 disease or if other factors may be associated with both asthma and COVID-19. One such potential confounder is exposure to air pollution, as air pollution is associated with both asthma and risk of death from COVID-19.[216]

Finally, laCFR was directly associated with the percentage of the population identifying as BAA in a county. Our model showed that a 1% increase in BAA was associated with a 0.97% increase in the laCFR. This likely reflects the effects of structural racism in the US, where BAAs have fewer economic and educational opportunities than White Americans and as a result are exposed to increased risk of morbidity and mortality from COVID-19.[217] Dalsania et al. also found that the social determinants of health contributed to an unequal impact of the COVID-19 pandemic for BAA at the county level.[218] A study by Golestaneh showed that US counties with BAA as the majority had three times the rate of infection and almost six times the rate of death as majority White counties.[219] Factors underlying this trend include years of structural racism resulting in a lack of financial resources, increased reliance on public transportation, housing instability, and dependence on low-paying retail jobs.[220] Our approach considered several other variables that might explain the effect but were either non-significant (e.g., household crowding, percentage of households without a vehicle, and county land area) or were correlated with percentage BAA (e.g. percentage single parent households and percentage living in poverty), further emphasizing the role of systemic racism in COVID-19 laCFR.

4.5.3 Excluded predictor variables

In reducing multicollinearity and using purposeful selection, several variables were surprisingly excluded. One of these excluded variables was population density, although higher population density had been hypothesized to increase contact rate and non-adherence with physical distancing.[200] Diabetes and cardiovascular disease were excluded, despite multiple studies reporting these conditions as risk factors for COVID-19 mortality.[221,222] While these factors are important at an individual level to assess the mortality risk, our model suggests that other variables may be more informative at the county-level, underscoring the value of ecological studies.

4.5.4 Study strengths and limitations

This study had several strengths besides the benefits of an ecological design when considering population interventions. First, the data were nationally representative, including over 50% of all US counties. Our model captured the variability in the data and accounted for the observed data distribution. The model also captured almost all outcomes within the prediction interval for both training and testing data sets, with similar accuracy between them, which indicates that our model is generalizable within the US. Additionally, our model based laCFR calculations on the distribution of times from hospitalization to death from US data,[192] which differed from earlier Chinese data.[214] Using US-based distribution of times likely improved our laCFR estimation for this study. The final model included several variables previously attributed to higher laCFR (such as older age)[211] and included a variable unique to the pandemic shutdown, i.e., banning religious gatherings, giving more nuanced insights into heterogeneous COVID-19 mortality rates across counties.

Despite these strengths, our study had several limitations. First, under-reporting of cases might affect the accuracy of CFR calculation.[223] The reported cases and deaths we used likely underestimated the true COVID-19 parameters. This underestimation was more among the asymptomatic and mild cases due to limited testing capacity and changes in testing practice; hence, the laCFR might have appeared inflated. Second, the type and timing of the tests used may have impacted the measured laCFR. Samples collected early during the infection can yield higher false negatives with RT-PCR tests.[224] False negatives in critically-ill patients who later die could decrease the measured laCFR unless probable COVID-19 deaths are reported, while false negatives in mild cases who are not retested later could increase the measured laCFR as survivable cases go undetected. These are challenges for any CFR study and highlight the ongoing need for improved COVID-19 testing. Third, COVID-19 reporting practices vary widely by state. For example, Florida was found to report fewer COVID-19 deaths in the official tally than the Medical Examiners Commission.[225] In addition to deliberate underreporting of deaths, states also vary in reporting of probable cases and deaths.[226] Without national standards in the COVID-19 response, comparing case counts and deaths across state line–let alone county–is deterred by lack of clarity about how these data differ.[226]

Beyond these, our study was also limited by the fact that relevant data were frequently unavailable, including data on non-pharmaceutical interventions (NPI) and comorbidities. To limit missingness in the NPI data, we used state-level data when available given that counties also enforce state-level orders. However, there may be heterogeneity between county- and state-level information making this a less effective approach. Other variables of interest were not available at the state- or county-level, including information on contact tracing efforts and community compliance with public health mandates. Funding to collect public health

information on more variables at a granular level would improve the information available to guide decision-making during emergencies. Another limitation was the highly correlated nature of the 64 variables considered for inclusion. Multicollinearity greatly affects the interpretability of coefficients and is rarely accounted for in epidemiologic studies.[227] Highly correlated variables in a model are unstable and can bias standard errors, leading to unreliable p-values and unrealistic interpretations.[227] Because we ensured our model interpretability by excluding highly correlated variables, not all of our collected 64 variables were screened for inclusion in the final model.

Finally, our study period ended in mid-June. Recent work has divided the COVID-19 pandemic in the USA into three waves, with the first wave running from late March 2020 until mid-June 2020.[228] The exact day of June 12, 2020 was chosen because [1] enough cases had occurred in the US to obtain reliable estimates of laCFR by county and [2] it preceded CDC reopening guidance and a shift in reporting to the HHS Protect system, which is less readily available to the public than the prior CDC reporting system.[229] The decision by the government to switch to the HHS Protect system hinders the ability of academic scientists to aid in the response to the on-going pandemic.[229] Making these data more readily available to the public would permit inclusion of additional data for future research.

4.6 Conclusion

This study highlights several variables that were associated with county-level laCFR during the first wave of COVID-19 in the US. Though further research is needed to examine the effects of additional NPIs, our work provides insights that may aid in targeting response and vaccination efforts for improved outcomes in subsequent waves.

Chapter 5 Discussion

Over the course of my time as a doctoral student, I have been reflecting on my training and growth as a scientist. Over time, my previous experiences gained earning degrees in biology, epidemiology, statistics, and bioinformatics have converged into my identity as a data scientist. This has been solidified by the training process I have been privy to the structure of our doctoral bioinformatics program at University of Michigan. Instead of focusing on specific problems, the program allowed us to be broad thinkers, focusing on identifying methodologies appropriate to different problems. This idea was exemplified in our Preliminary Exam process to become a doctoral candidate. Instead of presenting on what we would work on during our dissertation, we needed to defend an idea for a scientific study that was different from what our lab worked on. The point was to test if we could address a problem by using the appropriate methodology and be able to defend it to established scientists in the field, without relying on practices already established in our lab. This preparation allowed for an easier transition into working on a number of very different studies for my doctoral work. These different projects highlighted different problems, different scales, and different constraints that added to my breath of training and flexibility working in different environments.

In chapter 2, where I studied T cell activation in tuberculosis infection, I took advantage of a computational model of the immune system established in the lab for over a decade. Instead of starting from scratch, I started with a model with many moving parts, thousands of lines of code, that is actively worked on by others in the lab. The problem itself was incredibly

interesting, but working in this group environment proved to be an even more invaluable experience. Learning to work with an established computational model, contributing new features to the model, and using the collaborative knowledge in the lab to solve a problem helped me to become a more collaborative scientist. I worked within the constraints of an established system, but with the flexibility to change and improve. All the moving parts made the approach a little trickier, but the results were better for all the combined knowledge. This experience also allowed me the opportunity to contribute my statistical training to improve upon the analysis. While it wasn't the main focus of my project, I was happy to contribute statistical theory and improve some processes so that all people working on the project could benefit. This resulted in a number of other publications within the lab that I participated in.

In chapter 3, hospital VRE screening compliance, this project came from an idea I had while working with a large electronic health records dataset for a different project. While attempting to model the transmission rate of VRE within our hospital, it occurred to me our estimates were only as good as the data we have available and I wondered how compliant our hospital was with screening. These musings led to questions such as: how does screening compliance affect the models our lab was trying to build, and in turn, how much would it affect other aspects within the hospital surveillance programs, such as downstream isolation precaution effectiveness. Eventually these concepts began to coalesce over the next few years. Turning to the literature, I quickly found that very few publications mentioned the subject of compliance in VRE screening and it had never been explored in any great detail.

Once questions were pinned down, the methodology became a collaborative process, coming from both my advisor's extensive domain knowledge on the subject and my own training in statistics. The project became a balancing act of the sweet spot of how much data to show and

what information will be most useful in which context to which audience. Not knowing the main goal of the project from the beginning and having to discern the metrics of success while figuring out the goal made this one of the most challenging projects, but incredibly rewarding to see my findings help fill a hole in the literature, starting from a vague musing of data we had years earlier.

In chapter 4, where I investigated risk factors for COVID case fatality rate, I joined an project where the question was established, but where the methods were not. I first joined as a volunteer statistician. There were some ideas of methods, but the goals seemed to be muddled and the methods did not make sense for the type of data available. A few months in, I pulled the team back and created a presentation to walk through and define our aims, what could be done given the available data and what we hoped to contribute to the field. I found myself at the edge of my knowledge base and fell back on my statistical foundation to give me enough to go in the right direction and learn. I spent weeks researching methods, creating our workflow, correcting mistakes, and walking through interpretations. From that point on, I helped lead the group of volunteers in the design and execution of the project. I realized for the first time that I was leading a diverse group of graduate students and postdocs in research. At that moment, I knew I had what it takes to be a scholar and help create and shape our field, and that I could lead a team and strengthen our workforce.

5.1 Future directions of projects

Each of the projects contained in this dissertation not only advanced the field and my education, but they also created open directions for future work. In the Mtb project I explored T cell activation at the site of infection, and found recruitment of non-specific T cells and

granuloma spatial characteristics contributed to crowding out of the few Mtb specific T cells within the granuloma cell, reducing the chances they could interact with and be activated by infected macrophages. It might be hard to experimentally affect spatial characteristics of the granuloma to allow easier migration of T cells toward macrophages where they can be activated. However, it might be possible to increase the amount of Mtb-specific T cells that are recruited. There is also the question of the level of Mtb-specific T cells that would be needed to clear an infection. A recent study suggested non-specific T cells are required to keep the immune response going past the initial infection with Mtb.[230] A next step we could take would be to conduct more simulations of Mtb infections with higher levels of antigen specific T cells to find the level needed for Mtb clearance.

In the VRE project, I explored ways to improve detection of patients colonized with VRE within a hospital. My analysis revealed that increasing the compliance rate of screening under any current active surveillance strategy maximized the efficacy of identifying VRE cases. An ongoing question in the field is whether isolating patients reduces the spread of VRE.[231,232] Since few studies report the screening compliance of their hospitals and those that do are generally around 60-70%, it's difficult to address whether isolations procedures are effective.[116,142,144,146] If VRE cases are being missed in screening, it's likely these cases continue to contribute to hospital spread, undermining downstream interventions, and obscuring evidence for their efficacy. In the future our work should expand to include the role of compliance and downstream isolation procedures and investigate whether these have an effect on optimal strategies.

In the COVID project, we were interested in identifying risk factors for dying from COVID if infected, is the case fatality rates (CFR). My study agreed with previous findings, including relationships between asthma prevalence and CFR, but also contributed new findings

on risk factors, including reduced CFR with bans on religious gatherings. An important limitation of this study was that it only focused on the first wave of COVID in the United States. Future work to investigate whether these risk factors change, or stay the same as the US entered successive waves and COVID became more endemic. It would also be interesting to see how newer interventions and different sub lineages of COVID changed these risk factors and how quickly these shifts occurred.

5.2 Personal impressions

There were a number of impressions I was left with from the role as a data scientist on the projects in my dissertation and others in which I had supportive role.

1) The process is rarely linear. In chapter 1, the overarching approach for the data analysis was laid out. But projects rarely follow this approach step by step. In reality, I often ran into issues at various steps, used research methods, tried new things, and circled back. This process often generated new ideas, which were addressed within the current project or made clear for future endeavors.

2) Data analysis is often a compromise if you have no say in the data collection. There is a saying from the statistician Sir Ronald Fisher that goes along these lines: asking a statistician after the fact is like asking a mortician what could have been done.[233] Unfortunately, many times data scientists don't have a say in the creation of the data, as it has already been collected. In other cases, resources, methodology constraints, or even ethical reasons may limit data collection. In all these cases, you do your best to meet the needs of the project with the data you get. But, in doing so the assumptions incorporated into the work need to be clearly communicated. The data scientist may not be able to measure exactly what was asked (e.g. a

proxy), precision may be sacrificed, or you may only look at an aggregate level, among many other compromises that can arise.

3) Data quality checks are a must. As data streams expand, it becomes impossible to observe every single data point. Working out data assumptions and testing them becomes one of the most important steps in data processing, which is rarely covered in analysis classes.

4) It's important to know the boundaries of your understanding. New projects often expand into areas we have not worked in before. This can mean a new type of data, or odd quirks or assumptions that needs to be addressed in the data. Our training gives us a foundation to work from and expand upon. It is crucial to understanding that you have the ability to learn and to expand upon your knowledge, though it may take time. Knowing when the required domain knowledge is beyond your current abilities or if the time frame needed to develop sufficient skill is unfeasible for the project will help you know whether to decline a project or to reach out for help from a collaborator.

5.3 Future of the field

Looking ahead as bioinformatics and data science in general evolves, there are a number of things we need to be aware of. One is new types of data and analyses are being released all the time. The avalanche of data types has continued in the past few decades.[234] Along with this growth comes new, more complex methodologies to handle the data. Many people jump on a new methodology bandwagon without fully digging into the methods and assumptions that it was built upon. It is paramount to be cautious when applying a new methodology, especially one that is still under development, as all the kinks are rarely worked out completely in the first paper.

Another area that is often overlooked, is the amount of support needed for secondary analysis of data, especially using government datasets. Much of the public health field is working with secondary data that were acquired using complex sampling methodologies. Many of the new analysis methodologies being released have not been adapted to accommodate the assumptions of these datasets. An example of this can be seen in using various machine learning methods that sample from the data indiscriminately without considering the complex covariance structure generated by the sampling.[235] More training needs to be provided to data scientist on what can and cannot currently be done with this data, because these errors are propagated through studies conducted by both early and established career scientists.[236–238]

This leads me to one of the biggest challenges I see. In general, many students that become scientists are given the bare minimum of statistical training that inevitably lead to inappropriate statistical analysis and conclusions.[239–242] This can be especially problematic in the data science field, which is filled with boot camps that try to get you up to speed by attempting to teach foundational knowledge in just a few weeks. The literature is littered with studies that use methodology incorrectly,[243–250] which include peer reviewed papers. This has only increased by offering user friendly analysis packages, creating a black box of the methodologies, and users often using only default settings. More care needs to be put into explaining the issues, assumptions, and correct scenarios to use analysis methodologies. Understanding how to interpret statistics and judge their appropriateness needs to be taught more widely. We can start by including non-perfect data scenarios into training and teach how to work through those problems as this is almost always skipped over.

If we're not careful, all of these advances will produce a plethora of studies, but the signal of what is good and useful will be lost in the noise.

Appendix A Supporting Information for Chapter 2

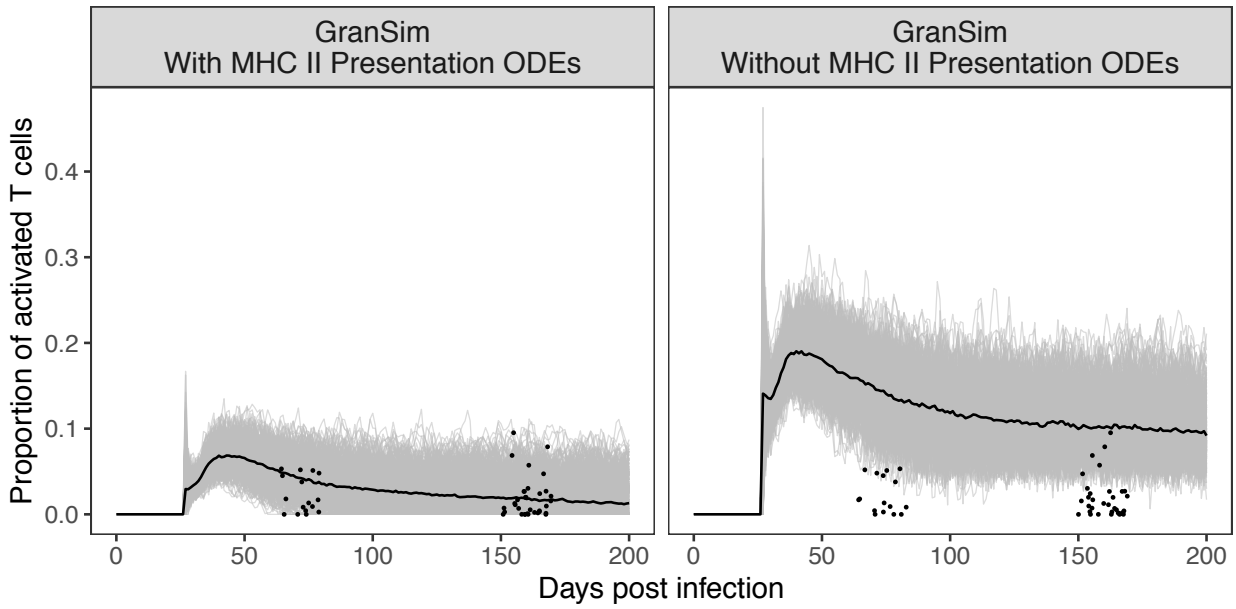

Figure A.1 Comparison of proportion of activated T cells in *GranSim* simulations with MHCII presentation ODEs and without. Median is shown in black, with simulation runs shown in gray. NHP data is overlaid for each output.[39]

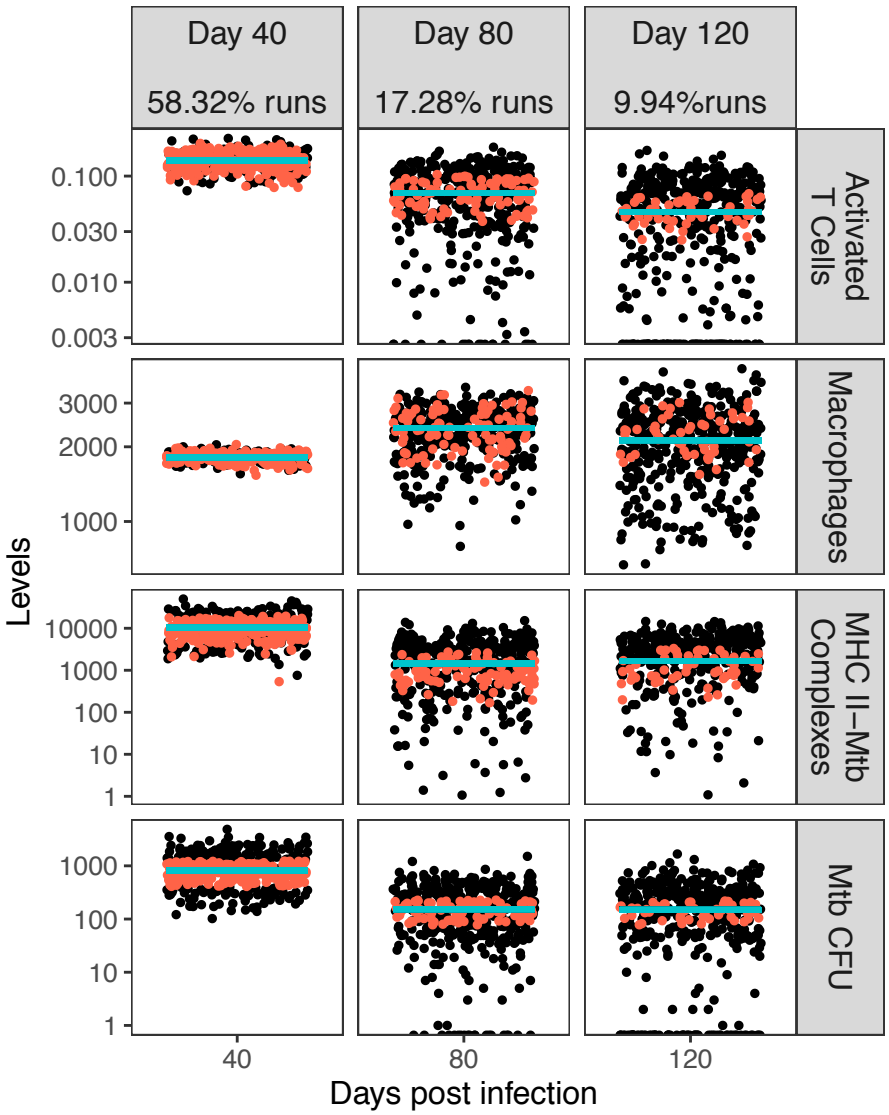

Figure A.2 Chosen runs of *GranSim* simulations (red) to best representation of the median values for outputs of interest (blue). *GranSim* runs that were not chosen are shown in black. Percentage of runs chosen is shown at the top of each column. For each 4 outputs, each of the red dots fit within a +/- 50% window of the median blue line and are chosen as good representations of all runs.

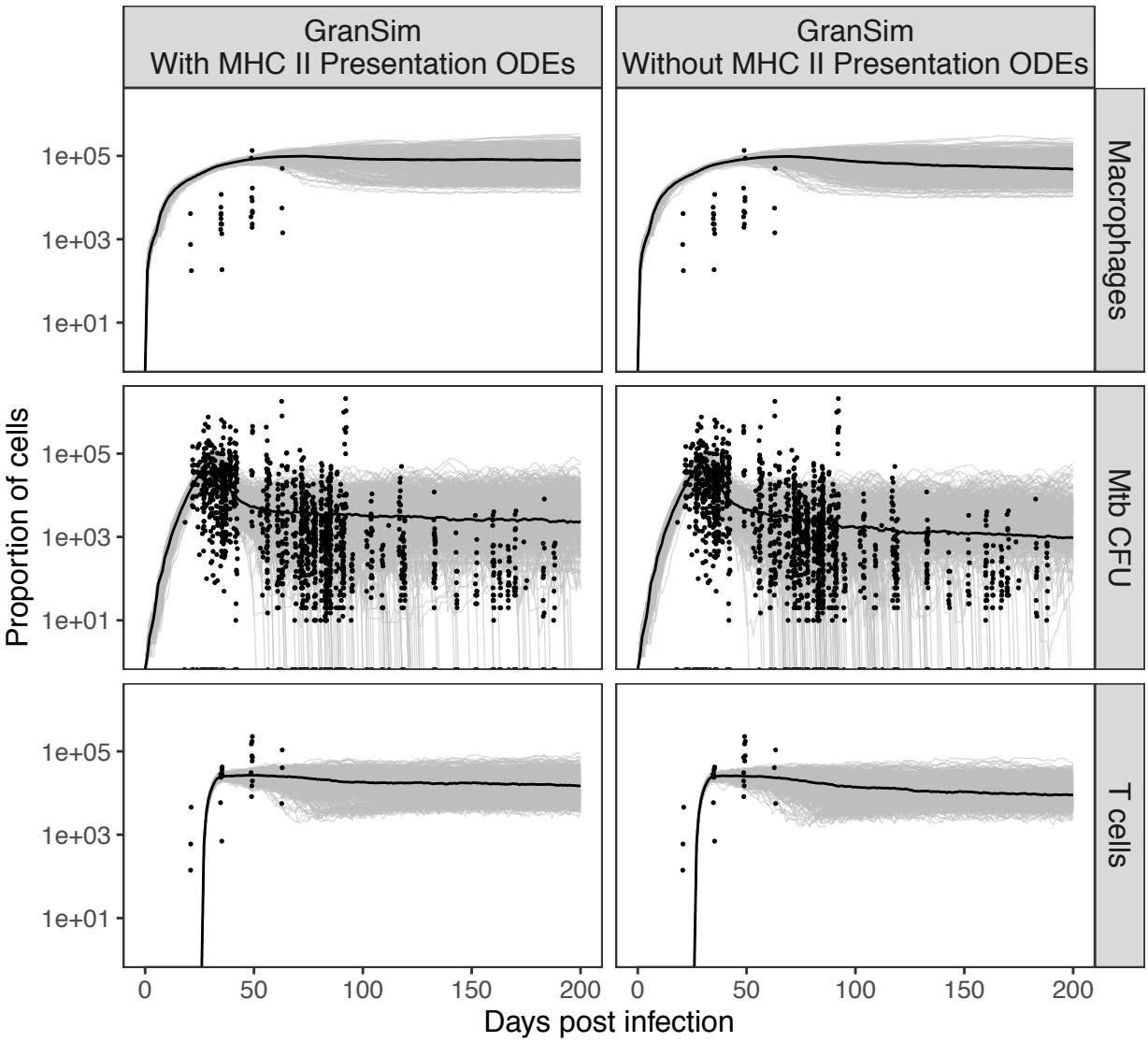

Figure A.3 Comparison of *GranSim* simulations with MHCII presentation ODEs and without. Median is shown in black, with simulation runs shown in gray. NHP data is overlaid for each output.[86]

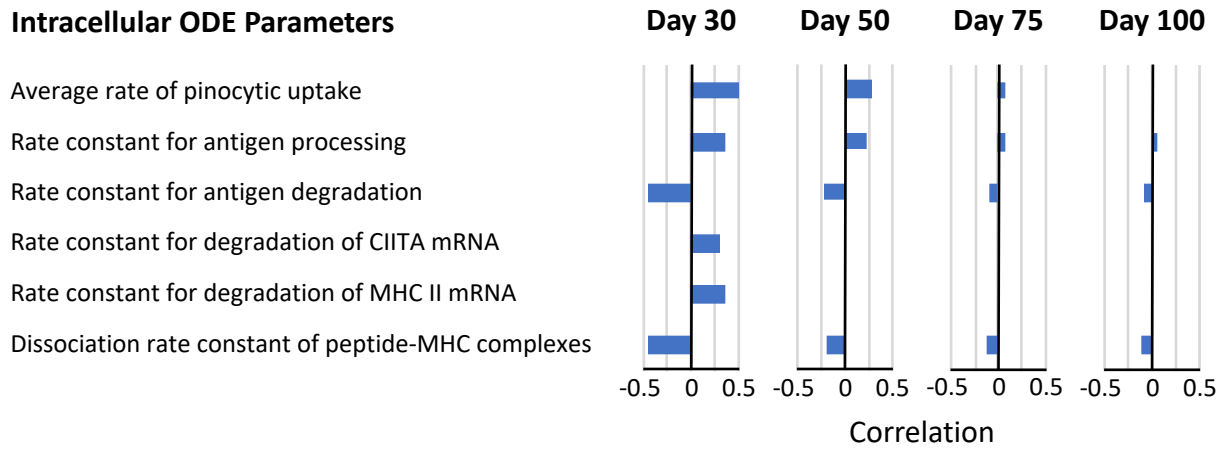

Figure A.4 Results of a cellular level intercompartment sensitivity analysis. Here we fixed *GranSim* parameters and varied parameters within the intracellular scale model with a readout in the *GranSim* tissue scale model. Partial rank correlation coefficients (see Methods) for four time points are shown for six model parameters inputs with their effects on numbers of Mtb-peptide-MHCII complexes on the surface of macrophages as the output variable.

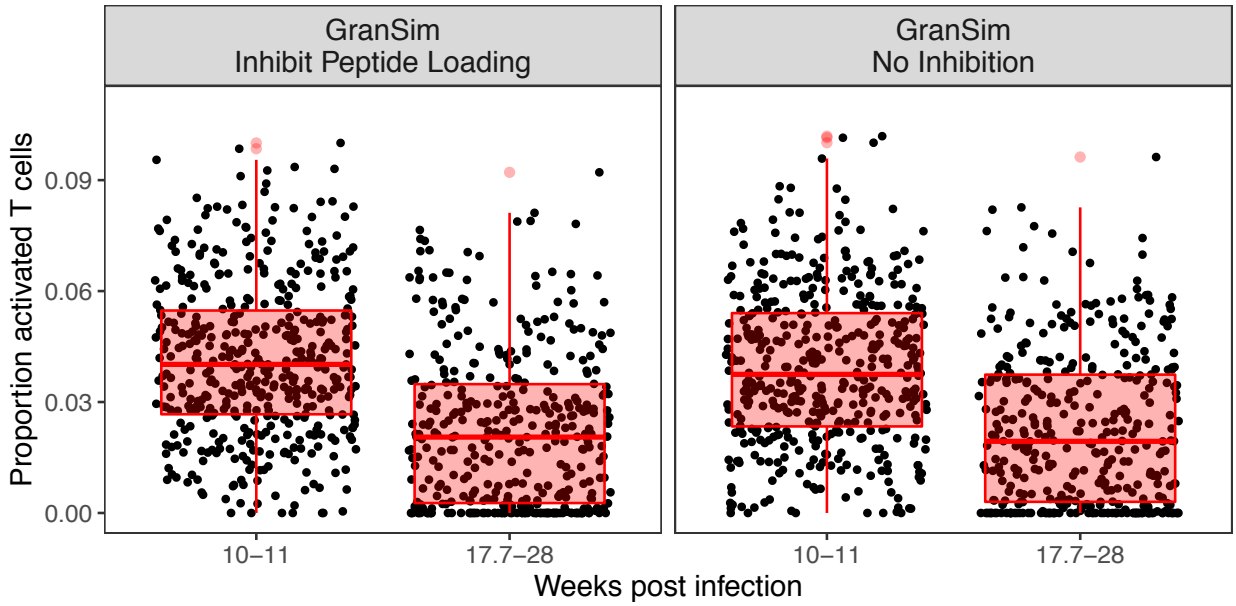

Figure A.5 Proportion of activated T cells in simulated granulomas in *GranSim* run with MHC II dynamics. These compare inhibition of peptide loading through MHCII downregulation and the negative control of no inhibition over the 28 weeks course of infection.

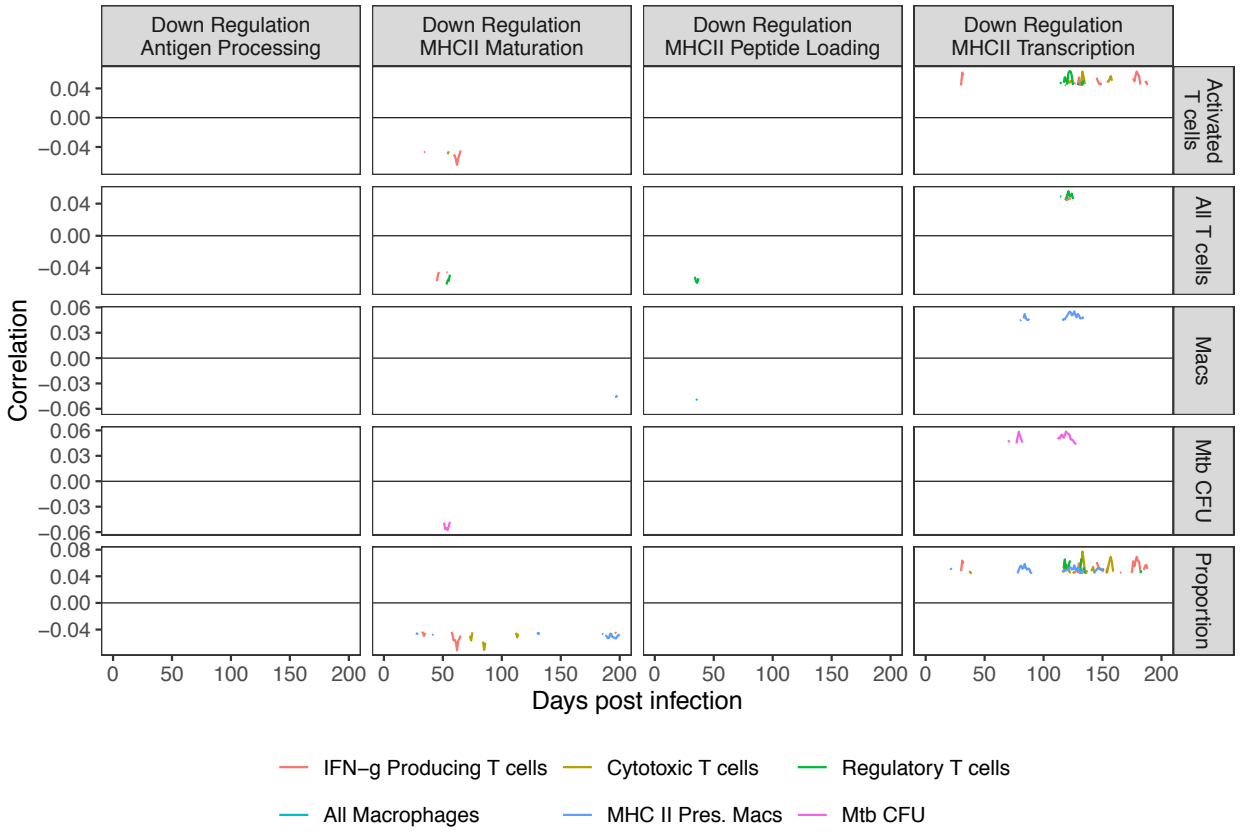

Figure A.6 Results of a tissue scale intracompartment sensitivity analysis. Here we fixed *GranSim* and MHCII sub model parameters and varied downregulation linking parameters with a readout in the *GranSim* tissue scale model. Partial rank correlation coefficients (see Methods) for entire virtual infection are shown for the four downregulation process inputs with their effects on numbers of T cells, macrophages, and Mtb as the output variables. Only significant correlations shown ($p<0.05$).

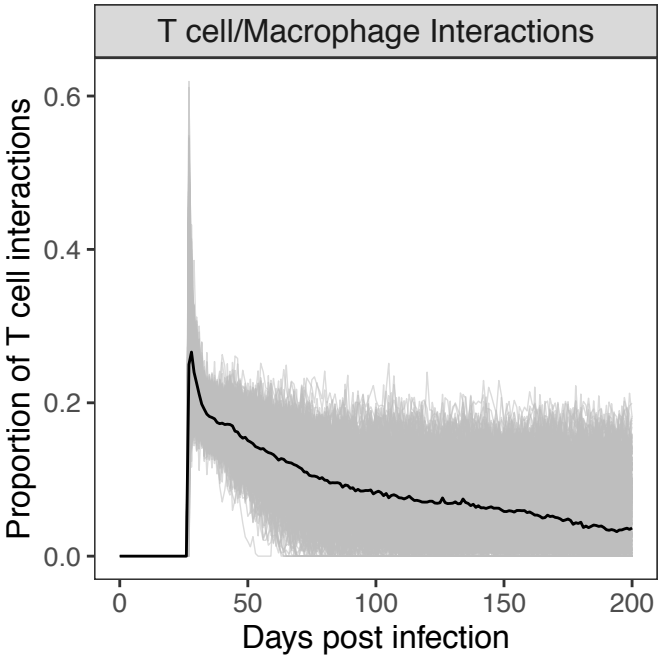

Figure A.7 Proportion of T cells that interacted with at least one macrophage. Median is shown in black, with simulation runs shown in gray. *GranSim* run with MHC II dynamics, no inhibition of antigen presentation, and all other *GranSim* parameters set to default.

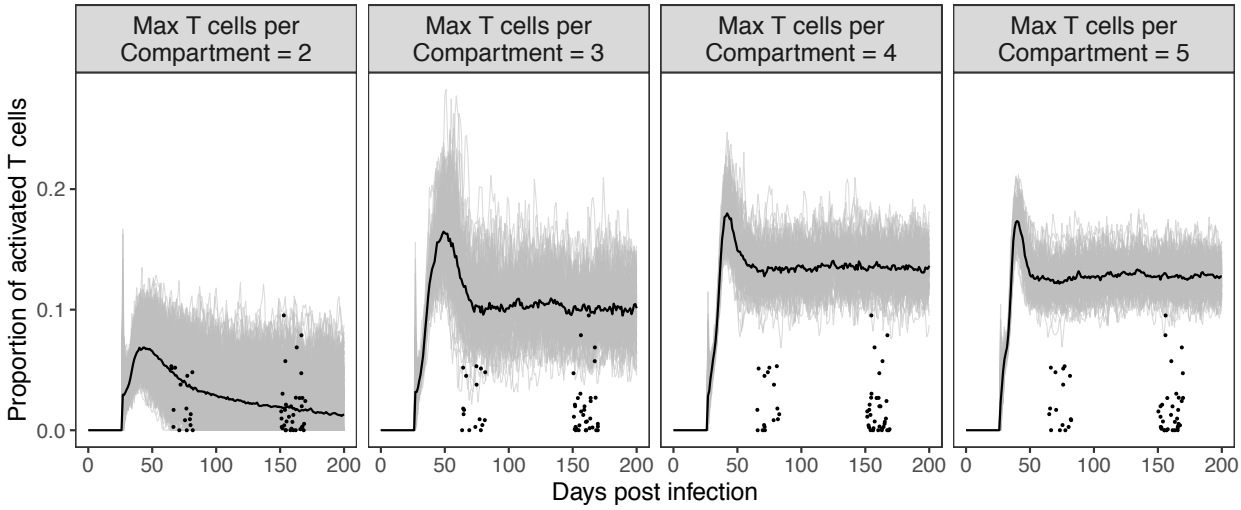

Figure A.8 Comparison of proportion of activated T cells in *GranSim* simulations with MHCII presentation ODEs, varying the maximum number of T cells that can fit within one grid compartment. Median is shown in black, with simulation runs shown in gray. NHP data is overlaid for each output (Gideon et al., 2015).[39]

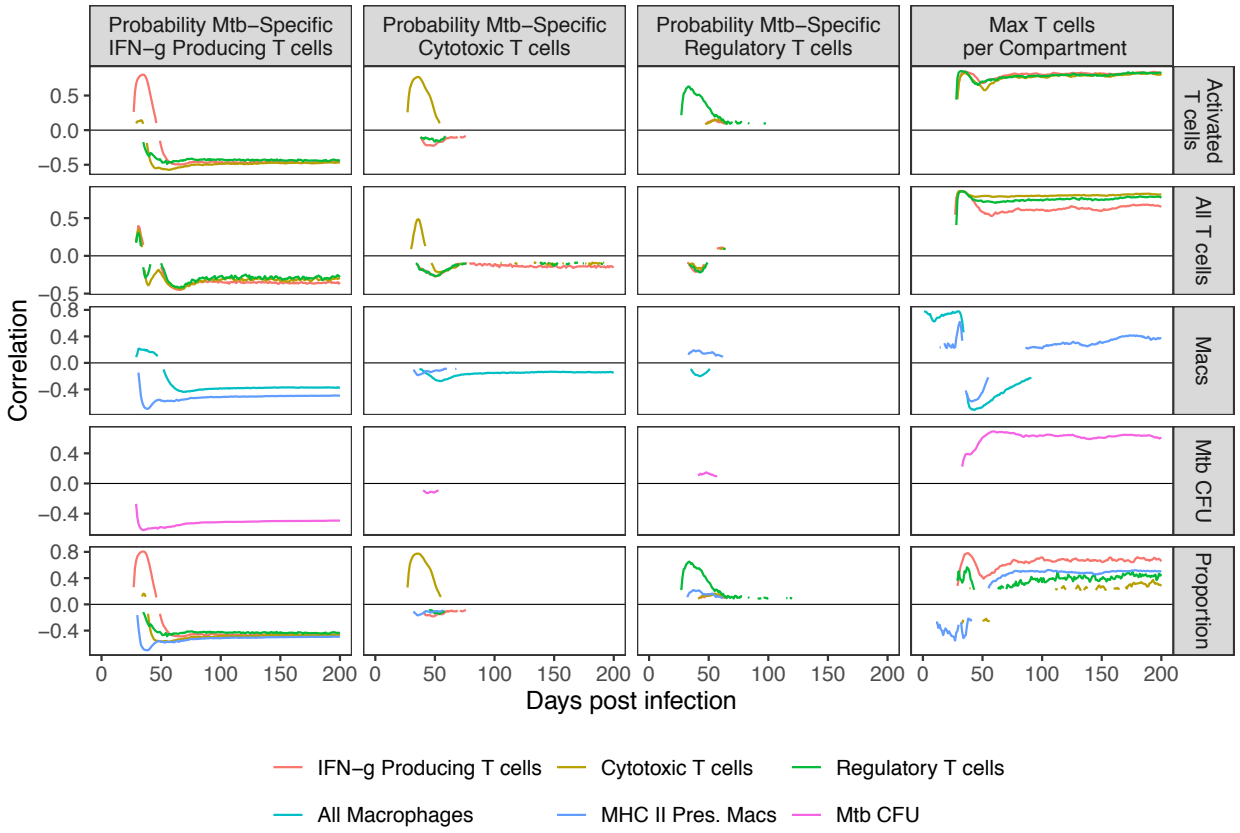

Figure A.9 Results of an intracompartment sensitivity analysis. Here we fixed the MHCII sub model parameters and varied GranSim parameters for probability for T cell Mtb-specificity and separately ran and varied T cell density, with a readout in the *GranSim* tissue scale model. Partial rank correlation coefficients (see Methods) for entire virtual infection are shown for the four parameter inputs with their effects on numbers of T cells, macrophages, and Mtb as the output variables. Only significant correlations shown ($p<0.05$).

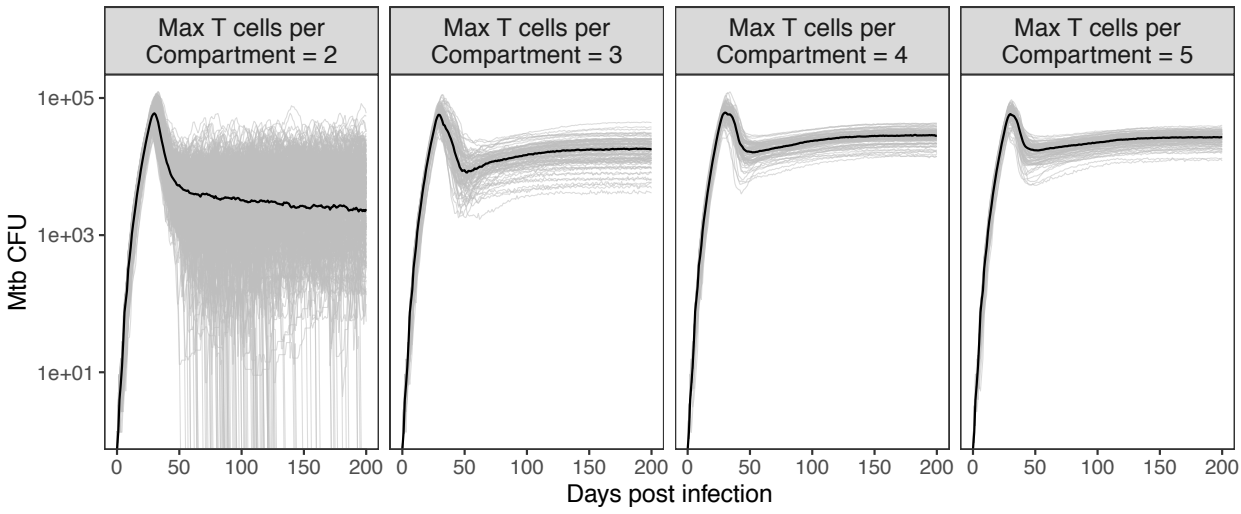

Figure A.10 Comparison of Mtb clearance in *GranSim* simulations with MHCII presentation ODEs, varying the maximum number of T cells that can fit within one grid compartment. Median is shown in black, with simulation runs shown in gray. NHP data is overlaid for each output.[39]

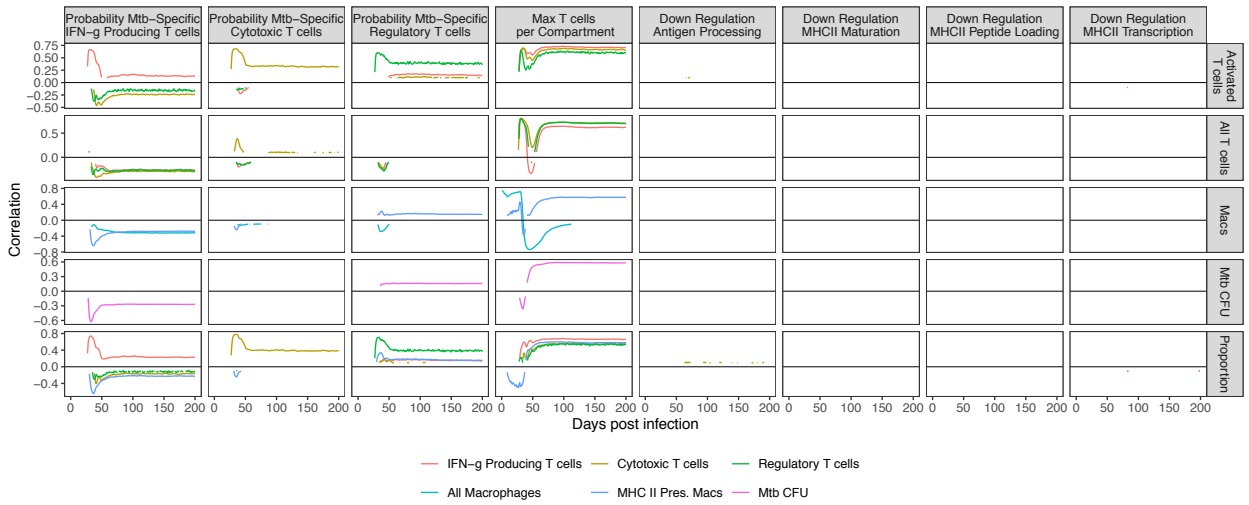

Figure A.11 Results of an intercompartment sensitivity analysis. Here we varied both simultaneously the *GranSim* parameters and the downregulation linking parameters with a readout in the *GranSim* tissue scale model. Partial rank correlation coefficients (see Methods) for entire virtual infection are shown for the eight parameter inputs with their effects on numbers of T cells, macrophages, and Mtb as the output variables. Only significant correlations shown ($p<0.05$).

Table A.1 Differential equations describing molecular scale MHCII dynamics

Equation description	Equation
IFN-γ Receptor Ligand Binding	
Change in number of free IFN-γ receptors	$\frac{dR}{dt} = -k_{\text{on-IFN-}\gamma} G \cdot R + k_{\text{off-IFN-}\gamma} C + k_{\text{recyc}} C$
Change in number of IFN-γ receptor ligand complexes	$\frac{dC}{dt} = k_{\text{on-IFN-}\gamma} G \cdot R - k_{\text{off-IFN-}\gamma} C - k_{\text{recyc}} C$
MHC Class II Transcription	
Change in levels of CIITA mRNA	$\frac{dT_1}{dt} = k_{\text{txn1}} \left(1 + \alpha \frac{C}{R_{\text{tot}}}\right) - k_{\text{deg-mRNA1}} T_1$
Change in levels of CIITA protein	$\frac{dP}{dt} = k_{\text{tsl1}} T_1 - k_{\text{deg-P}} P$
Change in levels of MHCII mRNA	$\frac{dT_2}{dt} = k_{\text{txn2}} P - k_{\text{deg-mRNA2}} T_2$
Mtb Antigens/Peptides and Self-Peptides	
Change in molar concentration of Mtb antigens in endosomal compartments	$\frac{dA}{dt} = \left(k_{\text{pino}} \frac{1}{v_{\text{MIIC}}}\right) A^* - k_{\text{deg-A}} A - k_{\text{lys}} A$
Change in molar concentration of Mtb antigen-derived peptides in endosomal compartments	$\frac{dE}{dt} = k_{\text{deg-A}} A + (k_{\text{off-MHC}} M_e - k_{\text{on-MHC}} M \cdot E) \cdot \frac{1}{N_A v_{\text{MIIC}}} - k_{\text{lys}} E$
Change in molar concentration of self-peptides in endosomal compartments	$\frac{dS}{dt} = k_{\text{source}} + \left[k_{\text{deg-MHC}} (M_s + M_s^*) - k_{\text{on-MHC}} M \cdot S + k_{\text{off-MHC}} M_s\right] \left(\frac{1}{N_A v_{\text{MIIC}}}\right) - k_{\text{lys}} S$
MHCII Translation & Peptide-MHCII Binding	
Change in number of free MHCII proteins in endosomal compartments	$\frac{dM}{dt} = k_{\text{ts12}} + \left(1 + \beta \frac{C}{R_{\text{tot}}}\right) T - k_{\text{on-MHC}} M \cdot S + k_{\text{off-MHC}} M_s - k_{\text{on-MHC}} M \cdot E + k_{\text{off-MHC}} M_e - k_{\text{out}} M + k_{\text{in}} M^* - k_{\text{deg-MHC}} M$
Change in number of free MHCII proteins on cell surface	$\frac{dM^*}{dt} = k_{\text{out}} M - k_{\text{in}} M^* - k_{\text{deg-MHC}} M$
Change in number of MHCII self-peptide complexes in endosomal compartments	$\frac{dM_s}{dt} = k_{\text{on-MHC}} M \cdot S - k_{\text{off-MHC}} M_s - k_{\text{out}} M_s + k_{\text{in}} M_s^* - k_{\text{deg-MHC}} M_s$
Change in number of surface MHCII self-peptide complexes	$\frac{dM_s^*}{dt} = k_{\text{out}} M_s - k_{\text{in}} M_s^* - k_{\text{deg-MHC}} M_s^*$
Change in number of MHCII Mtb-peptide complexes in endosomal compartments	$\frac{dM_e}{dt} = k_{\text{on-MHC}} M \cdot P - k_{\text{off-MHC}} M_e - k_{\text{out}} M_e + k_{\text{in}} M_e^* - k_{\text{deg-MHC}} M_e$
Change in number of surface MHCII Mtb-peptide complexes	$\frac{dM_e^*}{dt} = k_{\text{out}} M_e - k_{\text{in}} M_e^* - k_{\text{deg-MHC}} M_e^*$

Table A.2 Molecular scale MHCII dynamics parameters. (Estimated from Chang et al., 2005[64])

Parameter	Parameter description	Value (LHS Ranges)	
$k_{\text{on-IFN-}\gamma}$	Association rate constant of IFN-γ receptor-ligand complex	$1.2 \times 10^{7}/\text{s}$,	$(5.0 \times 10^{6}, 1.3 \times 10^{7})$
$k_{\text{off-IFN-}\gamma}$	Dissociation rate constant of IFN-γ receptor-ligand complex	$9.8 \times 10^{-3}/\text{s}$,	$(2.0 \times 10^{-3}, 1.0 \times 10^{-2})$
k_{recyc}	Rate constant for receptor internalization and recycling	$2.5 \times 10^{-4}/\text{s}$,	$(1.4 \times 10^{-5}, 6.0 \times 10^{-4})$
α	Scaling factor for CIITA transcription	4.5×10^{1},	$(1.0 \times 10^{1}, 5.0 \times 10^{1})$
$k_{\text{deg-mRNA1}}$	Rate constant for degradation of CIITA mRNA	$2.1 \times 10^{-5}/\text{s}$,	$(7.0 \times 10^{-7}, 7.0 \times 10^{-5})$
$k_{\text{deg-P}}$	Rate constant for degradation of CIITA protein	$5.5 \times 10^{-4}/\text{s}$,	$(2.0 \times 10^{-5}, 8.0 \times 10^{-4})$
$k_{\text{deg-mRNA2}}$	Rate constant for degradation of MHCII mRNA	$1.7 \times 10^{-5}/\text{s}$,	$(2.0 \times 10^{-6}, 1.8 \times 10^{-5})$
k_{pino}	Average rate of pinocytic uptake per cell	$6.1 \times 10^{-17}\text{L/s}$,	$(1.4 \times 10^{-17}, 6.0 \times 10^{-16})$
$k_{\text{deg-A}}$	Rate constant for antigen processing	$1.6 \times 10^{-3}/\text{s}$,	$(6.4 \times 10^{-5}, 2.4 \times 10^{-3})$
k_{lys}	Rate constant for antigen degradation	$1.3 \times 10^{-3}/\text{s}$,	$(3.0 \times 10^{-5}, 3.0 \times 10^{-3})$
$k_{\text{on-MHC}}$	Association rate constant of peptide-MHC complexes	$1.1 \times 10^{2}/\text{M} \cdot \text{s}$,	$(5.6 \times 10^{0}, 5.6 \times 10^{3})$
$k_{\text{off-MHC}}$	Dissociation rate constant of peptide-MHC complexes	$1.3 \times 10^{-2}/\text{s}$,	$(8.0 \times 10^{-4}, 4.0 \times 10^{-2})$
$k_{\text{deg-MHC}}$	Rate constant for degradation of peptide-MHC complexes	$9.3 \times 10^{-6}/\text{s}$,	$(5.0 \times 10^{-6}, 1.3 \times 10^{-5})$
β	Scaling factor for MHCII translation	4.2×10^{0},	$(4.0 \times 10^{0}, 8.0 \times 10^{0})$
k_{out}	Rate constant of MHCII protein transport from endosomes to the plasma membrane	$4.0 \times 10^{-4}/\text{s}$,	$(6.0 \times 10^{-5}, 3.0 \times 10^{-3})$
k_{source}	Rate of self-peptide synthesis	$k_{\text{lys}} \cdot S_0$	
k_{txn1}	Rate constant for CIITA transcription	$k_{\text{deg-mRNA1}} \cdot T_{1,0}$	
k_{tsl1}	Rate constant for CIITA translation	$k_{\text{deg-P}} \cdot \frac{P_0}{T_{1,0}}$	
k_{txn2}	Rate constant for MHCII transcription	$k_{\text{deg-mRNA2}} \cdot \frac{T_{2,0}}{P_0}$	
k_{ts12}	Rate constant representing MHCII translation	$k_{\text{deg-MHC}} \cdot \left(\frac{M_0+M_0^*+M_{s,0}+M_{s,0}^*}{T_{2,0}}\right)$	
k_{in}	Rate constant of MHCII protein internalization from the plasma membrane	$\frac{p_{\text{in}}}{1-p_{\text{in}}} k_{\text{out}} - k_{\text{deg-MHC}}$	
v_{MIIC}	Total volume of the MHCII-accessible endosomal compartments	$4.0 \times 10^{-16}\text{L}$	
R_{tot}	IFN-γ receptors per cell	1.0×10^{4}	
p_{in}	Fraction of all MHCII that are intracellular	3.3×10^{-1}	

p_{bound}	Fraction of all MHCII that are bound to self-peptide	8.0×10^{-1}
M_{tot}	Total number of MHCII proteins on and in a macrophage	2.0×10^{5}
N_A	Avogadro's number	6.02×10^{23}

Table A.3 Molecular scale MHCII dynamics variable initial conditions. (Estimated from Chang et al., 2005[64])

Variable	Variable description	Value
R_0	Number of free IFN- γ receptors	R_{tot}
C_0	Number of IFN- γ receptor ligand complexes	0
$T_{1,0}$	Levels of CIITA mRNA	1
P_0	Levels of CIITA protein	1
$T_{2,0}$	Number of MHCII mRNA	1
A_0	Molar conc. of Mtb antigens in endosomal compartments	0
E_0	Molar conc. of Mtb peptides in endosomal compartments	0
S_0	Molar conc. of self-peptides in endosomal compartments	$\frac{k_{\mathrm{deg-MHC}}\cdot\left(M_{s,0}+M_{s,0}^{*}\right)+k_{\mathrm{off-MHC}}\cdot M_{s,0}}{k_{\mathrm{on-MHC}}\cdot M_0}$
M_0	Number of MHCII proteins in endosomal compartments	$p_{\mathrm{in}}\cdot(1-p_{\mathrm{bound}})\cdot M_{\mathrm{tot}}$
M_0^*	Number of MHCII proteins on cell surface	$\frac{1-p_{\mathrm{in}}}{p_{\mathrm{in}}}\cdot M_0$
$M_{s,0}$	Number of MHCII self-peptide complexes in endosomal compartments	$\frac{p_{\mathrm{bound}}}{1-p_{\mathrm{bound}}}\cdot M_0$
$M_{s,0}^*$	Number of MHCII self-peptide complexes on cell surface	$\frac{1-p_{\mathrm{in}}}{p_{\mathrm{in}}}\cdot M_{s,0}$
$M_{e,0}$	Number of MHCII Mtb-peptide complexes in endosomal compartments	0
$M_{e,0}^*$	Number of MHCII Mtb-peptide complexes on cell surface	0

Table A.4 Molecular scale MHCII dynamics scaling factors used to represent Mtb down regulation of MHCII processes

Parameter	Parameter description	Value
$d_{\mathrm{deg-A}}$	Maximum rate of Mtb down regulation of antigen processing ($k_{\mathrm{deg-A}}$) achieved by the system, at saturating Mtb concentration	$0.25 - 1.00$
d_{ts12}	Maximum rate of Mtb down regulation of MHCII maturation (k_{ts12}) achieved by the system, at saturating Mtb concentration	$0.25 - 1.00$
$d_{\mathrm{on-MHC}}$	Maximum rate of Mtb down regulation of MHCII peptide loading ($k_{\mathrm{on-MHC}}$) achieved by the system, at saturating bacterial concentration	$0.25 - 1.00$
d_{txn1}	Maximum rate of Mtb down regulation of MHCII transcription (k_{txn1}) achieved by the system, at saturating Mtb concentration	$0.25 - 1.00$
d_{half}	Number of intracellular Mtb at which the reaction rate is half of Mtb down regulation	5

Table A.5 Mtb down regulation of MHCII processes equations

Equation description	Equation
Mtb down regulation of antigen processing	$\text{mtbDownReg}_{\text{deg-A}} = 1 - \frac{d_{\text{deg-A}} \cdot (\#\ \text{Mtb in Mac})}{d_{\text{half}} + (\#\ \text{Mtb in Mac})}$
Mtb down regulation of MHCII maturation	$\text{mtbDownReg}_{\text{ts12}} = 1 - \frac{d_{\text{ts12}} \cdot (\#\ \text{Mtb in Mac})}{d_{\text{half}} + (\#\ \text{Mtb in Mac})}$
Mtb down regulation of MHCII peptide loading	$\text{mtbDownReg}_{\text{on-MHC}} = 1 - \frac{d_{\text{on-MHC}} \cdot (\#\ \text{Mtb in Mac})}{d_{\text{half}} + (\#\ \text{Mtb in Mac})}$
Mtb down regulation of MHCII transcription	$\text{mtbDownReg}_{\text{txn1}} = 1 - \frac{d_{\text{txn1}} \cdot (\#\ \text{Mtb in Mac})}{d_{\text{half}} + (\#\ \text{Mtb in Mac})}$

Appendix B Supporting Information for Chapter 3

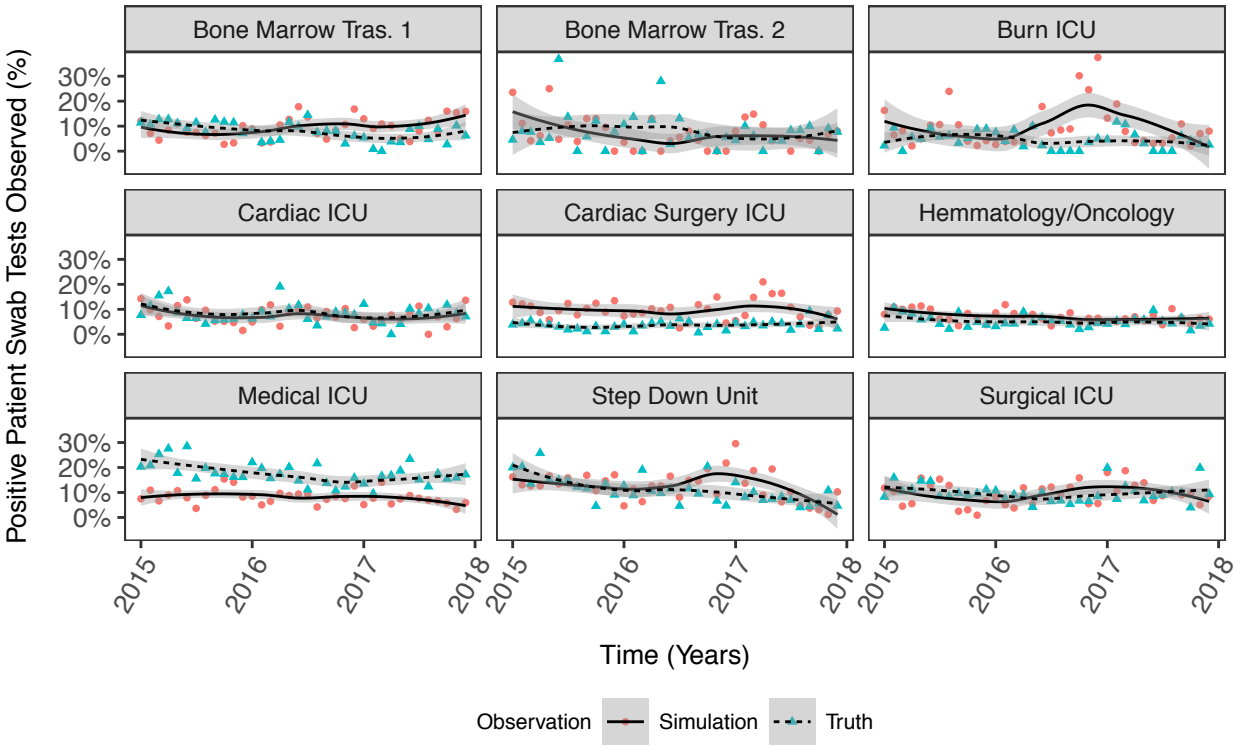

Figure B.1 Comparison of proportion of positive tests observed between actual events and simulated events in screening units. Loess smooth shown as black lines.

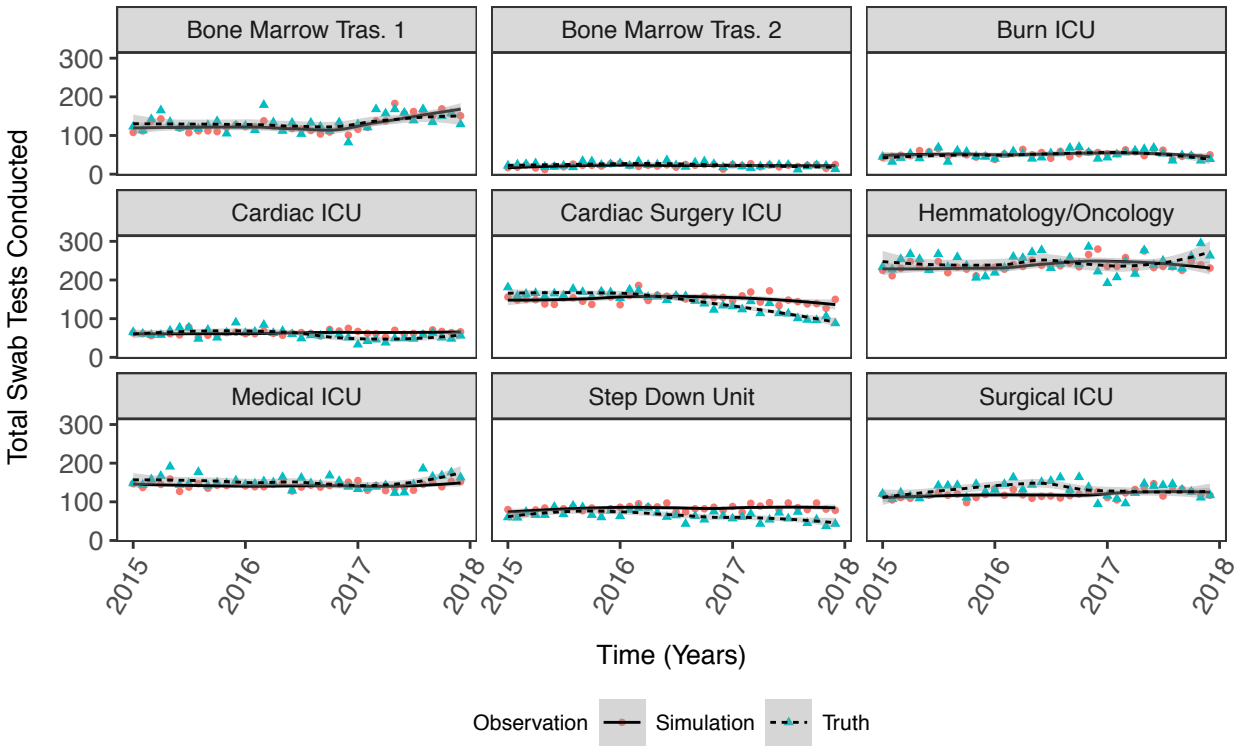

Figure B.2 Comparison of the overall number of tests conducted between actual events and simulated events in screening units. Loess smooth shown as black lines.

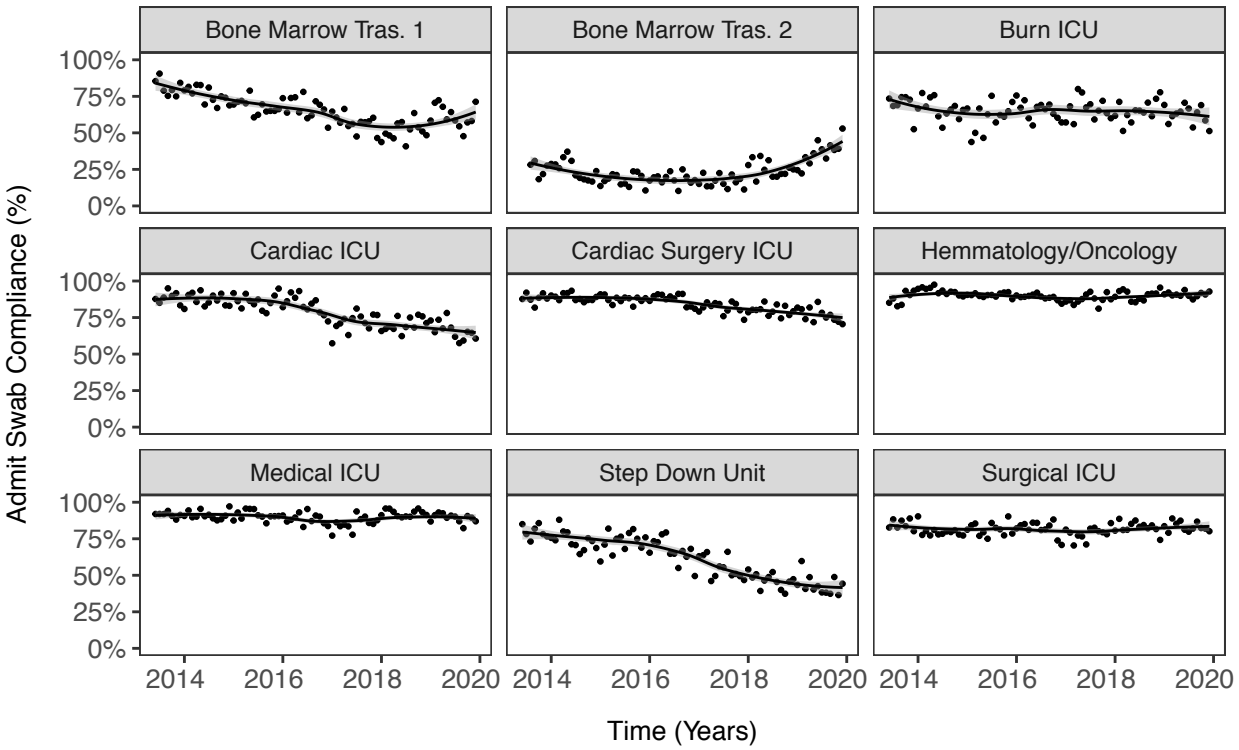

Figure B.3 Comparison of compliance rates of admission swabs performed in screening units by month. Loess smooth shown as black lines.

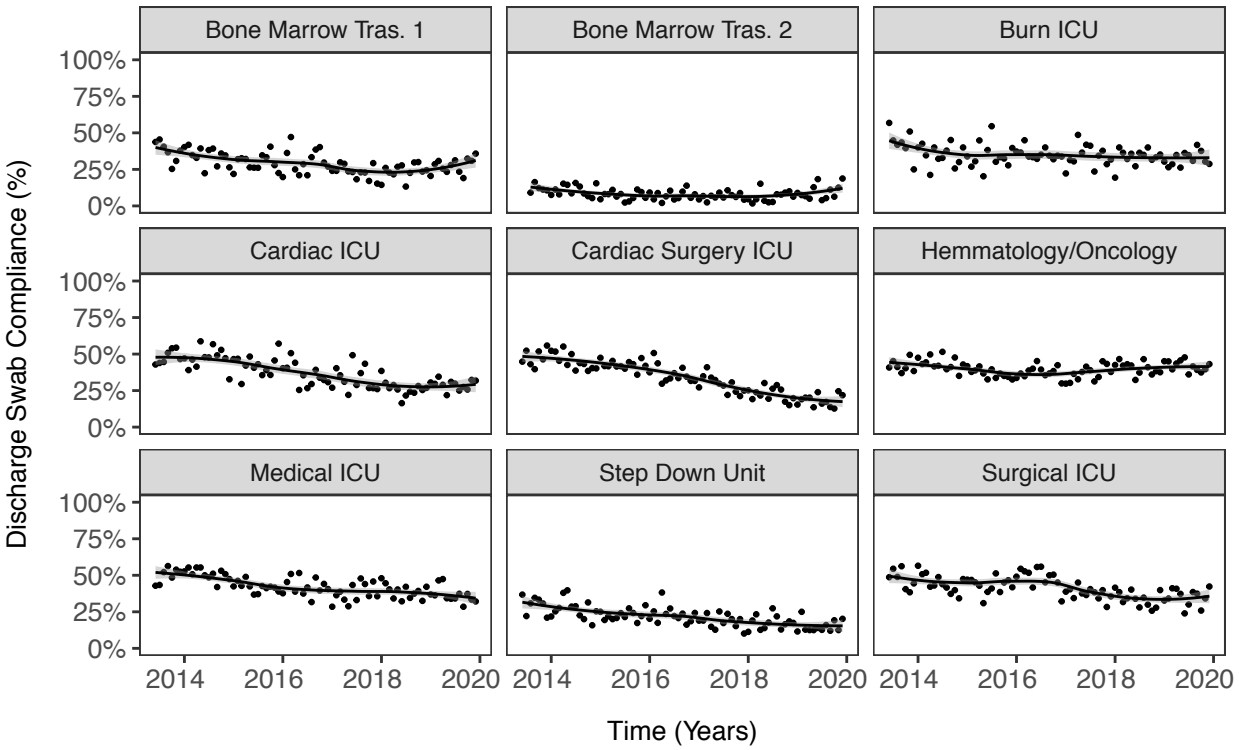

Figure B.4 Comparison of compliance rates of discharge swabs performed in screening units by month. Loess smooth shown as black lines.

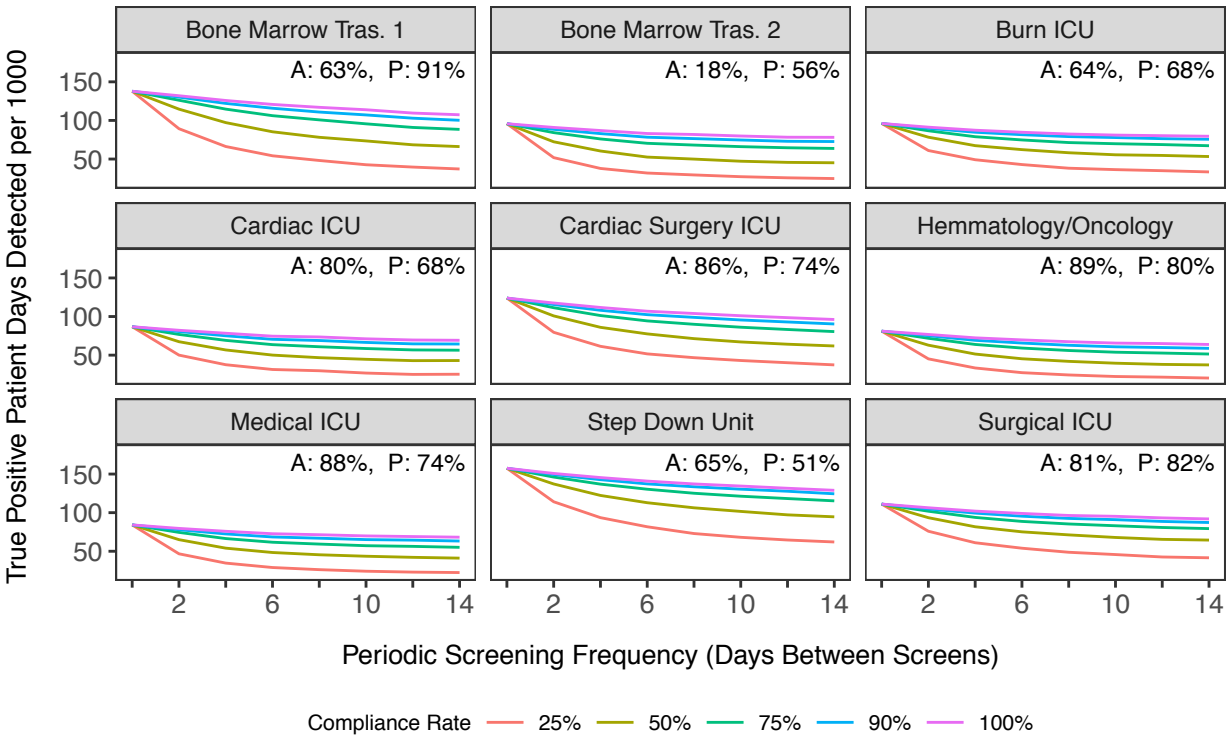

Figure B.5 Comparison of varying compliance rates and screening frequency on the number of VRE positive patient detected days per 1,000 in all screening units. A indicates the admission screening compliance for a unit. P indicates the periodic screening compliance for a unit.

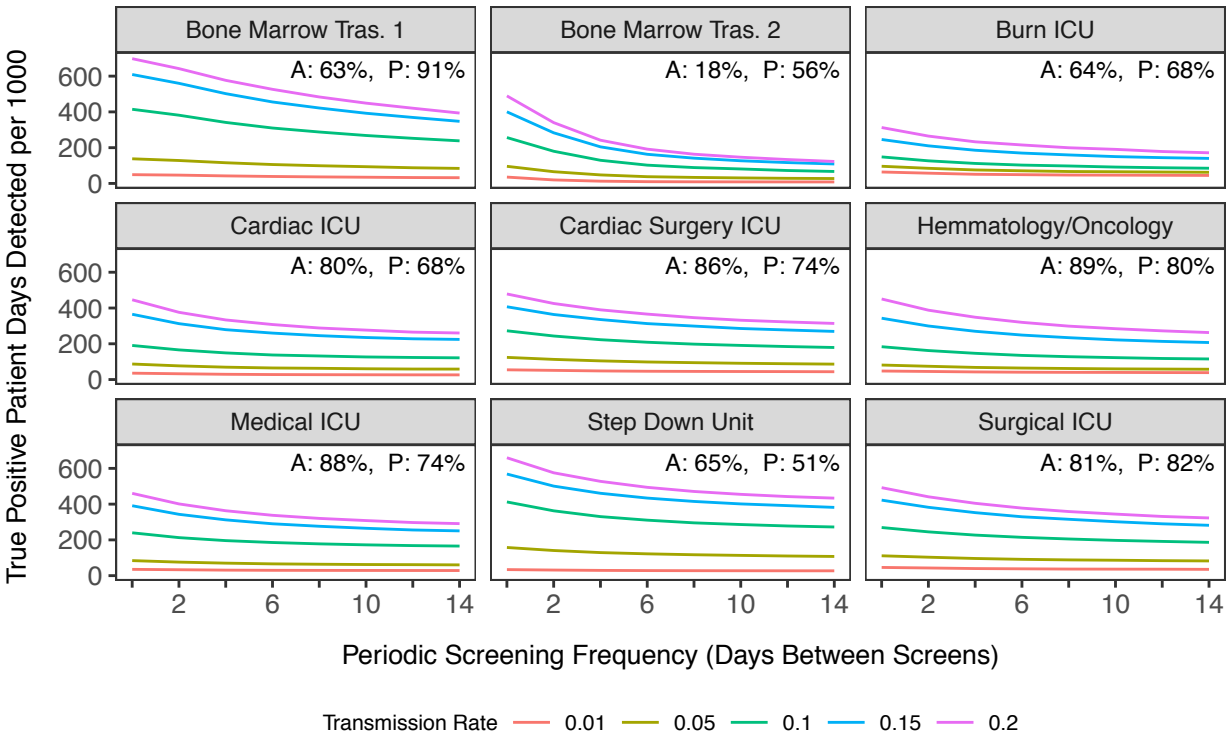

Figure B.6 Comparison of varying VRE transmission rates and screening frequency on the number of VRE positive patient detected days per 1,000 in all screening units. A indicates the admission screening compliance for a unit. P indicates the periodic screening compliance for a unit.

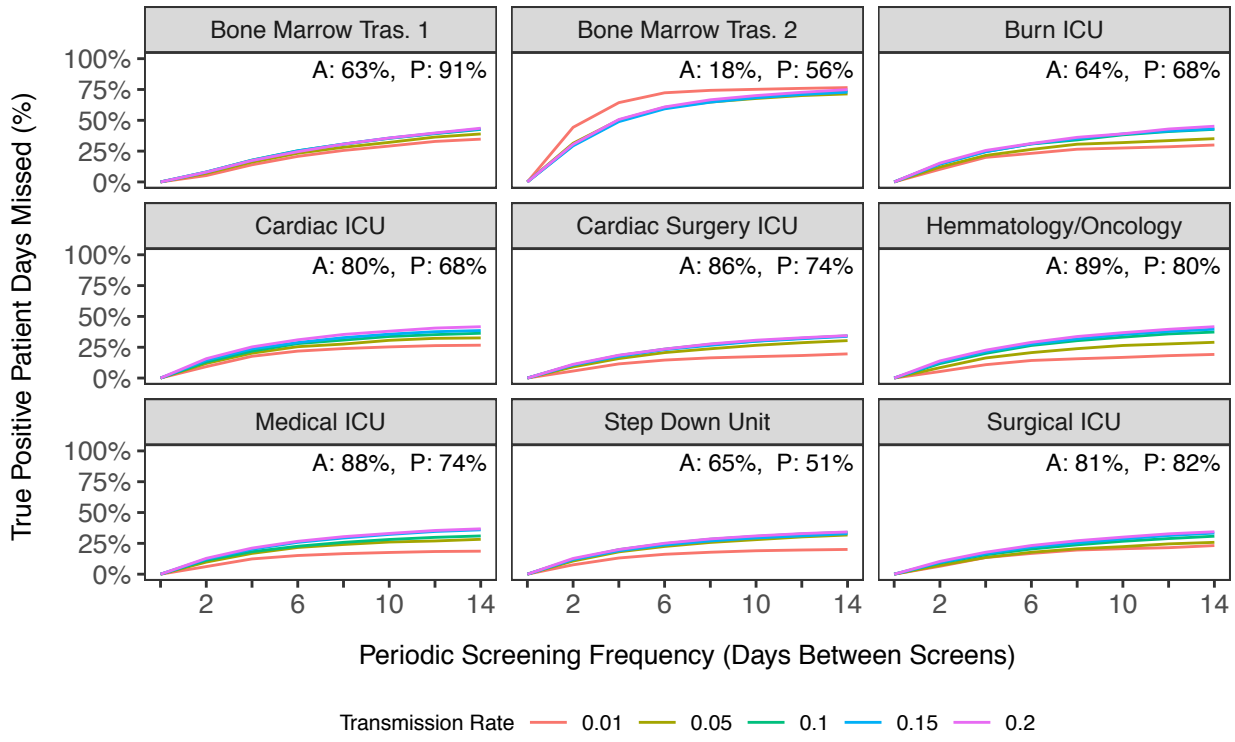

Figure B.7 Comparison of varying VRE transmission rates and screening frequency on the percent of VRE positive patient days missed in all screening units. A indicates the admission screening compliance for a unit. P indicates the periodic screening compliance for a unit.

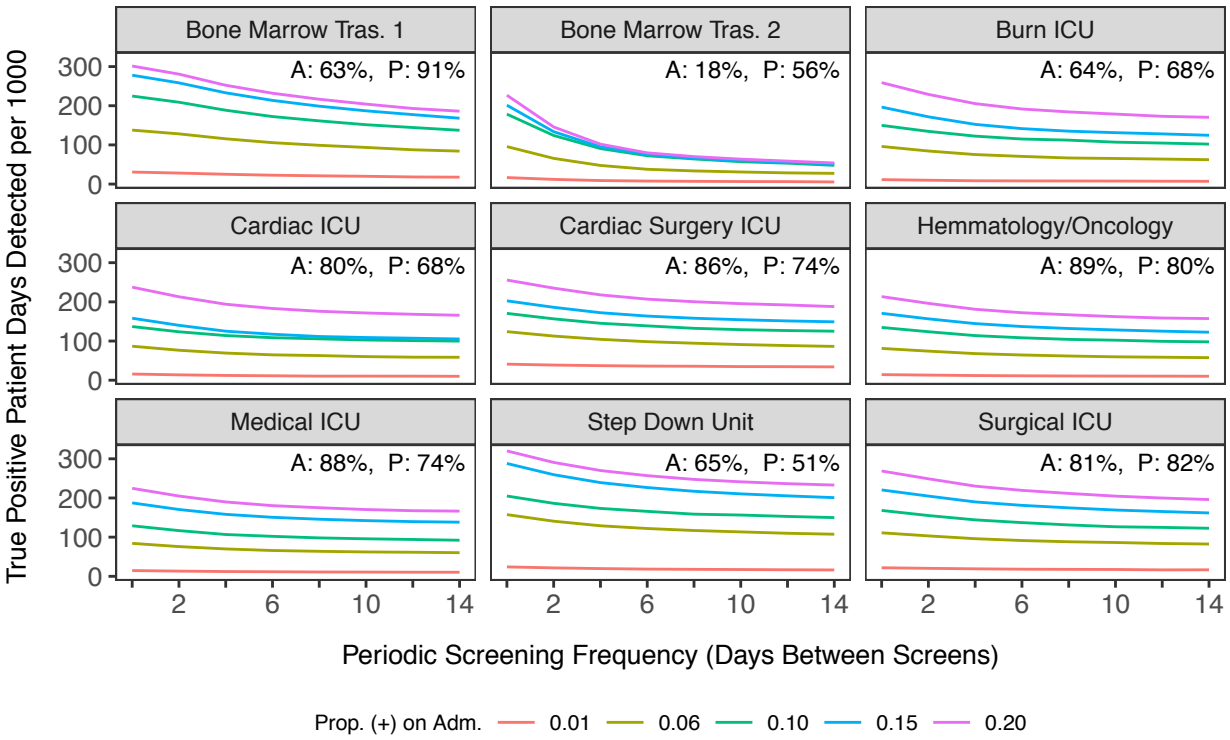

Figure B.8 Comparison of varying the proportion of patients that are positive for VRE on hospital admission and screening frequency on the number of VRE positive patient detected days per 1,000 in all screening units. A indicates the admission screening compliance for a unit. P indicates the periodic screening compliance for a unit.

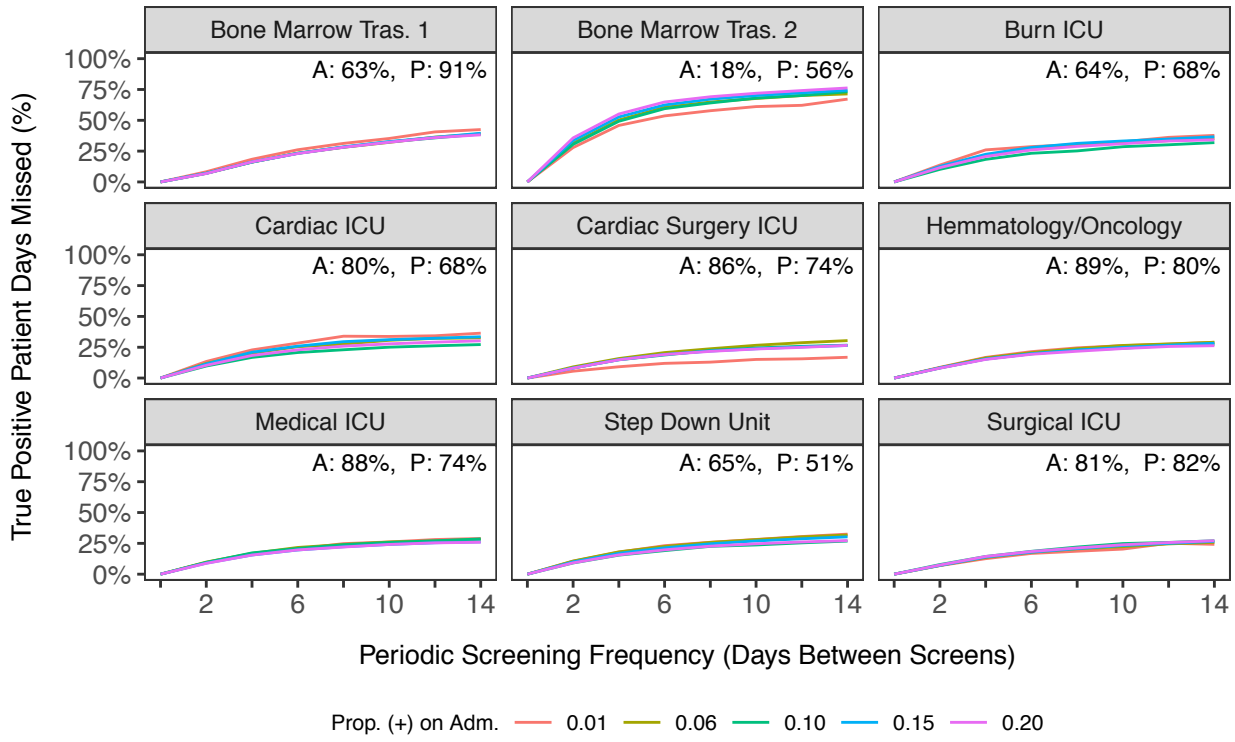

Figure B.9 Comparison of varying the proportion of patients that are positive for VRE on hospital admission and screening frequency on the percent of VRE positive patient days missed in all screening units. A indicates the admission screening compliance for a unit. P indicates the periodic screening compliance for a unit.

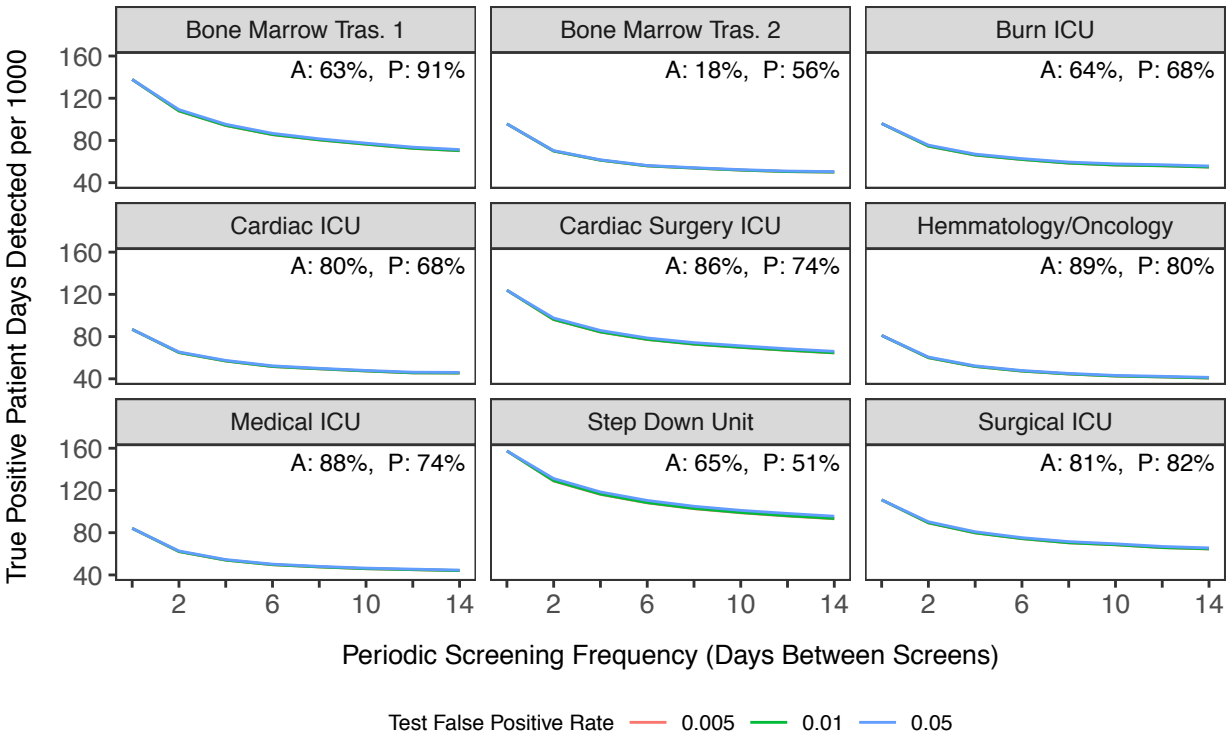

Figure B.10 Comparison of varying VRE swab false positive rates and screening frequency on the number of VRE positive patient detected days per 1,000 in all screening units. A indicates the admission screening compliance for a unit. P indicates the periodic screening compliance for a unit.

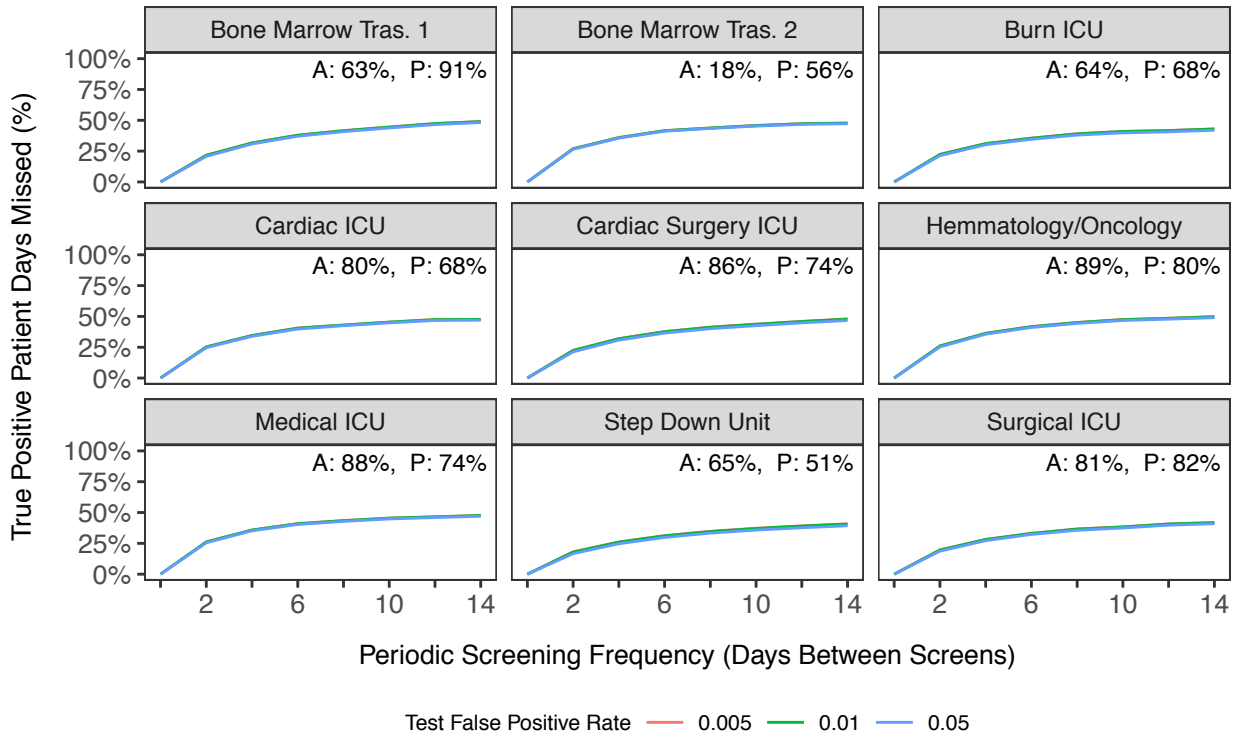

Figure B.11 Comparison of varying VRE swab false positive rates and screening frequency on the percent of VRE positive patient days missed in all screening units. A indicates the admission screening compliance for a unit. P indicates the periodic screening compliance for a unit.

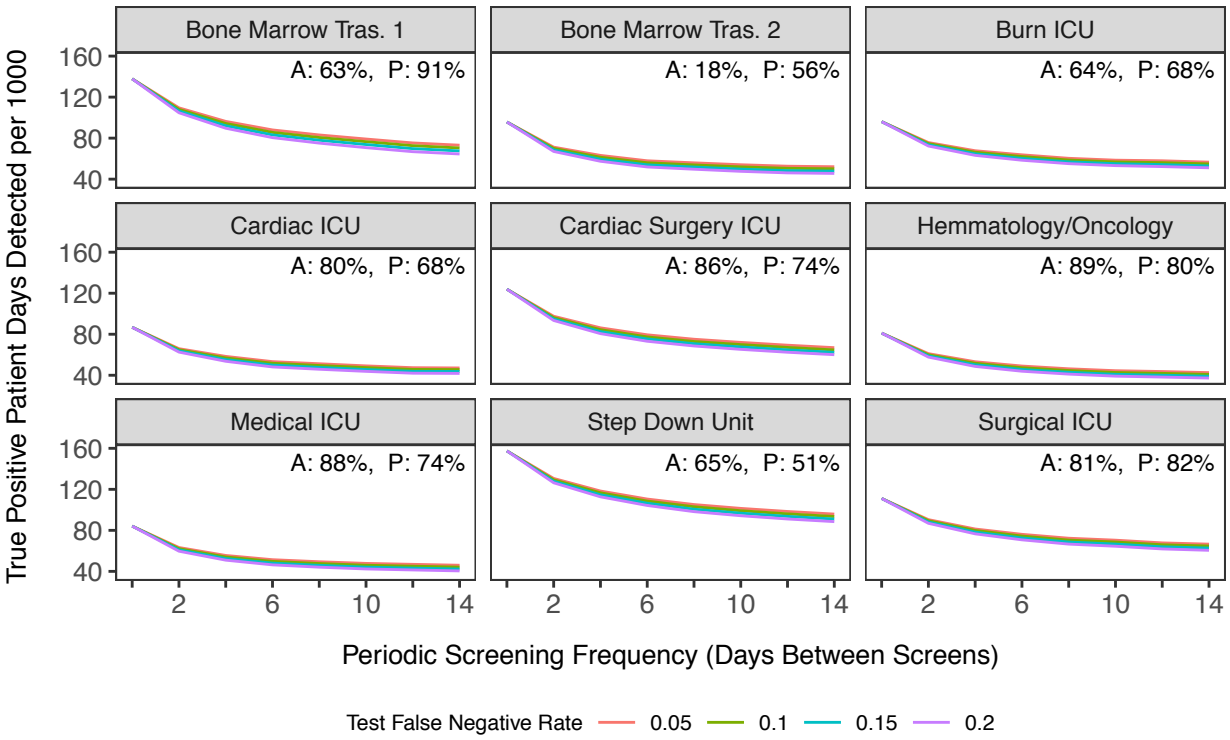

Figure B.12 Comparison of varying VRE swab false negative rates and screening frequency on the number of VRE positive patient detected days per 1,000 in all screening units. A indicates the admission screening compliance for a unit. P indicates the periodic screening compliance for a unit.

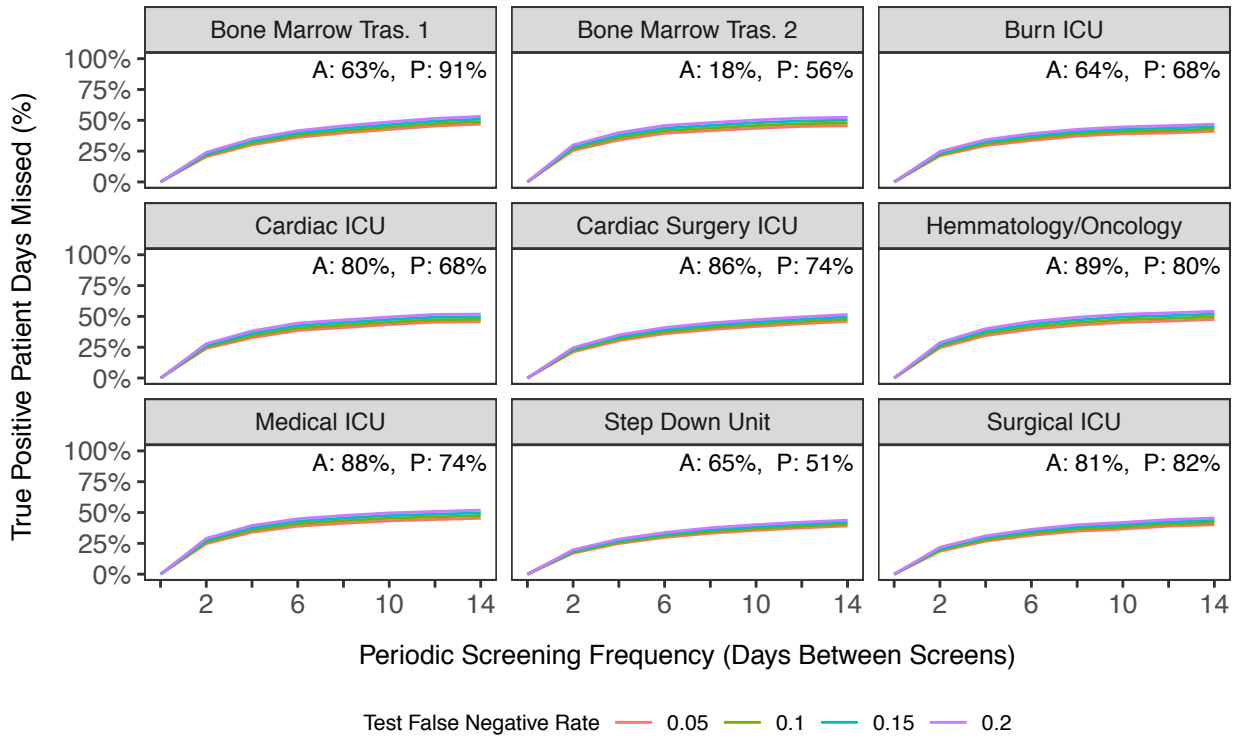

Figure B.13 Comparison of varying VRE swab false negative rates and screening frequency on the percent of VRE positive patient days missed in all screening units. A indicates the admission screening compliance for a unit. P indicates the periodic screening compliance for a unit.

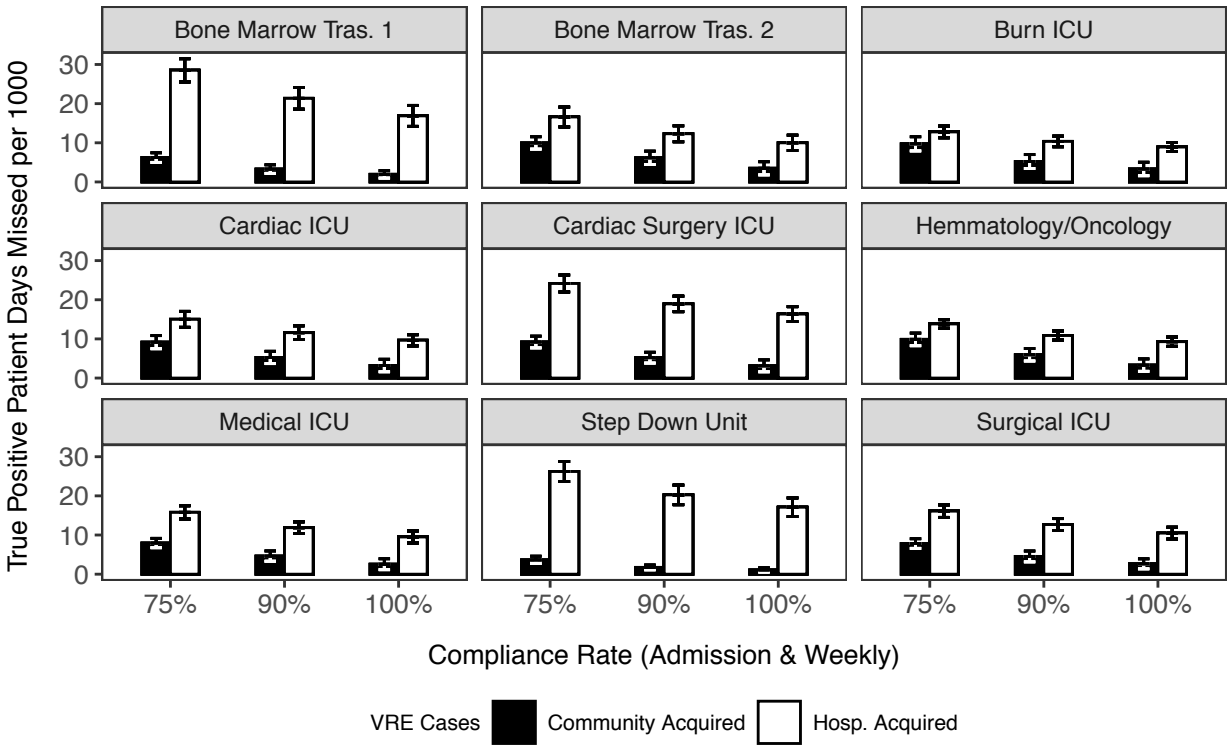

Figure B.14 Incremental reduction in missed VRE patient days per 1,000 as compliance increases in screening units. Standard deviations shown as error bars.

Table B.1 Parameters used to model VRE cases detected

Parameter description	**Baseline value (Range tested)**
Screening compliance rate	0.75, $(0.10,\ 1.00)$
Screening frequency (Biweekly period)	7, $(1,\ 14)$
Test sensitivity	0.90, $(0.80,\ 0.95)$
Test specificity	0.99, $(0.95,\ 0.995)$
Proportion of patients VRE positive on hospital admission	0.06, $(0.01,\ 0.20)$
Transmission rate	0.05, $(0.01,\ 0.20)$

Appendix C Supporting Information for Chapter 4

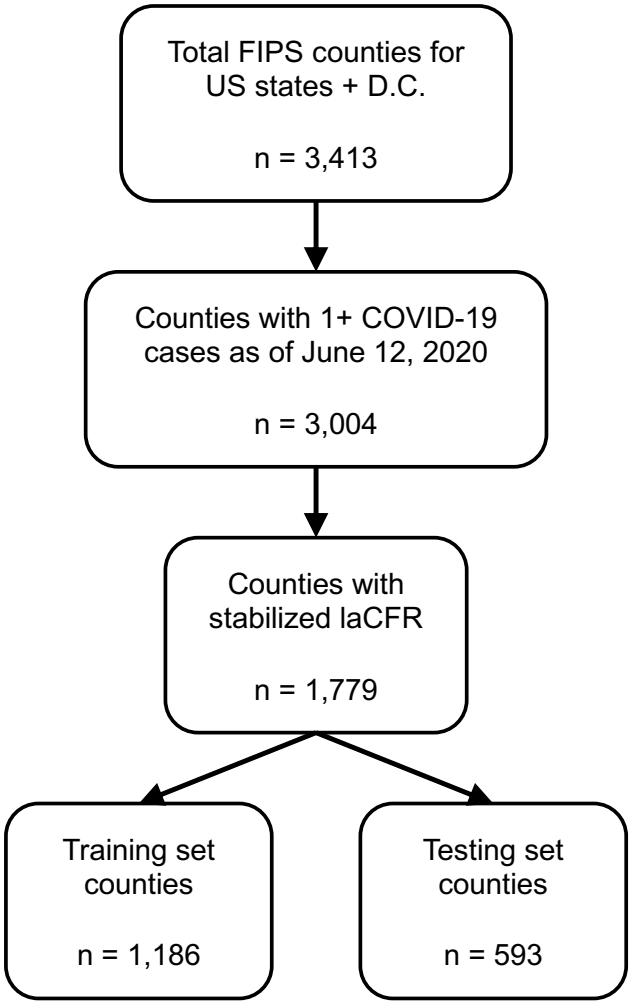

Figure C.1 COVID study population flow chart. Study population flow chart of counties included in the analysis.

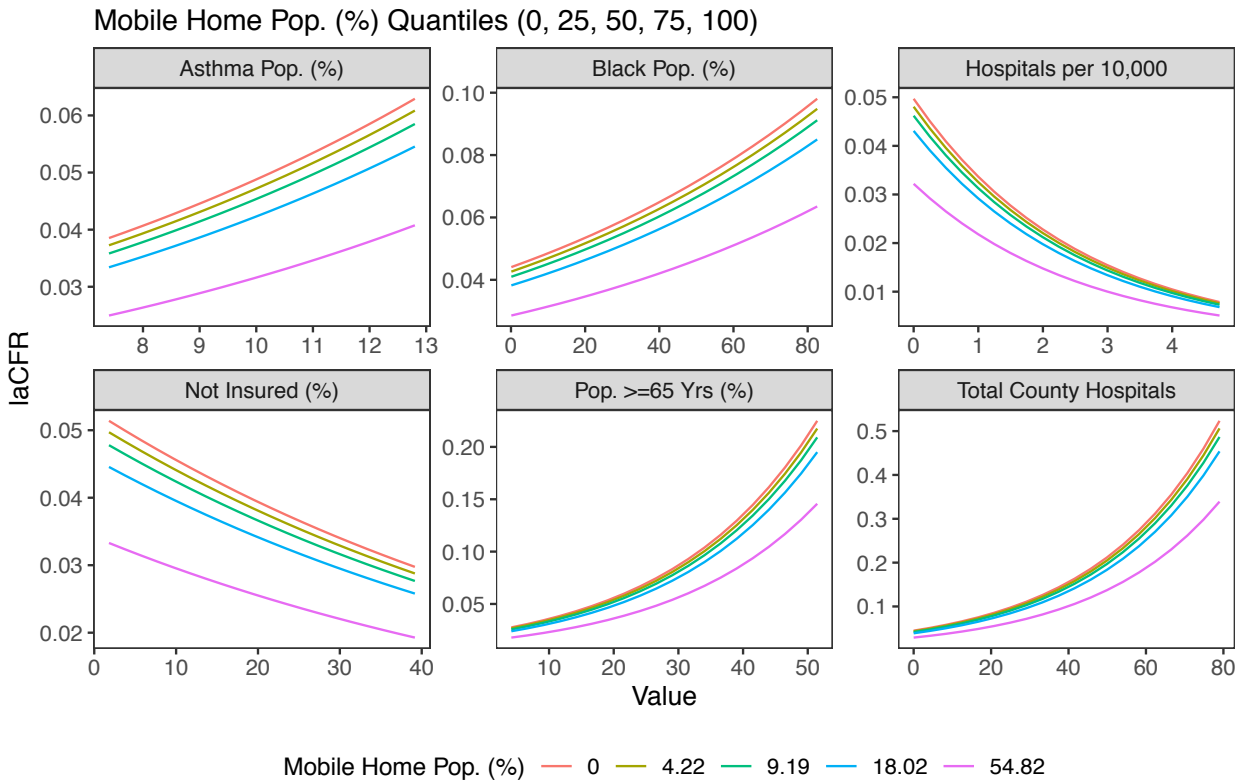

Figure C.2 Percentage of county populations living in mobile homes shown as quantiles, over the range of observed values for the second variable. All other variables set at their median (religious gatherings set to 0).

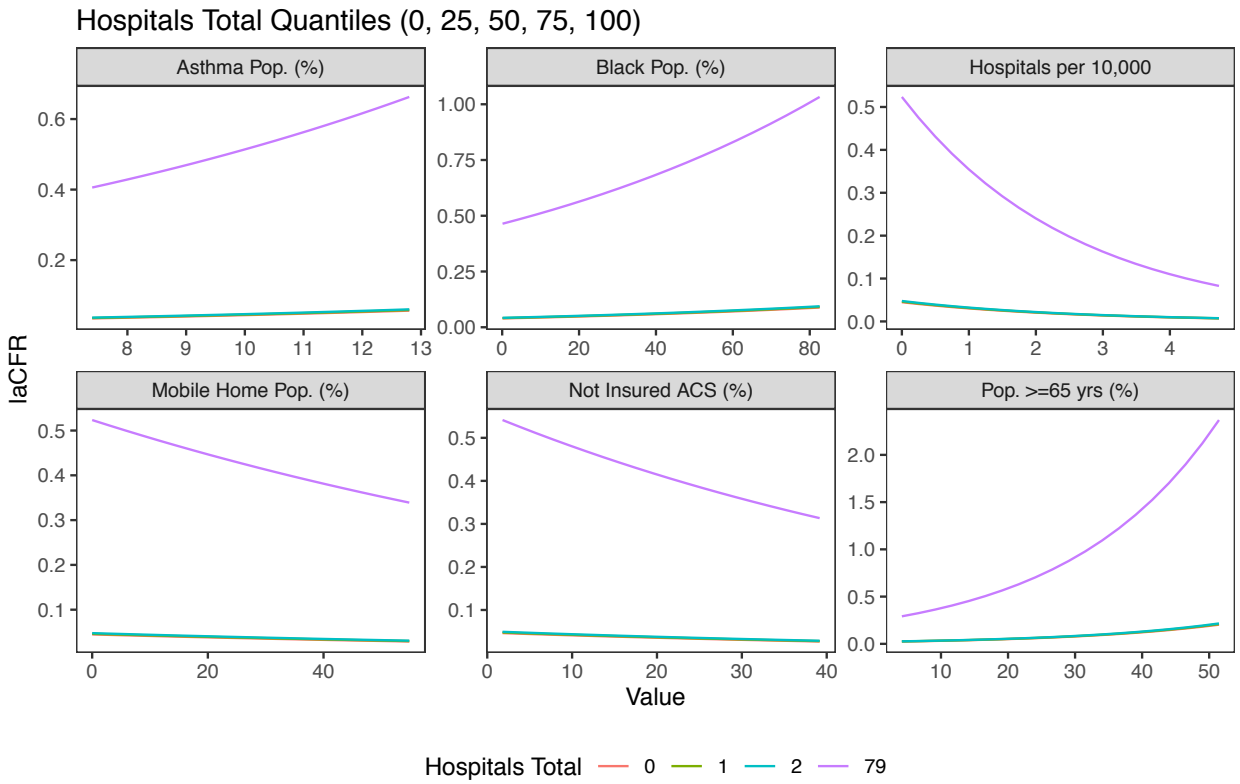

Figure C.3 Total number of hospitals within a county shown as quantiles, over the range of observed values for the second variable. All other variables set at their median (religious gatherings set to 0).

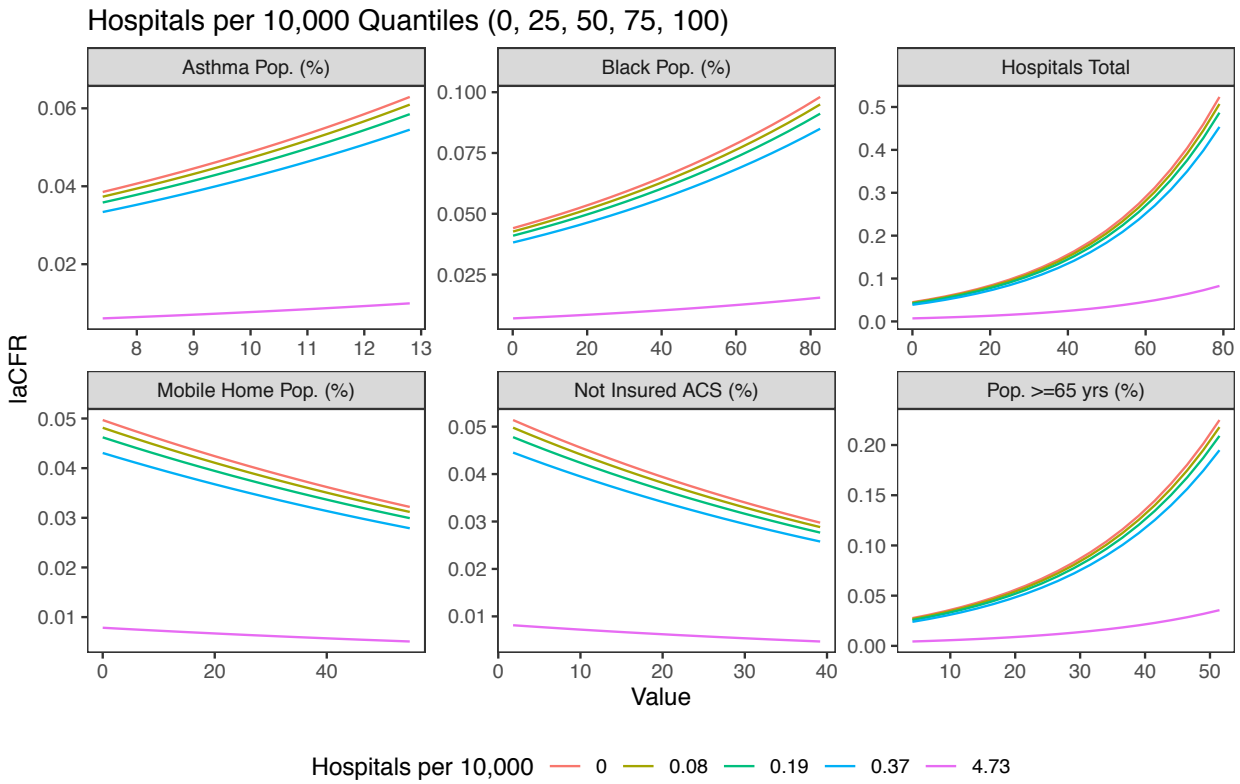

Figure C.4 Number of hospitals per 10,000 county residents shown as quantiles, over the range of observed values for the second variable. All other variables set at their median (religious gatherings set to 0).

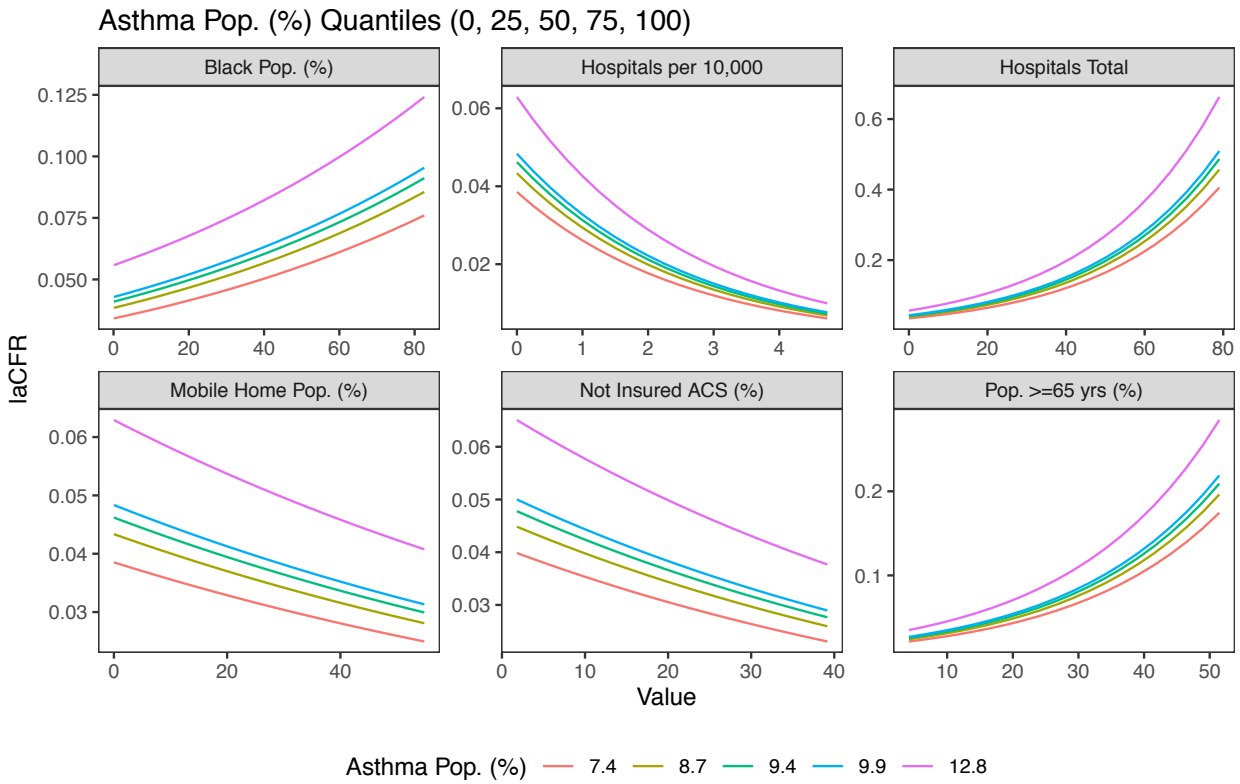

Figure C.5 Percentage asthma population within county shown as quantiles, over the range of observed values for the second variable. All other variables set at their median (religious gatherings set to 0).

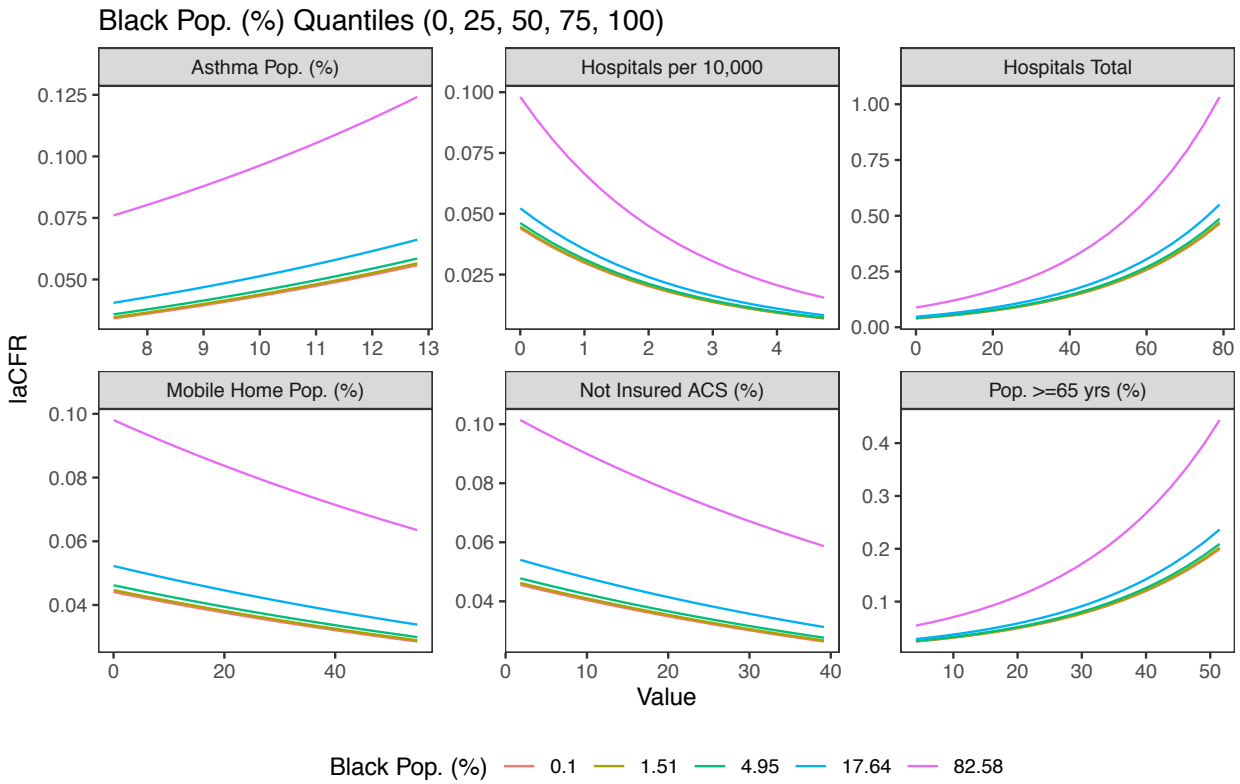

Figure C.6 Percentage Black population within county shown as quantiles, over the range of observed values for the second variable. All other variables set at their median (religious gatherings set to 0).

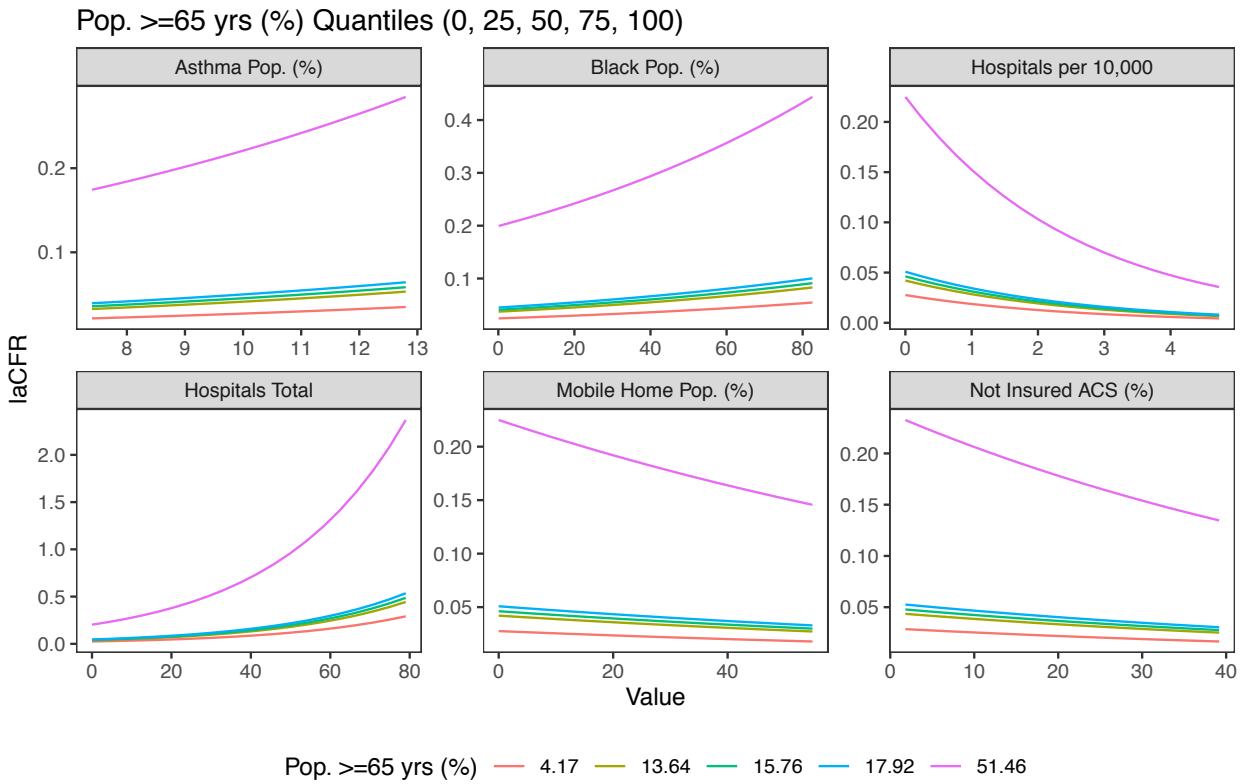

Figure C.7 Percentage population 65 years and older within county shown as quantiles, over the range of observed values for the second variable. All other variables set at their median (religious gatherings set to 0).

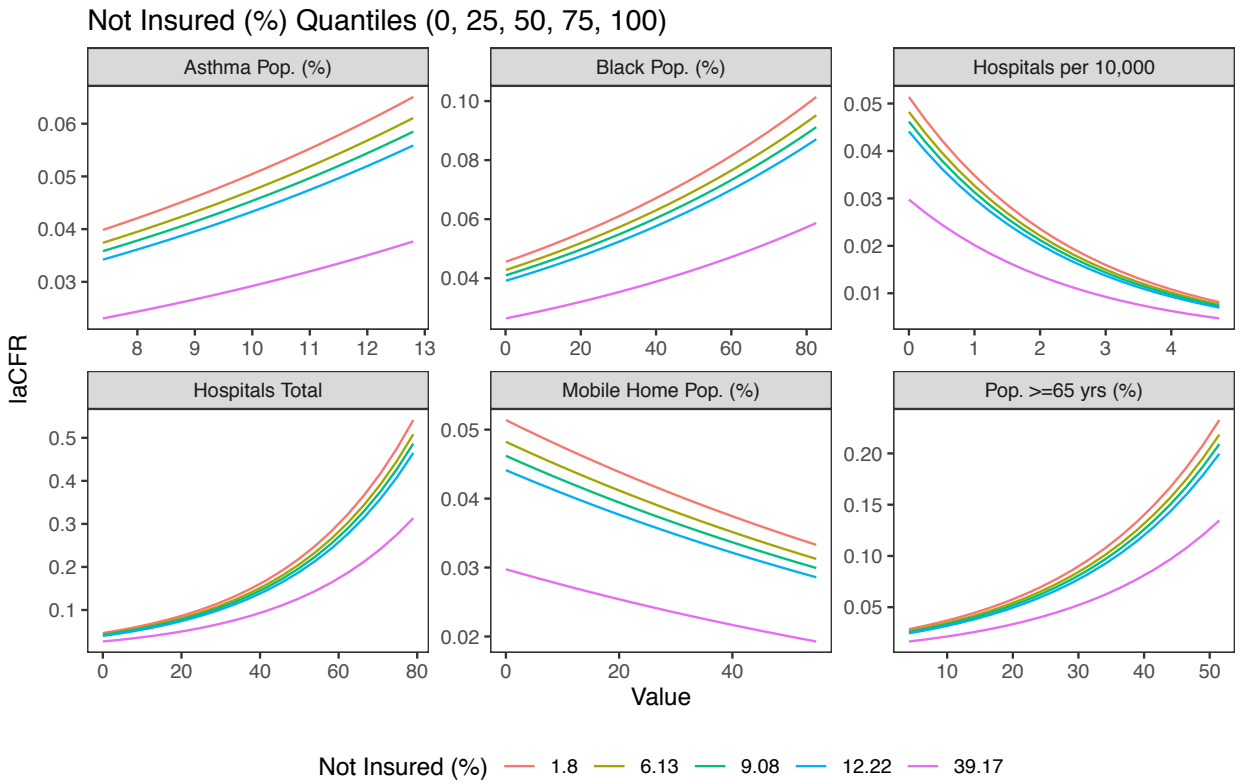

Figure C.8 Percentage population 65 uninsured within county shown as quantiles, over the range of observed values for the second variable. All other variables set at their median (religious gatherings set to 0).

Table C.1 Justifications for COVID related variable inclusion

Variable Topic	Variable Code	Variable Description	Justifications
Comorbidities	como_allheartdis_hosp	Heart Disease Hospitalization Rate per 1,000 Medicare Beneficiaries, 65+	This variable is associated with the severity of and mortality due to COVID-19.[251]
Comorbidities	como_allheartdis_mort	Heart Disease Death Rate per 100,000, 35+	This variable is associated with the severity of and mortality due to COVID-19.[251]
Comorbidities	como_asthma	Age-adjusted prevalence of adults who have been told they currently have asthma	This variable is associated with the severity of and mortality due to COVID-19.[252]
Comorbidities	como_cancer5yr	Age adjusted incidence of all cancer-5-year prevalence	This variable is associated with the severity of and mortality due to COVID-19.[253]
Comorbidities	como_COPD	Age-adjusted prevalence of adults diagnosed with chronic obstructive pulmonary disease	This variable is associated with the severity of and mortality due to COVID-19.[254]
Comorbidities	como_cvd_hosp	Total Cardiovascular Disease Hospitalization Rate per 1,000 Medicare Beneficiaries, 65+	This variable is associated with the severity of and mortality due to COVID-19.[251]
Comorbidities	como_cvd_mort	Total Cardiovascular Disease Death Rate per 100,000, All Ages	This variable is associated with the severity of and mortality due to COVID-19.[251]
Comorbidities	como_pdiabetes	Age-adjusted prevalence of adults aged 20+ years with diagnosed diabetes (in %) by county	This variable is associated with the severity of and mortality due to COVID-19.[255]
Comorbidities	como_htn_hosp	Hypertension Hospitalization Rate per	This variable is associated with the severity of and mortality due to COVID-19.[256]

		1,000 Medicare Beneficiaries, 65+	
Comorbidities	como_htn_mort	Hypertension Death Rate per 100,000 (any mention), 35+	This variable is associated with the severity of and mortality due to COVID-19.[256]
Comorbidities	como_ medicareheartdizprev	Prevalence (in %) of heart disease among Medicare beneficiaries	This variable is associated with the severity of and mortality due to COVID-19.[251]
Comorbidities	como_pobesity	Age-adjusted prevalence of adults aged 20+ years with obesity (in %) by county	This variable is associated with the severity of and mortality due to COVID-19.[257]
Comorbidities	como_smoking	Age-adjusted prevalence of adults who are current smokers (variable calculated from one or more BRFSS questions)	This variable is associated with the severity of and mortality due to COVID-19.[254]
Comorbidities	como_stroke_hosp	Stroke Hospitalization Rate per 1,000 Medicare Beneficiaries, 65+	This variable is associated with the severity of COVID-19.[258]
Comorbidities	como_stroke_mort	Stroke Death Rate per 1,000, 35+	This variable is associated with the severity of COVID-19.[258]
Demographics	demo_landarea	County land area in square meters	Historically, more rural areas saw a lower burden of infectious disease because smaller populations meant diseases were less likely to be circulating,[259] suggesting that counties with smaller populations or larger land areas may be less impacted if those who are unwell come into contact with fewer people allowing the disease to burn out before cases and CFR climb.
Demographics	demo_p60more	Percentage of population aged 60 years or older	COVID-19 has a higher fatality rate for older populations, while sparing younger ages from more severe forms of the disease.[260]

Demographics	demo_p65more	Percentage of population aged 65 years or older	COVID-19 has a higher fatality rate for older populations, while sparing younger ages from more severe forms of the disease.[260]
Demographics	demo_p45_64	Percentage of population aged 45 to 64	COVID-19 has a higher fatality rate for older populations, while sparing younger ages from more severe forms of the disease.[260]
Demographics	demo_popdensity	Population density	Population density may make social distancing more challenging and may also result in a higher effective contact rate.[200]
Demographics	demo_population	Total population of each county (same as demo_bridgedrace_total)	Historically, more rural areas saw a lower burden of infectious disease because smaller populations meant diseases were less likely to be circulating,[259] suggesting that counties with smaller populations or larger land areas may be less impacted if those who are unwell come into contact with fewer people allowing the disease to burn out before cases and CFR climb
Healthcare access & capacity	hc_hospitals	Number of Hospitals	Counties with greater healthcare resources available will presumably be able to manage a higher case-load before becoming overwhelmed.[261]
Healthcare access & capacity	hc_hospitals_per10000	Number of Hospitals per 10000	Counties with greater healthcare resources available will presumably be able to manage a higher case-load before becoming overwhelmed.[261]
Healthcare access & capacity	hc_icubeds_per1000	Number of ICU beds per 1000	Counties with greater healthcare resources available will presumably be able to manage a higher case-load before becoming overwhelmed.[261]
Healthcare access & capacity	hc_icubeds	Number of ICU beds per 1000	Counties with greater healthcare resources available will presumably be able to manage a higher case-load before becoming overwhelmed.[261]
Healthcare access & capacity	hc_icubeds_per60more	Number of ICU beds per >60 year resident	Counties with greater healthcare resources available will presumably be able to manage a higher case-load before becoming overwhelmed.[261]

Healthcare access & capacity	hc_icubeds_per65more	Number of ICU beds per >65 year resident	Counties with greater healthcare resources available will presumably be able to manage a higher case-load before becoming overwhelmed.[261]
Healthcare access & capacity	hc_medicaid	Medicaid eligible	Uninsured Americans are less likely to access health care when needed, more likely to delay treatment, are at higher risk of hospitalization, and also more likely to have preventable illnesses or uncontrolled chronic illnesses, which may put them at higher risk of serious COVID-19 illness.[262,263]
Healthcare access & capacity	hc_Pnotinsured_acs	Percentage without Health Insurance	Uninsured Americans are less likely to access health care when needed, more likely to delay treatment, are at higher risk of hospitalization, and also more likely to have preventable illnesses or uncontrolled chronic illnesses, which may put them at higher risk of serious COVID-19 illness.[262,263]
Healthcare access & capacity	hc_primarycare	Number of primary care physicians in the county	Counties with greater healthcare resources available will presumably be able to manage a higher case-load before becoming overwhelmed.[261]
Healthcare access & capacity	hc_primarycare_per1000	Primary Care Physician per capita	Counties with greater healthcare resources available will presumably be able to manage a higher case-load before becoming overwhelmed.[261]
Non-pharmaceutical intervention	npi_keystone_closing_of_public_venues	Government order closing venues such as restaurants, theaters, and bars	We expect non-pharmaceutical intervention (NPI) can have an influence on COVID-19 mortality. At the time of variable inclusion, there were no published results for COVID-19 NPI yet.
Non-pharmaceutical intervention	npi_keystone_gathering_size_10_0	Gathering size limited to 10 or fewer people	We expect non-pharmaceutical intervention (NPI) can have an influence on COVID-19 mortality. At the time of variable inclusion, there were no published results for COVID-19 NPI yet.
Non-pharmaceutical intervention	npi_keystone_gathering_size_100_to_26	Gathering size limited to 26 to 100 people	We expect non-pharmaceutical intervention (NPI) can have an influence on COVID-19 mortality. At the time of variable inclusion, there were no published results for COVID-19 NPI yet.

Non-pharmaceutical intervention	npi_keystone_gathering_size_25_to_11	Gathering size limited to 11 to 25 people	We expect non-pharmaceutical intervention (NPI) can have an influence on COVID-19 mortality. At the time of variable inclusion, there were no published results for COVID-19 NPI yet.
Non-pharmaceutical intervention	npi_keystone_gathering_size_500_to_101	Gather size limited to 101 to 500 people	We expect non-pharmaceutical intervention (NPI) can have an influence on COVID-19 mortality. At the time of variable inclusion, there were no published results for COVID-19 NPI yet.
Non-pharmaceutical intervention	npi_keystone_lockdown	Lockdown	We expect non-pharmaceutical intervention (NPI) can have an influence on COVID-19 mortality. At the time of variable inclusion, there were no published results for COVID-19 NPI yet.
Non-pharmaceutical intervention	npi_keystone_non_essential_services_closure	Government order closing non-essential services and shops	We expect non-pharmaceutical intervention (NPI) can have an influence on COVID-19 mortality. At the time of variable inclusion, there were no published results for COVID-19 NPI yet.
Non-pharmaceutical intervention	npi_keystone_Other	Other, unspecified, NPI	We expect non-pharmaceutical intervention (NPI) can have an influence on COVID-19 mortality. At the time of variable inclusion, there were no published results for COVID-19 NPI yet.
Non-pharmaceutical intervention	npi_keystone_religious_gatherings_banned	Cancellation of religious gatherings either explicitly or implicitly through gathering size limitations that do not exempt religious services	We expect non-pharmaceutical intervention (NPI) can have an influence on COVID-19 mortality. At the time of variable inclusion, there were no published results for COVID-19 NPI yet.
Non-pharmaceutical intervention	npi_keystone_school_closure	Closure of schools and university	We expect non-pharmaceutical intervention (NPI) can have an influence on COVID-19 mortality. At the time of variable inclusion, there were no published results for COVID-19 NPI yet.
Non-pharmaceutical intervention	npi_keystone_shelter_in_place	An order indicating that people should shelter in	We expect non-pharmaceutical intervention (NPI) can have an influence on COVID-19 mortality. At

		their homes except for essential reasons	the time of variable inclusion, there were no published results for COVID-19 NPI yet.
Non-pharmaceutical intervention	npi_keystone_social_distancing	Social distancing mandate of at least 6' between people	We expect non-pharmaceutical intervention (NPI) can have an influence on COVID-19 mortality. At the time of variable inclusion, there were no published results for COVID-19 NPI yet.
Socioeconomics	ses_hhincome	Median household income in the past 12 months	For most racial groups, increased income correlates with improved health.[264]
Socioeconomics	ses_pnohighschool	Percentage of no high school diploma by county	There is an established association of lower educational attainment and poorer health, including chronic illness and mortality,[265] along with the importance of high school education as a measure[266] of this association.
Socioeconomics	ses_ppoverty	Percentage of residents with income in the past 12 months below poverty level by county	For most racial groups, increased income correlates with improved health.[264]
Socioeconomics	ses_punemployed	Percentage of unemployed (not in labor force) by county	Unemployment is associated with poor health,[267] and also contributes to homelessness,[268] which is, in turn, a risk for COVID-19 infection.[269]
Social vulnerability	sv_groupquarterspop	Number of persons in institutionalized group quarters	This variable is used to construct the Social Vulnerability Index (SVI). Group living arrangements represent increased risk of SARS-CoV-2 transmission due to both difficulties in maintaining hygiene and social distancing in these settings and due to the risk that caregivers who visit multiple homes have an increased risk of both acquiring and spreading the virus.[270]
Social vulnerability	sv_p17below	Age 17 or younger	This variable is used to construct the Social Vulnerability Index (SVI). COVID-19 has a higher fatality rate for older populations, while sparing younger ages from more severe forms of the disease.[260]

Social vulnerability	sv_pcrowding	Percentage of occupied housing units with more people than rooms	This variable is used to construct the Social Vulnerability Index (SVI). This variable represents increased challenges to social distancing. For instance, an individual falling ill in a crowded apartment will have more difficulty in self-isolating than someone living in a spacious home. Individuals living in apartment complexes will have more difficulty in maintaining a 6-foot distance when outside than individuals with access to backyards.[271,272]
Social vulnerability	sv_pdisability	Older than age 5 with disability	This variable is used to construct the Social Vulnerability Index (SVI). This variable represents barriers to healthcare access and increased likelihood of greater health needs and worse outcomes from existing health conditions.[273]
Social vulnerability	sv_penglish	Percentage of population (age 5+) who speak English "less than well"	This variable is used to construct the Social Vulnerability Index (SVI). Populations with poor English skills are likely to have increased difficulty in accessing accurate health information and decreased visits to healthcare professionals.[274]
Social vulnerability	sv_pminority	Percentage minority (not non-Hispanic White)	This variable is used to construct the Social Vulnerability Index (SVI). The percentage of the county population who belongs to the racial minority is included for the same reason that different races/ethnicities were included as part of the demographic data above (Essentially, this is a grouping that includes all except non-Hispanic White. It can be understood as it might be better to use a composite variable to increase statistical power). Historically, more rural areas saw a lower burden of infectious disease because smaller populations meant diseases were less likely to be circulating,[259] suggesting that counties with smaller

			populations or larger land areas may be less impacted if those who are unwell come into contact with fewer people allowing the disease to burn out before cases and CFR climb.
Social vulnerability	sv_pmobilehome	Percentage of total housing units with mobile home	This variable is used to construct the Social Vulnerability Index (SVI). This variable represents increased challenges to social distancing. For instance, an individual falling ill in a crowded apartment will have more difficulty in self-isolating than someone living in a spacious home. Individuals living in apartment complexes will have more difficulty in maintaining a 6-foot distance when outside than individuals with access to backyards.[267,268]
Social vulnerability	sv_pmultiunit	Percentage of total housing units with 10 or more units	This variable is used to construct the Social Vulnerability Index (SVI). This variable represents increased challenges to social distancing. For instance, an individual falling ill in a crowded apartment will have more difficulty in self-isolating than someone living in a spacious home. Individuals living in apartment complexes will have more difficulty in maintaining a 6-foot distance when outside than individuals with access to backyards.[267,268]
Social vulnerability	sv_pnovehicle	Percentage of households with no vehicle available	This variable is used to construct the Social Vulnerability Index (SVI). Increased reliance on public transport will create more crowded transport and a higher risk of transmission.[275]
Social vulnerability	sv_singleparent	Single-parent household with children under 18	This variable is used to construct the Social Vulnerability Index (SVI). These households are likely to experience increased difficulties finding childcare. The potential impact on absenteeism for

			healthcare workers could lead to higher mortality rates[276] if more of the workforce are single parents.
Demographics	demo_bridgedrace_p_ american_indians_alaskan	Percentage of (non Hispanic) American Indian or Alaska Native	In the US structural racism leads to racial/ethnic populations' lack of access to health care and receipt of low-quality health care, contributing to substantial health disparities,[277] which may, in turn, result in worse outcomes for COVID-19 patients.
Demographics	demo_bridgedrace_p_ asians_pacific	Percentage of (non-Hispanic) Asian or Pacific Islander.	In the US structural racism leads to racial/ethnic populations' lack of access to health care and receipt of low-quality health care, contributing to substantial health disparities,[277] which may, in turn, result in worse outcomes for COVID-19 patients.
Demographics	demo_bridgedrace_p_ blacks	Percentage of (non-Hispanic) Black or African American	In the US structural racism leads to racial/ethnic populations' lack of access to health care and receipt of low-quality health care, contributing to substantial health disparities,[277] which may, in turn, result in worse outcomes for COVID-19 patients.
Demographics	demo_bridgedrace_p_hisp	Percentage of Hispanic or Latino	In the US structural racism leads to racial/ethnic populations' lack of access to health care and receipt of low-quality health care, contributing to substantial health disparities,[277] which may, in turn, result in worse outcomes for COVID-19 patients.
Demographics	demo_bridgedrace_p_ whites	Percentage of (non-Hispanic) White	In the US structural racism leads to racial/ethnic populations' lack of access to health care and receipt of low-quality health care, contributing to substantial health disparities,[277] which may, in turn, result in worse outcomes for COVID-19 patients.
Demographics	demo_bridgedrace_total	Total population of each county (same as demo_population)	Historically, more rural areas saw a lower burden of infectious disease because smaller populations meant diseases were less likely to be circulating,[259] suggesting that counties with smaller populations or larger land areas may be less impacted if those who are unwell come into contact with fewer

			people allowing the disease to burn out before cases and CFR climb
Time	days_since_first_case	Number of days between first detected COVID19 case and final date of included case data, New York City data by county from NYC public health website, Kansas City counties were excluded	We expect this variable can have an influence on COVID-19 mortality. At the time of variable inclusion, there were no published results for this variable yet.

Table C.2 COVID related variables and data sources

Model Inclusion Status	Variable Code	Level Data	Data Source	Year(s) Collected	Variable Unit Description
Included, linking variable	FIPS	County	US Census TIGER shapefile	2018	ID number
Excluded, highly correlated	como_allheartdis_hosp	County	CDC, Interactive Atlas of Heart Disease and Stroke	2016-2018	Incidence per 1000, 65+
Excluded, highly correlated	como_allheartdis_mort	County	CDC, Interactive Atlas of Heart Disease and Stroke	2014-2016	Incidence per 100,000
Included in final model	como_asthma	State	BRFSS	2018	Adjusted prevalence %
Excluded, non-significant in multivariate model	como_cancer5yr	County if available, o/w State	NIH, National Cancer Institute, State Cancer Profiles	2012-2016	5-year incidence
Excluded, highly correlated	como_COPD	State	BRFSS	2018	Adjusted prevalence %
Excluded, highly correlated	como_cvd_hosp	County	CDC, Interactive Atlas of Heart Disease and Stroke	2016-2018	Incidence per 1000, ages 65+
Excluded, highly correlated	como_cvd_mort	County	CDC, Interactive Atlas of Heart Disease and Stroke	2016-2018	Incidence per 100,000
Excluded, non-significant in bivariate model	como_pdiabetes	County	CDC, US Diabetes Surveillance System	2016	Percentage (%)
Excluded, highly correlated	como_htn_hosp	County	CDC, Interactive Atlas of Heart Disease and Stroke	2016-2018	Incidence per 1000, 65+
Excluded, non-significant in bivariate model	como_htn_mort	County	CDC, Interactive Atlas of Heart Disease and Stroke	2016-2018	Incidence per 1000
Excluded, non-significant in multivariate model	como_ medicareheartdizprev	County	CDC, Interactive Atlas of Heart Disease and Stroke	2018	Prevalence per 1000, medicare beneficiaries

Excluded, highly correlated	como_pobesity	County	CDC, Interactive Atlas of Heart Disease and Stroke	2016	Percentage (%)
Excluded, non-significant in bivariate model	como_smoking	State	BRFSS	2018	Age adjusted prevalence
Excluded, highly correlated	como_stroke_hosp	County	CDC, Interactive Atlas of Heart Disease and Stroke	2018	Incidence per 1000
Excluded, highly correlated	como_stroke_mort	County	CDC, Interactive Atlas of Heart Disease and Stroke	2014-2016	Incidence per 1000
Excluded, non-significant in bivariate model	demo_landarea	County	US Census TIGER shapefile	2018	Area in square kilometer
Excluded, highly correlated	demo_p60more	County	Bridged race	2010-2018 average	Percentage (%)
Included in final model	demo_p65more	County	Bridged race	2010-2018 average	Percentage (%)
Excluded, highly correlated	demo_p45_64	County	Bridged race	2010-2018 average	Percentage (%)
Excluded, highly correlated	demo_popdensity	County	Bridged race	2018	Person per square kilometer
Excluded, highly correlated	demo_population	County	Bridged race	2018	Count
Included in final model	hc_hospitals	County	CDC, Interactive Atlas of Heart Disease and Stroke and demo_population variable	2018	Total number of hospitals in county
Included in final model	hc_hospitals_per10000	County	CDC, Interactive Atlas of Heart Disease and Stroke and demo_population variable	2018	Number of hospitals per 10,000 people in county
Excluded, highly correlated	hc_icubeds_per1000	County	Kaiser Health News analysis of hospital cost reports filed to the Centers for Medicare	2018/2019	Number per 1000 persons in the county

			& Medicaid Services American Community Survey, (5-year estimate)		
Excluded, highly correlated	hc_icubeds	County	Kaiser Health News analysis of hospital cost reports filed to the Centers for Medicare & Medicaid Services American Community Survey, (5-year estimate)	2018/2019	Number of ICU beds in county
Excluded, non-significant in multivariate model	hc_icubeds_per60more	County	Kaiser Health News analysis of hospital cost reports filed to the Centers for Medicare & Medicaid Services	2018/2019	Number per 1000 persons aged 60+
Excluded, highly correlated	hc_icubeds_per65more	County	Kaiser Health News analysis of hospital cost reports filed to the Centers for Medicare & Medicaid Services	2018/2019	Number per 1000 persons aged 65+
Excluded, highly correlated	hc_medicaid	County	CDC, Interactive Atlas of Heart Disease and Stroke	2018	Percentage (%)
Included in final model	hc_Pnotinsured_acs	County	US Census American Community Survey 5-Year Data	2018	Percentage (%)
Excluded, non-significant in bivariate model	hc_primarycare	County	Health Resources and Services Administration, (Area Health Resources File)	2016	Count
Excluded, highly correlated	hc_primarycare_per1000	County	Health Resources and Services Administration, (Area Health Resources File)	2016	Adjusted incidence rate per 1000
Excluded, highly correlated	npi_keystone_closing_of_public_venues	County	KeyStone Coronavirus City and County Non-Pharmaceutical Intervention Rollout Date Dataset	2020	Date

Excluded, highly correlated	npi_keystone_gathering_size_10_0	County	KeyStone Coronavirus City and County Non-Pharmaceutical Intervention Rollout Date Dataset	2020	Date
Excluded, highly correlated	npi_keystone_gathering_size_100_to_26	County	KeyStone Coronavirus City and County Non-Pharmaceutical Intervention Rollout Date Dataset	2020	Date
Excluded, highly correlated	npi_keystone_gathering_size_25_to_11	County	KeyStone Coronavirus City and County Non-Pharmaceutical Intervention Rollout Date Dataset	2020	Date
Excluded, highly correlated	npi_keystone_gathering_size_500_to_101	County	KeyStone Coronavirus City and County Non-Pharmaceutical Intervention Rollout Date Dataset	2020	Date
Excluded, highly correlated	npi_keystone_lockdown	County	KeyStone Coronavirus City and County Non-Pharmaceutical Intervention Rollout Date Dataset	2020	Date
Excluded, highly correlated	npi_keystone_non_essential_services_closure	County	KeyStone Coronavirus City and County Non-Pharmaceutical Intervention Rollout Date Dataset	2020	Date
Excluded, highly correlated	npi_keystone_Other	County	KeyStone Coronavirus City and County Non-Pharmaceutical Intervention Rollout Date Dataset	2020	Date
Included in final model	npi_keystone_religious_gatherings_banned	County	KeyStone Coronavirus City and County Non-Pharmaceutical Intervention Rollout Date Dataset	2020	Date

Excluded, highly correlated	npi_keystone_school_closure	County	KeyStone Coronavirus City and County Non-Pharmaceutical Intervention Rollout Date Dataset	2020	Date
Excluded, highly correlated	npi_keystone_shelter_in_place	County	KeyStone KeyStone Coronavirus City and County Non-Pharmaceutical Intervention Rollout Date Dataset	2020	Date
Excluded, non-significant in multivariate model	npi_keystone_social_distancing	County	KeyStone Coronavirus City and County Non-Pharmaceutical Intervention Rollout Date Dataset	2020	Date
Excluded, highly correlated	ses_hhincome	County	US Census American Community Survey 5-Year Data	2018	Median income in US Dollars
Excluded, highly correlated	ses_pnohighschool	County	US Census American Community Survey 5-Year Data	2018	Percentage (%)
Excluded, highly correlated	ses_ppoverty	County	US Census American Community Survey 5-Year Data	2018	Percentage (%)
Excluded, non-significant in multivariate model	ses_punemployed	County	US Census American Community Survey 5-Year Data	2018	Percentage (%)
Excluded, highly correlated	sv_groupquarterspop	County	US Census American Community Survey 5-Year Data	2018	Count
Excluded, highly correlated	sv_p17below	County	US Census American Community Survey 5-Year Data	2018	Percentage (%)

Excluded, non-significant in multivariate model	sv_pcrowding	County	US Census American Community Survey 5-Year Data	2018	Percentage (%)
Excluded, highly correlated	sv_pdisability	County	US Census American Community Survey 5-Year Data	2018	Percentage (%)
Excluded, highly correlated	sv_penglish	County	US Census American Community Survey 5-Year Data	2018	Percentage (%)
Excluded, highly correlated	sv_pminority	County	CDC SVI	2018	Percentage (%)
Included in final model	sv_pmobilehome	County	US Census American Community Survey 5-Year Data	2018	Percentage (%)
Excluded, highly correlated	sv_pmultiunit	County	US Census American Community Survey 5-Year Data	2018	Percentage (%)
Excluded, non-significant in multivariate model	sv_pnovehicle	County	US Census American Community Survey 5-Year Data	2018	Percentage (%)
Excluded, highly correlated	sv_singleparent	County	US Census American Community Survey 5-Year Data	2018	Percentage (%)
Excluded, non-significant in bivariate model	demo_bridgedrace_p_ american_indians_alaskan	County	CDC, National Center for Health Statistics	2010-2018	Percentage (%)
Excluded, non-significant in bivariate model	demo_bridgedrace_p_ asians_pacific	County	CDC, National Center for Health Statistics	2010-2018	Percentage (%)
Included in final model	demo_bridgedrace_p_ blacks	County	CDC, National Center for Health Statistics	2010-2018	Percentage (%)
Excluded, highly correlated	demo_bridgedrace_p_hisp	County	CDC, National Center for Health Statistics	2010-2018	Percentage (%)

Excluded, highly correlated	demo_bridgedrace_p_ whites	County	CDC, National Center for Health Statistics	2010-2018	Percentage (%)
Excluded, highly correlated	demo_bridgedrace_total	County	CDC, National Center for Health Statistics	2010-2018	Percentage (%)
Excluded, highly correlated	days_since_first_case	County	NYT COVID-19 Dataset	Jan 21-Jun 12, 2020	Days

Table C.3 Descriptive statistics for county variables retained for analysis. Median and range shown for continuous variables.

Variable Code	Training Set (n=1186)		Testing Set (n=593)		Excluded (n=1364)	
como_asthma (%)	9.4	(7.4–12.8)	9.2	(7.4–12.8)	9.2	(7.4–12.8)
como_cancer5yr (cases/5yr)	463.3	(272.1–1135)	457.8	(241–592.1)	451.0	(130.1–677.2)
como_pdiabetes (%)	10.1	(1.5–33.0)	10.0	(1.7–24.6)	9.5	(1.8–32.3)
como_htn_mort (deaths/100,000)	120.2	(20.4–400.6)	120.2	(18.7–442.2)	129.7	(26.5–592.1)
como_medicareheartdizprev (%)	36.0	(22.1–55.2)	35.7	(19.5–49.3)	35.3	(18.0–53.5)
como_smoking (%)	17.3	(9.0–26.8)	17.3	(9.0–26.8)	17.7	(9.0–26.8)
demo_landarea (m^2)	1535.8	(6.5–64008)	1505.1	(38.8–51954)	1743.8	(5.3–377034)
demo_p65more (%)	15.8	(4.2–51.5)	15.6	(6.5–27.6)	19.1	(5.1–36.1)
demo_bridgedrace_p_american_indians_alaskan (%)	0.3	(0.0–73.4)	0.3	(0.1–92.2)	0.5	(0.0–93.9)
demo_bridgedrace_p_asians_pacific (%)	0.9	(0.1–65.5)	1.0	(0.1–30.4)	0.5	(0.0–59.1)
demo_bridgedrace_p_blacks (%)	4.6	(0.1–82.6)	5.5	(0.2–78.5)	1.1	(0.0–85.7)
hc_hospitals (hospitals)	1	(0–32)	1	(0–79)	1	(0–8)
hc_hospitals_per10000 (hospitals/10,000)	0.2	(0.0–3.8)	0.2	(0.0–4.7)	0.4	(0.0–8.5)
hc_icubeds_per60more (beds/>60yr resident)	0.6	(0.0–8.2)	0.7	(0.0–7.0)	0.0	(0.0–101.1)
hc_pnotinsured_acs (%)	9.0	(1.8–39.2)	9.4	(2.0–35.6)	9.3	(1.7–45.6)
hc_primarycare (physicians)	1.9	(0.2–46.6)	1.8	(0.4–17.9)	2.3	(0.2–19.9)
npi_keystone_religious_gatherings_banned (% counties that ban religious gatherings)	52.4		54.8		47.3	
npi_keystone_social_distancing (days since 1st county case)	3	(-46–76)	3	(-30–82)	-8	(-84–46)
ses_punemployed (%)	3.3	(0.5–9.5)	3.3	(0.5–13.6)	2.9	(0.0–16.5)
sv_pcrowding (%)	0.8	(0.0–10.1)	0.9	(0.0–15.4)	0.9	(0.0–35.5)
sv_pmobilehome (%)	9.5	(0.0–54.8)	8.7	(0.0–51.2)	12.2	(0.0–59.3)
sv_pnovehicle (%)	5.8	(1.4–77.0)	5.9	(1.0–32.2)	5.4	(0.0–87.8)

Table C.4 Regression results. County-level predictors of COVID-19 laCFR in the United States. $R^2 = 0.8620$.

Variable	Exp (Coeff.)	95% CI	Std. Error	Wald	*p*-value	VIF
Intercept	0.0111	(0.0062, 0.0200)	0.2961	-15.2016	<0.0001	
Hospitals per 10,000	0.6773	(0.5549, 0.8248)	0.1013	-3.8485	0.0001	1.0636
Religious Gatherings Ban	0.8752	(0.7894, 0.9702)	0.0526	-2.5315	0.0114	1.0592
Pop. Not Insured (%)	0.9855	(0.9715, 0.9998)	0.0075	-1.9444	0.0519	1.6001
Mobile Home Pop. (%)	0.9921	(0.9854, 0.9989)	0.0035	-2.2539	0.0242	1.7227
Asthma Pop. (%)	1.0951	(1.0400, 1.1533)	0.0264	3.4378	0.0006	1.1489
Pop. >= 65 Yrs. (%)	1.0453	(1.0308, 1.0605)	0.0069	6.4394	<0.0001	1.1405
Total Hospitals in County	1.0316	(1.0100, 1.0522)	0.0099	3.1329	0.0017	1.1663
Black Pop. (%)	1.0097	(1.0063, 1.0133)	0.0018	5.5012	<0.0001	1.2210

Appendix D Modulation of Bacterial Fitness and Virulence Through Antisense RNAs

This appendix is a published work:

Millar JA, Raghavan R. (2021). Modulation of bacterial fitness and virulence through antisense RNAs. Front Cell Infect Microbiol. 10:596277.

D.1 Abstract

Regulatory RNAs contribute to gene expression control in bacteria. Antisense RNAs (asRNA) are a class of regulatory RNAs that are transcribed from opposite strands of their target genes. Typically, these untranslated transcripts bind to cognate mRNAs and rapidly regulate gene expression at the post-transcriptional level. In this article, we review asRNAs that modulate bacterial fitness and increase virulence. We chose examples that underscore the variety observed in nature including, plasmid- and chromosome-encoded asRNAs, a riboswitch-regulated asRNA, and asRNAs that require other RNAs or RNA-binding proteins for stability and activity. We explore how asRNAs improve bacterial fitness and virulence by modulating plasmid acquisition and maintenance, regulating transposon mobility, increasing resistance against bacteriophages, controlling flagellar production, and regulating nutrient acquisition. We conclude with a brief discussion on how this knowledge is helping to inform current efforts to develop new therapeutics.

D.2 Introduction

A major breakthrough in biology was the discovery of non-coding RNAs (ncRNAs) that regulate gene expression instead of coding for proteins. ncRNAs play important regulatory roles in all domains of life.[278–281] In bacteria, ncRNAs regulate gene expression at the post-transcriptional level by binding to messenger RNAs (mRNAs) to control several processes, including pathogenesis.[282–284] Typically, ncRNAs that are encoded on the opposite strands of target genes (complementary to sense transcript) are known as cis-acting antisense RNAs (asRNAs), while an ncRNA that is encoded in a separate part of the genome in relation to its target mRNAs is called a trans-acting small RNA (sRNA).[280,285,286] Regulatory RNAs generally have an advantage over regulatory proteins because their synthesis require lower energy and they act rapidly. In addition, their co-degradation along with target mRNAs allow precise control of regulatory circuits, which is key for bacteria to quickly adapt to host immune response.[287,288] asRNAs are particularly useful for rapid gene regulation because they bind to target mRNAs with perfect complementarity, whereas sRNAs form imperfect complementarity with target mRNAs and often require chaperone proteins such as Hfq and ProQ for stability and function.[280,289–291]

Initially, asRNAs were thought to be rare in bacteria, and the pervasive antisense transcription observed in microarray-based studies were assumed to be experimental artifacts.[280,292,293] Even with the advent of high-throughput sequencing, it was initially difficult to differentiate between bona fide asRNAs and transcriptional noise because of low sequence coverage.[294–296] With the increase in sequencing resolution, recent studies have confirmed the presence of abundant asRNAs in bacteria and have revealed it to be a genome-wide phenomenon.[280,297–299]

asRNAs have been shown to modulate bacterial pathogenicity by either regulating the expression of virulence genes[300] or by controlling biochemical processes that improve bacterial fitness, which in turn boosts virulence.[282,301] In this mini review, we focus on the latter. In particular, we cover examples where we generally understand the mechanism of action and where the genome locations of asRNAs have been determined. We also chose examples that underscore the variety observed in nature, including short asRNAs and long asRNAs, those found in plasmids and those encoded on genomes, asRNAs that require binding stability from other RNAs or proteins, and asRNAs that work in concert with riboswitches. These examples are presented in several sections based on the main roles asRNAs play in pathogenesis: 1) acquisition and regulation of virulence plasmids, 2) modulation of transposon mobility, 3) increasing resistance against bacteriophages, 4) controlling flagellar production, and 5) regulating nutrient acquisition.

D.3 Modes of antisense RNA-based improvement in fitness and virulence

D.3.1 Acquisition and maintenance of virulence plasmids

A major avenue through which bacteria acquire new virulence factors is by obtaining new plasmids *via* conjugation, a process that involves asRNAs. An example of this is F-like plasmids, which are part of a large group of conjugative plasmids frequently found in *Escherichia coli* and throughout Enterobacteriaceae.[302] These plasmids frequently harbor accessory genes, including antibiotic resistance genes, enterotoxins, and other virulence genes.[303] Conjugation is encoded by the Tra-operon, with initiation requiring TraJ. Initiation is regulated by FinOP, which consists of the RNA binding protein FinO and FinP, an asRNA. FinP attaches to the ribosome binding site of *traJ*, inhibiting its translation and promoting mRNA degradation

(Figure D.1A). FinO contributes by helping promote FinP binding to *traJ*, as well as protecting FinP from RNase E cleavage.[304] FinP levels are controlled by RNase E digestion, preventing binding to *traJ*. These processes play out temporally during conjugation, starting with initial high level of *traJ* expression, followed by dampening and repression, maintaining bacterial fitness by reducing the metabolic burden of the plasmid.[305]

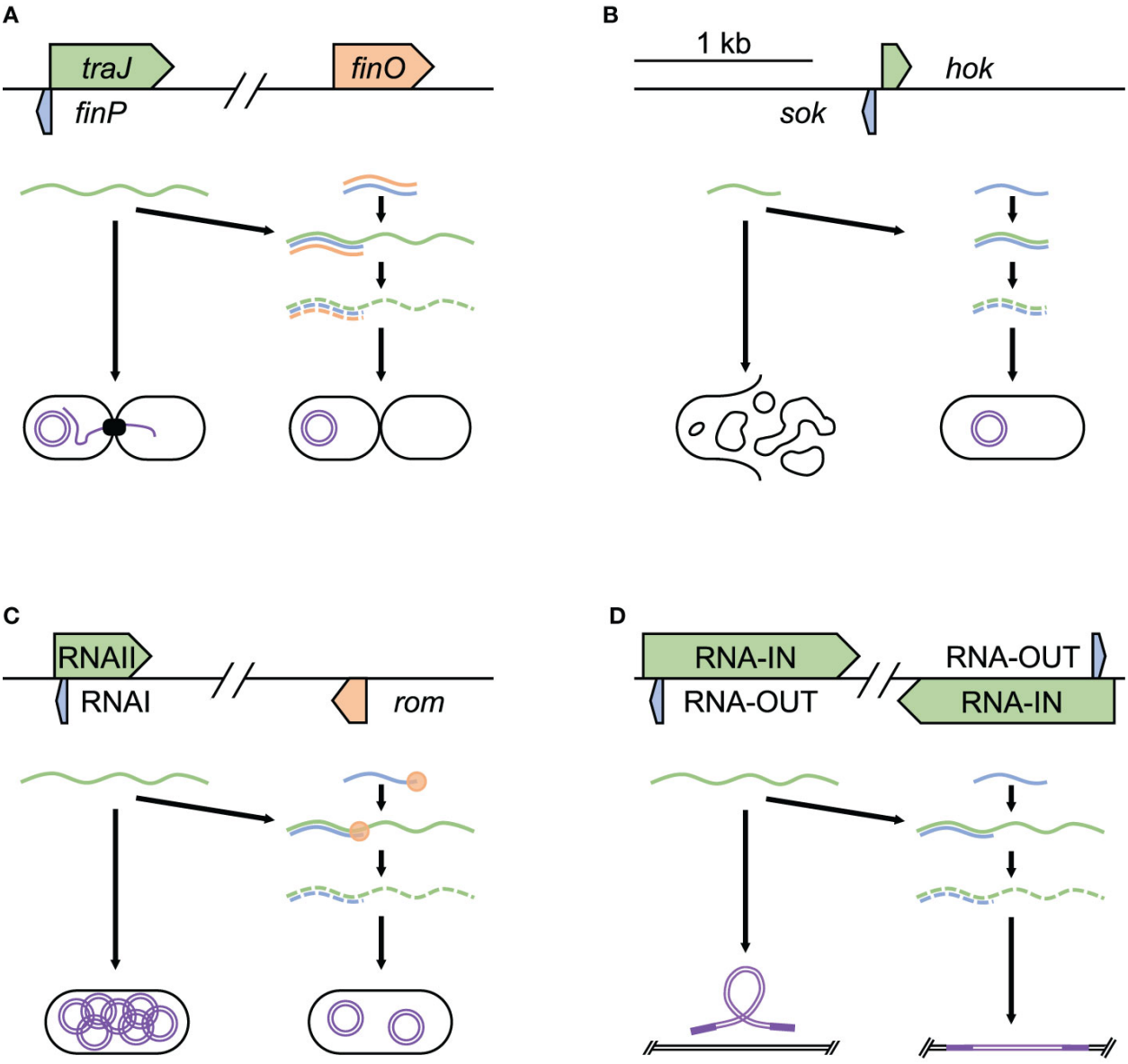

Figure D.1 Acquisition and regulation of virulence plasmids and modulation of transposons. Examples of asRNA post-transcriptional regulation of acquisition and regulation of virulence

plasmids and modulation of transposons. (A) *finP* control of plasmid acquisition through conjugation in *E. coli* F plasmid (NC_002483.1), (B) *hok/sok* toxin-antitoxin to maintain plasmids in *S. flexneri* R100 plasmid (NC_002134.1), (C) RNAI control of plasmid copy number in *E. coli* pColK plasmid (NC_006881.1), and (D) RNA-OUT regulation of transposase expression, controlling transposon movement in *S. flexneri* R100 plasmid (NC_002134.1). Each panel shows the relative position and size of the asRNA (blue), the target it regulates (green), and any factors required for stable binding (orange). In these examples, the successful binding of asRNA to its target promotes degradation.

Once bacteria acquire advantageous virulence factors through plasmids, some plasmids are retained through toxin/antitoxin systems.[306] These systems function by encoding a toxin and a paired strong antitoxin — many of which function as asRNAs, on the plasmid. During cell division, loss of the plasmid in a daughter cell results in loss of the strong antitoxin, leading to the death of cells without a plasmid copy. A well-studied system is the *hok/sok* system of R1 plasmids in *E. coli* and R100 in *Shigella flexneri*, known for harboring various antibiotic resistance genes.[307,308] This system encodes the Hok (host killing) toxin, which leads to cell death by depolarization of the cell membrane,[309] and Sok (suppression of killing), an asRNA antitoxin, which degrades very quickly.[306] Sok acts by binding to *hok* mRNA to block translation of the toxin (Figure D.1B). Within the *E. coli* chromosome, *sok* gene has a very weak promoter, resulting in the production of small amounts of antitoxin that are degraded quickly and are unable to keep up with the Hok toxin, leading to cell death. On the R1 plasmid, the *sok* gene has a strong promoter, producing many times more of Sok than Hok.[310] Hence, if the cell contains an R1 plasmid, excess Sok continues to bind all of Hok and prevent cell death. This ensures that after cell division, *E. coli* daughter cells will survive only if they maintain the plasmid. Thus, the Sok asRNA maintains bacterial fitness by promoting the retention of the R1/R100 plasmid, which has been found to improve bacterial stress response and growth in growth-limiting conditions.[311]

Replication control is another asRNA-based mechanism used by bacteria to maintain plasmids. The presence of too many copies of a plasmid can increase the metabolic burden of the cell, lowering fitness through reduced growth rate and weakened competitiveness.[312,313] However, too few plasmid copies could result in the loss of a potentially useful plasmid in subsequent generations.[314,315] Bacteria encode plasmid copy number control systems in order to maintain optimal number of plasmids. One that has been widely studied is found in ColE1-related plasmids[316] present in *E. coli*. The plasmid is named for containing the gene that encodes Colicin E1, the product of which is active against *E. coli*, as well as containing a gene for conferring immunity to Colicin E1. Under stressful conditions such as nutrient depletion, overcrowding, or antibiotics *E. coli* express Colicin E1, which promotes bacterial proliferation in mixed microbe niches such as the intestinal tract.[317,318] To replicate the plasmid, RNAII (a pre-primer) attaches to DNA at origin of replication. RNAII is then trimmed into a primer, which initializes plasmid replication. The 5′ region of RNAII contains the asRNA RNAI, which inhibits ColE1 plasmid replication (Figure D.1C). RNAI inhibits plasmid replication with the help of the Rom protein by binding to RNAII, preventing RNAII from binding to the plasmid origin of replication. As the copy number of ColE1 plasmid increases, so does the concentration of RNAI, resulting in a balance of copy number through negative control.[319] This ensures that there are enough copies of the virulence plasmid to pass on to daughter cells, while maintaining fitness by reducing the metabolic burden of what are often large — sometimes hundreds of kilobases long, plasmids.[320]

D.3.2 Regulation of transposon mobility

Another role for asRNAs in maintaining bacterial fitness and virulence is by controlling the movement of transposons, which are genetic elements that move from one position to another within a genome.[321] Insertion of transposons in virulence associated genes could reduce a bacterium's pathogenicity or increase its susceptibility to antibiotics.[322–324] In addition, transposable elements could modulate virulence by affecting biofilm formation[325–327] and reduce fitness by interrupting metabolic genes.[328–330] Bacteria defend against this by controlling transposases, the enzymes required for transposons' mobility. An example of inhibition of bacterial transposase can be seen in the Tn10 transposable element, which is found in *S. flexneri* and other Enterobacteriaceae.[331] Tn10 contains a number of tetracycline resistance genes and a pair of IS10 insertion sequences that each encode transposases that promote transposon mobility. IS10s also encode antisense RNA (RNA-OUT), which is found in the 5′-most segment of the transposase mRNA (RNA-IN). RNA-OUT inhibits transposase translation by binding RNA-IN and blocking ribosome binding site (Figure D.1D). As the Tn10 copy number increases, asRNA increases to suppress the transposition of the transposon;[332] thus, bacteria are able to maintain fitness by reducing the chance of mobile elements disrupting essential genes or virulence factors.

D.3.3 Modulation of bacteriophages

During infection, pathogenic bacteria have to outcompete other bacteria that share their niche and defend against both internal and external threats. One of the major dangers that bacteria face is from bacteriophages, many of which have lysogenic and lytic growth cycles.[333] In the lysogenic cycle, phage DNA is integrated into the bacterial chromosome, allowing replication of the phage to occur more passively along with that of the bacterium. In lytic reproduction, the phage actively creates a large number of progeny and quickly lyse the bacterial

cell to continue its lifecycle. Temperate phages include both cycles and are found in about half of microbial genomes currently sequenced.[334] While attempting to block infection by phages with lytic growth cycles often reduces virulence and fitness, allowing temperate phages to stay in the lysogenic cycle benefit bacteria by delaying eventual cell lysis.[335–338] An example of this can be seen in *Salmonella*'s maintenance of lysogeny in P22 phages.[339] In P22, lytic growth is inhibited by the regulatory protein C2, which blocks the transcription of proteins needed for the development of lytic cycle. The switch to lytic growth is brought on by the anti-repressor protein Ant, which blocks C2 binding to the O_R and O_L operators of the P22 phage. Repression of this progression into the lytic replication cycle can be accomplished through Sar, an asRNA in the intergenic region of *arc-ant* mRNA.[340] Sar blocks the *ant* ribosome binding site, which results in a failure to produce Ant and thus increasing bacterial fitness by preventing the escape of the prophage from the lysogenic state (Figure D.2A).

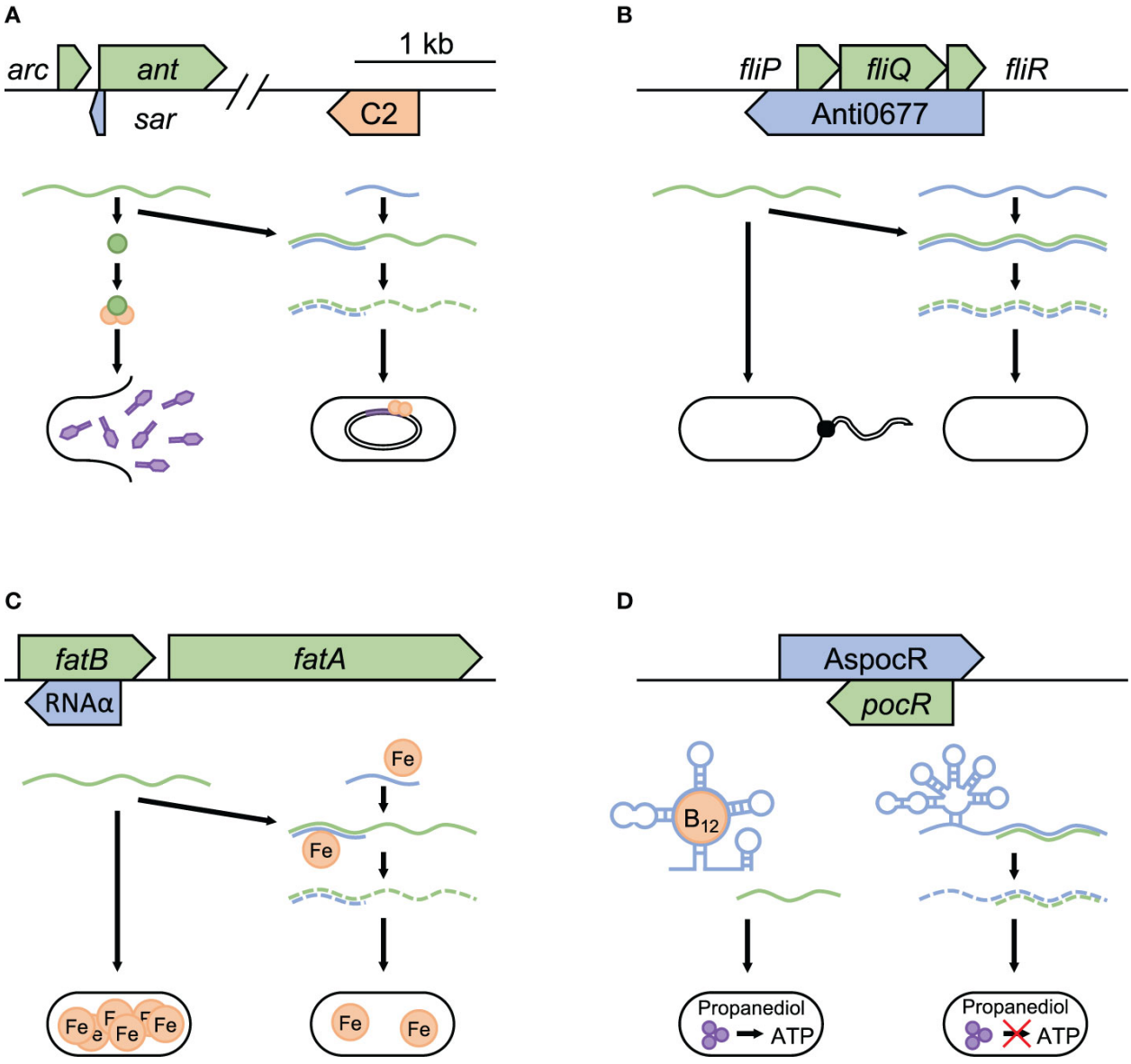

Figure D.2 Modulation of bacteriophages, flagella production, and nutrient acquisition. Examples of asRNA post-transcriptional regulation in pathogenic bacteria to increase fitness during infection. Repression of temperate bacteriophages is seen in (A) *sar* repression of lytic growth in Enterobacteria P22 phage (NC_002371.2). *sar* blocks the anti-repressor *ant* from binding to C2 (orange), which prevents escape from lysogenic growth. Temperature dependent regulation in bacteria is represented by (B) Anti0677 control of flagella formation in response to temperature change in *L. monocytogenes* (NC_003210.1). Examples that highlight regulation of nutrient acquisition can be seen in (C) RNAα control of iron acquisition in *V. anguillarum* pJM1 plasmid (NC_005250.1) and (D) AspocR/B_{12} riboswitch control of opportunistic propanediol catabolism in *L. monocytogenes* (NC_003210.1). Each panel shows the relative position and size of the asRNA (blue), and the target it regulates (green). In (A–C), the successful binding of the asRNA to its target promotes mRNA degradation, with iron required in (C) for stable binding. In (D), when B_{12} is absent, a full-length version of the asRNA is transcribed, which binds to pocR

mRNA and blocks propanediol fermentation. When B_{12} is present and binds to the riboswitch, transcription ends prematurely, resulting in the production of PocR, which promotes propanediol fermentation.

D.3.4 Control of flagella production

By modifying their outer structures bacteria evade immune response and improve persistence within hosts. An example of asRNA's involvement in this process is observed in *Listeria monocytogenes*, where a long asRNA regulates flagella production in response to temperature.[341] At 30°C, *L. monocytogenes* expresses flagella on its surface and exhibits swimming motility. Producing flagella requires the expression of a number of genes, including the flagellum export apparatus genes (*fliP*, *fliQ*, and *fliR*). When the temperature rises to 37°C, the motility gene repressor *mogR* switches off flagella formation.[342] Overlapping *fliP*, *fliQ*, and *fliR* is a large asRNA, Anti0677, which negatively regulates their expression by promoting mRNA degradation by direct interaction (Figure D.2B). Additionally, the end of Anti0677 both contains the coding sequence for and drives the expression of MogR. These two effects — the antisense component Anti0677 and increased expression of MogR — together suppress flagella formation within the host, possibly reducing the host inflammatory response attempting to lyse the invading bacteria.[343] The term "excludon" has been proposed for transcripts such as Anti0677 that both code for proteins and regulate the expression of multiple genes or operons encoded divergently from them.[344]

D.3.5 Regulation of nutrient acquisition

Bacteria can optimize their growth rates by modulating nutrient acquisition. This capability bestows increased fitness to pathogens by allowing them to survive under nutrient-

poor conditions such as infections.[345,346] An example of this phenomenon is iron uptake suppression in the fish pathogen *Vibrio anguillarum*.[344,347] Iron is an essential nutrient for most bacteria because it plays critical roles in numerous metabolic processes.[348] *V. anguillarum* contains the pJM1 plasmid, which encodes most genes necessary for iron-anguibactin siderophore transport and biosynthesis.[349] Among these are transport proteins FatA and FatB, which are encoded by genes *fatA* and *fatB* that are transcribed together as a polycistronic mRNA. An asRNA termed RNAα, which is encoded within *fatB*, is expressed in response to increasing iron levels. It binds to the *fatB* portion of the *fatA-fatB* mRNA and represses the translation of both genes (Figure D.2C). Iron further stabilizes the binding of RNAα to *fatA-fatB* mRNA, leading to its degradation. This system helps to reduce the fitness cost associated with metabolic burden by synthesizing iron siderophores only when confronted with iron-poor conditions, thereby allowing the bacterium to optimize its resources to outcompete other bacteria.

Another example of improving metabolic fitness through regulating bacterial nutrient acquisition is the regulation of propanediol catabolism in *L. monocytogenes*.[350] Propanediol is a byproduct of the fermentation of rhamnose and fucose, and is often produced by commensal bacteria in host intestines.[351,352] Propanediol fermentation is facilitated through a coenzyme B12-dependent process and can support bacterial growth by providing ATP.[353] Some studies suggest that propanediol catabolism gives bacteria a competitive advantage, with mutations in related genes resulting in a virulence defect.[354] Within *L. monocytogenes*, the presence of propanediol activates the transcription factor PocR, which controls the expression of propanediol catabolism genes that require vitamin B_{12} as a cofactor. On the opposite strand of *pocR* gene, there is a vitamin B_{12} riboswitch-regulated asRNA, AspocR (Figure D.2D). When vitamin B_{12} is absent, a full-length version of AspocR is transcribed, which inhibits *pocR* expression. When vitamin B_{12}

is bound to the riboswitch, AspocR transcript ends prematurely, and hence cannot inhibit *pocR*. This leads to the production of PocR, which promotes the expression of propanediol catabolism genes. Thus, the riboswitch-regulated asRNA allows the expression of propanediol fermentation genes only when both propanediol and B_{12} are present, thereby reducing the fitness cost associated with unnecessary metabolic burden.

D.4 Conclusions

asRNAs are ubiquitous in bacteria and are involved in a multitude of pathogenesis-related mechanisms. The wide range of asRNA functions span the control of intra- and extra-chromosomal DNA, as well as adaption strategies to improve persistence under changing environments. Some asRNAs are only found in specific bacterial species, while others are found across bacteria. Because asRNAs play important roles in modulating the fitness of pathogenic bacteria, current research is focused not only on identifying new asRNAs, but also to use them to our advantage by developing novel asRNA-based therapeutics. For instance, bacterial antibiotic resistance genes can be targeted with synthetic asRNAs, resulting in antibiotic sensitive bacteria.[355,356] Other possible applications include using asRNA to silence bacterial metabolism or ribosomal protein coding genes (successfully shown in *E. coli*) and protection from bacteriophages in the production of live mucosal vaccines.[357–359] Applications of these techniques beyond *in vitro* studies have been limited due to difficulties in delivering asRNAs to the site of infection.[289,360] As these impediments are addressed, the use of asRNAs in therapeutics will likely expand and contribute to the understanding of the rich landscape of bacterial control systems.

Appendix E The GRE in Public Health Admissions: Barriers, Waivers, and Moving Forward

This appendix is a published work:

Millar JA. (2020). The GRE in public health admissions: Barriers, waivers, and moving forward. Front Public Health. 8:609599.

E.1 Abstract

In the wake of COVID-19, there is an urgent need for a diverse public health work force to address problems presented or exacerbated by the global pandemic. Educational programs that create our work force both train and shape the makeup of access through graduate applications. The Graduate Record Exam has a number of standing issues, with additional barriers created by the pandemic. We trace the GRE waiver movement over several years, focusing on the gradual adoption in CEPH accredited programs and the rapid expansion of temporary waivers as a response to testing access. Going forward, we need to consider gaps in waivers during the pandemic and how this data can be used to shape our future use of the GRE.

E.2 Introduction

As we move forward in the profession of public health, many of the problems presented or exacerbated by the COVID-19 global pandemic may require new solutions and a diversity of

thought and approaches to problem solving. The future of our public health workforce is shaped by decisions from our educational programs that decide who can access public health training and who designs the trainings. It is important that we consider how components of our graduate school applications currently shape our workforce and the possible barriers we create by the inclusion of testing metrics, such as the Graduate Record Exam (GRE). As the pandemic has unfolded, I have been active in collecting and sharing data on issues with the use of the GRE in public health admissions and I believe now is the time to re-think its problematic role in our public health workforce.

E.3 Issues with the GRE

Over the last decade, issues with how the GRE increases barriers to graduate education have been more widely discussed. The GRE may not predict academic success, with correlation between GRE and academic success appearing to be weak at best.[361] Boston University School of Public Health found no significant difference in mean GRE component scores for achieving >3.0 GPA in 1st year MPH students.[362] The Association of Schools and Programs of Public Health also found no correlation between GRE scores and final GPA after public health degree completion at several of its member schools.[363] Colorado SPH found GRE scores to be a weak predictor of degree completion, with other variables such as undergraduate GPA better predictors of success.[364] And University of Minnesota conducted a randomized assessment, finding GRE score didn't substantially influence admissions decisions.[365] Because of the financial burden and gender & racial/ethnic biases within the test, use of the GRE in public health admissions may create barriers for underrepresented groups. One of the direct barriers the GRE creates is a financial burden, with testing costing $205, and $27 per school submission. Another issue is the

impact on diversity and inclusion efforts. Variation in scores by race and gender has been reported, with women and members of underrepresented racial and ethnic minority groups scoring lower than white and Asian men.[366] Given that the GRE is not a convincing predictor of graduate school success, these barriers to entry are unnecessary.

The inequalities in testing have been increased during the COVID-19 pandemic. In March 2020 as lockdowns began and testing centers closed, ETS rolled out a solution to testing access: the GRE at Home. This version allowed testing to continue online, but came with a number of hurdles.[367] The requirement of a desktop or laptop and stable internet connection to take the GRE at Home are particularly problematic, given the digital divide that has become more consequential during the pandemic. A 2019 Pew Research Center survey found one in four American adults lack access to high-speed internet. This increases to half for adults with an annual income <$30,000 in major US cities.[368] In another study looking at undergrads at a large Midwestern university, although 98% of students had access to laptops, 20% of students still had difficulty accessing necessary education technology.[369] These technology barriers create a further divide to accessing the test and shut out many students.[367]

E.4 GRE waiver movement

Recently, the practice of waiving the GRE in graduate applications has spread. In 2016, the American Astronomical Society recommended the elimination of the GRE due to the test's poor predictions of success, correlations with gender, race, and socioeconomic status, and financial burden.[370] In November 2017 Joshua Hall director of the Biological & Biomedical Science Program at University of North Carolina at Chapel Hill, created a list of Bio/Biomedical programs that waive the GRE requirement.[371] By the end of 2018, almost half of all top 50

ranked molecular biology programs had waived this requirement, with the practice spreading to more STEM disciplines.[372] In 2019, some of the first high-ranked public health programs started to waive the GRE, including Boston University SPH and University of Colorado SPH.[362,364] In October 2019, a public health GRE waiver list of degrees/concentrations was created by Jess Millar, an Epidemiology MPH student at University of Michigan.[373] At the time of its creation, 48 CEPH accredited programs (one in four) had at least one GRE waiver.

As COVID-19 started to spread in the United States and lockdowns were initiated, public health programs began to consider the possibility of temporarily waiving the GRE in light of barriers to the GRE at Home. By the beginning of April, Rutgers allowed temporary waivers for Fall 2020.[374] By the end of May, at least 9 CEPH accredited programs participated in the temporary waiver, with Emory extending their waivers to Fall 2021.[375,376] The public health GRE waiver list increased 68% during its first 7 months, going from 145 to 243 entries by May 2020. By the time the SOPHAS application opened in August, the list had increased another 350% to include 880 entries.[377] As of September 20th, 2020, the list contains 1,201 entries from 150 CEPH programs (just over three quarters). 560 of the entries are for concentrations/ degrees with a permanent GRE waiver, while 641 are temporary for COVID-19 (https://www.frontiersin.org/articles/10.3389/fpubh.2020.609599/full#supplementary-material).

E.5 Public health GRE waiver coverage not evenly spread

The coverage of GRE waivers in public health programs is not equal, with very few programs allowing a blanket waiver to all graduate degrees. Among top 50 public health programs ranked by US World News, only 15 have a waiver for all degrees (permanent or temporary) as of September 20, 2020. Inclusion of a waiver also varies by concentration, with

some programs including permanent waivers for specific degrees or concentrations, and temporary or no waivers for others. Seventy percentage of degrees in CEPH accredited programs currently offer at least a temporary GRE waiver, but that percent drops as low as 59% for admissions to biostatistics-specific degree programs (Figure E.1A). These numbers drop to 33 and 23%, respectively, when only counting permanent waivers. The divide in GRE waivers is more apparent between masters and doctoral degrees. Among CEPH accredited programs, only 50% of degrees offer at least a temporary GRE waiver offered for doctoral degrees and 16% offer a permanent waiver (Figure E.1B). Most doctoral programs require doctoral interviews for admission. These can and have been used by other disciplines—such as the aforementioned STEM programs–to devise interview questions to help identity characteristics found in successful doctoral researchers and rely less on the GRE.[378]

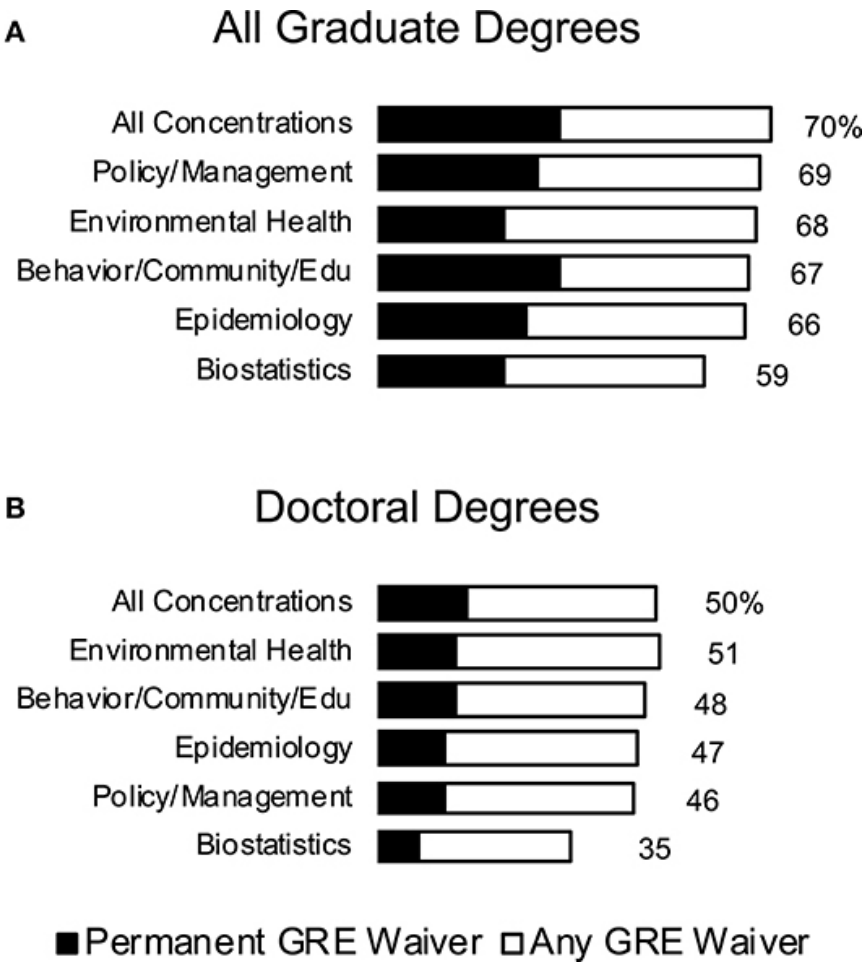

Figure E.1 GRE Waiver Snapshot. Percent of degrees at CEPH accredited public health programs at U.S. universities that have a GRE waiver. All graduate degrees are shown in (A), while only doctoral degrees are shown in (B). Permanent GRE waivers are shown super imposed (black) over all GRE waivers (white). Percentage of degrees with any GRE waiver explicitly stated at the right of each bar. Not all degrees and concentrations were offered at each program.

E.6 Conclusions

The pandemic has made inequities in access to education more visible, through the digital divide, financial concerns, and resulting conversations of barriers for minority groups. As we make our way into the 2020–2021 academic year, public health programs that have not done so may want to consider instigating or expanding temporary GRE waivers to more degrees and

concentrations. Moving beyond the pandemic, there is a great deal more hesitancy on allowing more permanent GRE waivers. Several programs, such as Cornell, University of Iowa, and Ohio State, converted their temporary MPH GRE waivers to permanent.[379,380] University of Washington took it a step further, and converted their temporary GRE waivers to permanent for both masters and doctoral degrees.[381]

Few studies on the GRE in public health programs have been conducted, but that is beginning to change. Boston University and University of North Carolina at Chapel Hill are currently conducting three-year studies to look at the impact of removing the GRE requirement on diversity and student success.[362,382] Several other programs are currently conducting 1-year pilot studies on the GRE waiver effect.[383,384] Temporary waivers are an opportunity for public health programs to test the relevance of GRE scores in the application process and their prediction of student success. We have the opportunity to test how removing a barrier to public health education will affect the professionals we create and I hope we take it.

Bibliography

1. Daley DJ, Gani J. *Epidemic Modeling: An Introduction*. Cambridge University Press; 2005.

2. Niu Y, Li Z, Meng L, et al. The collaboration between infectious disease modeling and public health decision-making based on the COVID-19. *Journal of Safety Science and Resilience*. 2021;2(2):69-76.

3. Olesen SW, Trabert E. Infectious Disease Modeling: Recommendations for Public Health Decision-Makers. *Disaster med public health prep*. Published online June 2, 2022:1-3.

4. Kretzschmar M. Disease modeling for public health: added value, challenges, and institutional constraints. *J Public Health Pol*. 2020;41(1):39-51.

5. Meerschaert MM. *Mathematical Modeling*. Academic Press; 2013.

6. Siettos CI, Russo L. Mathematical modeling of infectious disease dynamics. *Virulence*. 2013;4(4):295-306.

7. Anderson RM, May RM. *Infectious Diseases of Humans: Dynamics and Control*. OUP Oxford; 1992.

8. Moran KR, Fairchild G, Generous N, et al. Epidemic Forecasting is Messier Than Weather Forecasting: The Role of Human Behavior and Internet Data Streams in Epidemic Forecast. *The Journal of Infectious Diseases*. 2016;214(suppl_4):S404-S408.

9. Wei Y, Sha F, Zhao Y, Jiang Q, Hao Y, Chen F. Better modelling of infectious diseases: lessons from covid-19 in China. *BMJ*. 2021;375:n2365.

10. Savill NJ, Shaw DJ, Deardon R, et al. Effect of data quality on estimates of farm infectiousness trends in the UK 2001 foot-and-mouth disease epidemic. *Journal of The Royal Society Interface*. 2007;4(13):235-241.

11. Snyder J. Data Cleansing: An Omission from Data Analytics Coursework. *Information Systems Education Journal*. 2019;17(6):22.

12. Kim AY, Hardin J. "Playing the Whole Game": A Data Collection and Analysis Exercise With Google Calendar. *Journal of Statistics and Data Science Education*. 2021;29(sup1):S51-S60.

13. Gentemann CL, Holdgraf C, Abernathey R, et al. Science Storms the Cloud. *AGU Advances*. 2021;2(2):e2020AV000354.

14. Cai L, Zhu Y. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Science Journal*. 2015;14(0):2.

15. Dror IE. Cognitive and Human Factors in Expert Decision Making: Six Fallacies and the Eight Sources of Bias. *Anal Chem*. 2020;92(12):7998-8004.

16. Broeck JV den, Cunningham SA, Eeckels R, Herbst K. Data Cleaning: Detecting, Diagnosing, and Editing Data Abnormalities. *PLOS Medicine*. 2005;2(10):e267.

17. Gudivada V, Apon A, Ding J. Data Quality Considerations for Big Data and Machine Learning: Going Beyond Data Cleaning and Transformations. *International Journal on Advances in Software*. 2017;10:1-20.

18. Kerr K, Norris T. The Development of a Healthcare Data Quality Framework and Strategy. In: ; 2004:218-233.

19. World Health Organization. Tuberculosis fact sheet. World Health Organization.

20. Flynn JL, Chan J, Lin PL. Macrophages and control of granulomatous inflammation in tuberculosis. *Mucosal Immunology*. 2011;4(3):271-278.

21. Flynn JL, Gideon HP, Mattila JT, Lin PL. Immunology studies in non-human primate models of tuberculosis. *Immunological Reviews*. 2015;264(1):60-73.

22. Ehlers S, Schaible UE. The granuloma in tuberculosis: Dynamics of a host–pathogen collusion. *Frontiers in Immunology*. 2013;3:411.

23. Kumar P. IFNγ-producing CD4+ T lymphocytes: the double-edged swords in tuberculosis. *Clinical and Translational Medicine*. 2017;6(1):e21.

24. O'Garra A, Redford PS, McNab FW, Bloom CI, Wilkinson RJ, Berry MPR. The Immune Response in Tuberculosis. *Annual Review of Immunology*. 2013;31(1):475-527.

25. Russell DG. Who puts the tubercle in tuberculosis? *Nature Reviews Microbiology*. 2007;5(1):39-47.

26. Chen CY, Huang D, Wang RC, et al. A critical role for CD8 T cells in a nonhuman primate model of tuberculosis. Bishai W, ed. *PLoS Pathogens*. 2009;5(4):e1000392.

27. Lin PL, Rutledge T, Green AM, et al. CD4 T cell depletion exacerbates acute *Mycobacterium tuberculosis* while reactivation of latent infection is dependent on severity of tissue depletion in cynomolgus macaques. *AIDS Research and Human Retroviruses*. 2012;28(12):1693-1702.

28. Rossouw M, Nel HJ, Cooke GS, van Helden PD, Hoal EG. Association between tuberculosis and a polymorphic NFκB binding site in the interferon γ gene. *The Lancet*. 2003;361(9372):1871-1872.

29. Yao S, Huang D, Chen CY, Halliday L, Wang RC, Chen ZW. CD4+ T cells contain early extrapulmonary tuberculosis (TB) dissemination and rapid TB progression and sustain multieffector functions of CD8+ T and CD3- lymphocytes: mechanisms of CD4+ T cell immunity. *The Journal of Immunology*. 2014;192(5):2120-2132.

30. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. Helper T Cells and Lymphocyte Activation. In: *Molecular Biology of the Cell*. 4th ed. Garland Science; 2002.

31. Chan ED, Chan J, Schluger NW. What is the Role of Nitric Oxide in Murine and Human Host Defense against Tuberculosis? *Am J Respir Cell Mol Biol*. 2001;25(5):606-612.

32. Rozot V, Vigano S, Mazza-Stalder J, et al. Mycobacterium tuberculosis-specific CD8+ T cells are functionally and phenotypically different between latent infection and active disease. *European Journal of Immunology*. 2013;43(6):1568-1577.

33. Harari A, Rozot V, Enders FB, et al. Dominant TNF-α+ Mycobacterium tuberculosis–specific CD4+ T cell responses discriminate between latent infection and active disease | Nature Medicine. *Nature Medicine*. 2011;17:372-376.

34. Cooper AM, Flynn JL. The protective immune response to Mycobacterium tuberculosis. *Current Opinion in Immunology*. 1995;7(4):512-516.

35. Flynn JL, Bloom BR. Role of Tl and T2 Cytokines in the Response to Mycobacterium tuberculosis. *Annals of the New York Academy of Sciences*. 1996;795(1):137-146.

36. Flynn JL, Chan J. Immunology of Tuberculosis. *Annual Review of Immunology*. 2001;19(1):93-129.

37. Maartens G, Wilkinson RJ. Tuberculosis. *The Lancet*. 2007;370(9604):2030-2043.

38. Green AM, DiFazio R, Flynn JL. IFN-γ from CD4 T Cells Is Essential for Host Survival and Enhances CD8 T Cell Function during Mycobacterium tuberculosis Infection. *The Journal of Immunology*. 2013;190(1):270-277.

39. Gideon HP, Phuah J, Myers AJ, et al. Variability in tuberculosis granuloma T cell responses exists, but a balance of pro- and anti-inflammatory cytokines is associated with sterilization. Lewinsohn DM, ed. *PLOS Pathogens*. 2015;11(1):e1004603.

40. Day CL, Kaufmann DE, Kiepiela P, et al. PD-1 expression on HIV-specific T cells is associated with T-cell exhaustion and disease progression. *Nature*. 2006;443(7109):350-354.

41. D'Souza M, Fontenot AP, Mack DG, et al. Programmed Death 1 Expression on HIV-Specific CD4+ T Cells Is Driven by Viral Replication and Associated with T Cell Dysfunction. *The Journal of Immunology*. 2007;179(3):1979-1987.

42. Onlamoon N, Rogers K, Mayne AE, et al. Soluble PD-1 rescues the proliferative response of simian immunodeficiency virus-specific CD4 and CD8 T cells during chronic infection. *Immunology*. 2008;124(2):277-293.

43. Velu V, Titanji K, Zhu B, et al. Enhancing SIV-specific immunity in vivo by PD-1 blockade. *Nature*. 2009;458(7235):206-210.

44. Wong EA, Joslyn L, Grant NL, et al. Low levels of T cell exhaustion in tuberculous lung granulomas. Bäumler AJ, ed. *Infection and Immunity*. 2018;86(9):e00426-18.

45. Athman JJ, Sande OJ, Groft SG, et al. *Mycobacterium tuberculosis* Membrane Vesicles Inhibit T Cell Activation. *The Journal of Immunology*. 2017;198(5):2028-2037.

46. Mahon RN, Sande OJ, Rojas RE, Levine AD, Harding CV, Henry Boom W. *Mycobacterium tuberculosis* ManLAM inhibits T-cell-receptor signaling by interference with ZAP-70, Lck and LAT phosphorylation. *Cellular Immunology*. 2012;275(1-2):98-105.

47. Pancholi P, Mirza A, Bhardwaj N, Steinman RM. Sequestration from immune CD4+ T cells of mycobacteria growing in human macrophages. *Science*. 1993;260(5110):984-986.

48. Hmama Z, Gabathuler R, Jefferies WA, Jong G de, Reiner NE. Attenuation of HLA-DR Expression by Mononuclear Phagocytes Infected with Mycobacterium tuberculosis Is Related to Intracellular Sequestration of Immature Class II Heterodimers. *The Journal of Immunology*. 1998;161(9):4882-4893.

49. Noss EH, Harding CV, Boom WH. Mycobacterium tuberculosis Inhibits MHC Class II Antigen Processing in Murine Bone Marrow Macrophages. *Cellular Immunology*. 2000;201(1):63-74.

50. Segovia-Juarez JL, Ganguli S, Kirschner D. Identifying control mechanisms of granuloma formation during *M. tuberculosis* infection using an agent-based model. *Journal of Theoretical Biology*. 2004;231(3):357-376.

51. Ray JCJ, Wang J, Chan J, Kirschner DE. The timing of TNF and IFN-γ signaling affects macrophage activation strategies during Mycobacterium tuberculosis infection. *Journal of Theoretical Biology*. 2008;252(1):24-38.

52. Joslyn LR, Renardy M, Weissman C, et al. Temporal and Spatial Analyses of TB Granulomas to Predict Long-Term Outcomes. In: Vodovotz Y, An G, eds. *Complex Systems and Computational Biology Approaches to Acute Inflammation: A Framework for Model-Based Precision Medicine*. Springer International Publishing; 2021:273-291.

53. Mattila JT, Maiello P, Sun T, Via LE, Flynn JL. Granzyme B-expressing neutrophils correlate with bacterial load in granulomas from Mycobacterium tuberculosis-infected cynomolgus macaques. *Cellular Microbiology*. 2015;17(8):1085-1097.

54. Co DO, Hogan LH, Karman J, et al. Interactions between T Cells Responding to Concurrent Mycobacterial and Influenza Infections. *The Journal of Immunology*. 2006;177(12):8456-8465.

55. Egen JG, Rothfuchs AG, Feng CG, Winter N, Sher A, Germain RN. Macrophage and T Cell Dynamics during the Development and Disintegration of Mycobacterial Granulomas. *Immunity*. 2008;28(2):271-284.

56. Egen JG, Rothfuchs AG, Feng CG, Horwitz MA, Sher A, Germain RN. Intravital Imaging Reveals Limited Antigen Presentation and T Cell Effector Function in Mycobacterial Granulomas. *Immunity*. 2011;34(5):807-819.

57. Ogongo P, Porterfield JZ, Leslie A. Lung Tissue Resident Memory T-Cells in the Immune Response to Mycobacterium tuberculosis. *Frontiers in Immunology*. 2019;10:992.

58. Ariotti S, Haanen JB, Schumacher TN. Chapter 8 - Behavior and Function of Tissue-Resident Memory T cells. In: Melief CJM, ed. *Advances in Immunology*. Vol 114. Synthetic Vaccines. Academic Press; 2012:203-216.

59. Marakalala MJ, Raju RM, Sharma K, et al. Inflammatory signaling in human tuberculosis granulomas is spatially organized. *Nat Med*. 2016;22(5):531-538.

60. Lin PL, Ford CB, Coleman MT, et al. Sterilization of granulomas is common in active and latent tuberculosis despite within-host variability in bacterial killing. *Nat Med*. 2014;20(1):75-79.

61. Fallahi-Sichani M, El-Kebir M, Marino S, Kirschner DE, Linderman JJ. Multiscale Computational Modeling Reveals a Critical Role for TNF-α Receptor 1 Dynamics in Tuberculosis Granuloma Formation. *The Journal of Immunology*. 2011;186:3472-3483.

62. Cilfone NA, Perry CR, Kirschner DE, Linderman JJ. Multi-Scale Modeling Predicts a Balance of Tumor Necrosis Factor-α and Interleukin-10 Controls the Granuloma Environment during Mycobacterium tuberculosis Infection. *PLOS ONE*. 2013;8(7):e68680.

63. Pienaar E, Matern WM, Linderman JJ, Bader JS, Kirschner DE. Multiscale Model of Mycobacterium tuberculosis Infection Maps Metabolite and Gene Perturbations to Granuloma Sterilization Predictions. *Infection and Immunity*. 2016;84(5):1650-1669.

64. Chang ST, Linderman JJ, Kirschner DE. Multiple mechanisms allow *Mycobacterium tuberculosis* to continuously inhibit MHC class II-mediated antigen presentation by macrophages. *Proceedings of the National Academy of Sciences*. 2005;102(12):4530-4535.

65. Cilfone NA, Kirschner DE, Linderman JJ. Strategies for Efficient Numerical Implementation of Hybrid Multi-scale Agent-Based Models to Describe Biological Systems. *Cel Mol Bioeng*. 2015;8(1):119-136.

66. Marino S, Hult C, Wolberg P, Linderman JJ, Kirschner DE. The Role of Dimensionality in Understanding Granuloma Formation. *Computation*. 2018;6(4):58.

67. Kaufmann SHE. Cell-mediated immunity: Dealing a direct blow to pathogens. *Current Biology*. 1999;9(3):R97-R99.

68. Kirschner DE, Hunt CA, Marino S, Fallahi-Sichani M, Linderman JJ. Tuneable resolution as a systems biology approach for multi-scale, multi-compartment computational models. *WIREs Systems Biology and Medicine*. 2014;6(4):289-309.

69. Woodbury JL, Barrow WW. Radiolabelling of Mycobacterium avium Oligosaccharide Determinant and Use in Macrophage Studies. *Microbiology*. 1989;135(7):1875-1884.

70. Linderman JJ, Riggs T, Pande M, Miller M, Marino S, Kirschner DE. Characterizing the Dynamics of CD4+ T Cell Priming within a Lymph Node. *The Journal of Immunology*. 2010;184(6):2873-2885.

71. Santana MA, Esquivel-Guadarrama F. Cell Biology of T Cell Activation and Differentiation. *International Review of Cytology*. 2006;250:217-274.

72. Kabilan L, Andersson G, Lolli F, Ekre H peter, Olsson T, Troye-blomberg M. Detection of intracellular expression and secretion of interferon-γ at the single-cell level after activation of human T cells with tetanus toxoid in vitro. *European Journal of Immunology*. 1990;20(5):1085-1089.

73. Farrar MA, Schreiber RD. The molecular cell biology of interferon-gamma and its receptor. *Annu Rev Immunol*. 1993;11:571-611.

74. R&D Systems. Unit Conversion Table. R&D Systems.

75. Hirschfield GR, McNeil M, Brennan PJ. Peptidoglycan-associated polypeptides of Mycobacterium tuberculosis. *Journal of Bacteriology*. 1990;172(2):1005-1013.

76. Young DB. Structure of mycobacterial antigens. *British Medical Bulletin*. 1988;44(3):562-583.

77. Beste DJV, Peters J, Hooper T, Avignone-Rossa C, Bushell ME, McFadden J. Compiling a Molecular Inventory for Mycobacterium bovis BCG at Two Growth Rates: Evidence for Growth Rate-Mediated Regulation of Ribosome Biosynthesis and Lipid Metabolism. *Journal of Bacteriology*. 2005;187(5):1677-1684.

78. Marino S, Myers A, Flynn JL, Kirschner DE. TNF and IL-10 are major factors in modulation of the phagocytic cell environment in lung and lymph node in tuberculosis: A next-generation two-compartmental model. *Journal of Theoretical Biology*. 2010;265(4):586-598.

79. Marino S, El-Kebir M, Kirschner D. A hybrid multi-compartment model of granuloma formation and T cell priming in Tuberculosis. *Journal of Theoretical Biology*. 2011;280(1):50-62.

80. Gong C, Mattila JT, Miller M, Flynn JL, Linderman JJ, Kirschner D. Predicting lymph node output efficiency using systems biology. *Journal of Theoretical Biology*. 2013;335:169-184.

81. Marino S, Hogue IB, Ray CJ, Kirschner DE. A methodology for performing global uncertainty and sensitivity analysis in systems biology. *Journal of Theoretical Biology*. 2008;254(1):178-196.

82. Blower SM, Dowlatabadi H. Sensitivity and Uncertainty Analysis of Complex Models of Disease Transmission: An HIV Model, as an Example. *International Statistical Review / Revue Internationale de Statistique*. 1994;62(2):229-243.

83. Marino S, Gideon HP, Gong C, et al. Computational and Empirical Studies Predict Mycobacterium tuberculosis-Specific T Cells as a Biomarker for Infection Outcome. *PLOS Computational Biology*. 2016;12(4):e1004804.

84. VanderWeele TJ, Mathur MB. Some desirable properties of the Bonferroni correction: is the Bonferroni correction really so bad? *American Journal of Epidemiology*. 2019;188(3):617-618.

85. Renardy M, Wessler T, Blemker S, Linderman J, Peirce S, Kirschner D. Data-Driven Model Validation Across Dimensions. *Bull Math Biol*. 2019;81(6):1853-1866.

86. Wessler T, Joslyn LR, Borish HJ, et al. A computational model tracks whole-lung Mycobacterium tuberculosis infection and predicts factors that inhibit dissemination. *PLOS Computational Biology*. 2020;16(5):e1007280.

87. Cicchese JM, Dartois V, Kirschner DE, Linderman JJ. Both Pharmacokinetic Variability and Granuloma Heterogeneity Impact the Ability of the First-Line Antibiotics to Sterilize Tuberculosis Granulomas. *Frontiers in Pharmacology*. 2020;11:333.

88. Patankar YR, Sutiwisesak R, Boyce S, et al. Limited recognition of Mycobacterium tuberculosis-infected macrophages by polyclonal CD4 and CD8 T cells from the lungs of infected mice. *Mucosal Immunol*. 2020;13(1):140-148.

89. Pai RK, Askew D, Boom WH, Harding CV. Regulation of Class II MHC Expression in APCs: Roles of Types I, III, and IV Class II Transactivator. *The Journal of Immunology*. 2002;169(3):1326-1333.

90. Cullell-Young M, Barrachina M, López-López C, et al. From transcription to cell surface expression, the induction of MHC class II I-Aα by interferon-γ in macrophages is regulated at different levels. *Immunogenetics*. 2001;53(2):136-144.

91. Buus S, Werdelin O. Large, but Not Small, Antigens Require Time-and Temperature-Dependent Processing in Accessory Cells Before They Can Be Recognized by T Cells. *Acta Pathologica Microbiologica Scandinavica Series C: Immunology*. 1986;94C(1-6):17-24.

92. Pienaar E, Dartois V, Linderman JJ, Kirschner DE. In silico evaluation and exploration of antibiotic tuberculosis treatment regimens. *BMC Systems Biology*. 2015;9(1):79.

93. Riggs T, Walts A, Perry N, et al. A comparison of random vs. chemotaxis-driven contacts of T cells with dendritic cells during repertoire scanning. *Journal of Theoretical Biology*. 2008;250(4):732-751.

94. Hilhorst M, Shirai T, Berry G, Goronzy JJ, Weyand CM. T Cell–Macrophage Interactions and Granuloma Formation in Vasculitis. *Frontiers in Immunology*. 2014;5:432.

95. Kauffman KD, Sallin MA, Sakai S, et al. Defective positioning in granulomas but not lung-homing limits CD4 T-cell interactions with Mycobacterium tuberculosis-infected macrophages in rhesus macaques. *Mucosal Immunol*. 2018;11(2):462-473.

96. DiazGranados CA, Zimmer SM, Mitchel K, Jernigan JA. Comparison of Mortality Associated with Vancomycin-Resistant and Vancomycin-Susceptible Enterococcal Bloodstream Infections: A Meta-analysis. *Clinical Infectious Diseases*. 2005;41(3):327-333.

97. Edmond MB, Ober JF, Dawson JD, Weinbaum DL, Wenzel RP. Vancomycin-Resistant Enterococcal Bacteremia: Natural History and Attributable Mortality. *Clinical Infectious Diseases*. 1996;23(6):1234-1239.

98. Hayden MK. Insights into the Epidemiology and Control of Infection with Vancomycin-Resistant Enterococci. *CLIN INFECT DIS*. 2000;31(4):1058-1065.

99. Chavers LS, Moser SA, Benjamin WH, et al. Vancomycin-resistant enterococci: 15 years and counting. *Journal of Hospital Infection*. 2003;53(3):159-171.

100. Arias CA, Murray BE. The rise of the Enterococcus: beyond vancomycin resistance. *Nature Reviews Microbiology*. 2012;10(4):266-278.

101. Chiang HY, Perencevich EN, Nair R, et al. Incidence and Outcomes Associated With Infections Caused by Vancomycin-Resistant Enterococci in the United States: Systematic Literature Review and Meta-Analysis. *Infect Control Hosp Epidemiol*. 2017;38(2):203-215.

102. Hayakawa K, Marchaim D, Palla M, et al. Epidemiology of Vancomycin-Resistant Enterococcus faecalis: a Case-Case-Control Study. *Antimicrob Agents Chemother*. 2013;57(1):49-55.

103. Ghanem G, Hachem R, Jiang Y, Chemaly RF, Raad I. Outcomes for and Risk Factors Associated With Vancomycin-Resistant *Enterococcus faecalis* and Vancomycin-Resistant *Enterococcus faecium* Bacteremia in Cancer Patients. *Infection Control & Hospital Epidemiology*. 2007;28(9):1054-1059.

104. Hayakawa K, Marchaim D, Martin ET, et al. Comparison of the Clinical Characteristics and Outcomes Associated with Vancomycin-Resistant Enterococcus faecalis and Vancomycin-Resistant E. faecium Bacteremia. *Antimicrobial Agents and Chemotherapy*. 2012;56(5):2452-2458.

105. Centers for Disease Control and Prevention. Nosocomial Enterococci Resistant to Vancomycin—United States, 1989-1993. *Morbidity and Mortality Weekly Report*. 1993;42(30):597-599.

106. Centers for Disease Control and Prevention. Recommendations for Preventing the Spread of Vancomycin Resistance Recommendations of the Hospital Infection Control Practices Advisory Committee (HICPAC). *Morbidity and Mortality Weekly Report*. 1995;44(RR12):1-13.

107. Werner G, Coque TM, Hammerum AM, et al. Emergence and spread of vancomycin resistance among enterococci in Europe. *Eurosurveillance*. 2008;13(47):19046.

108. Deshpande LM, Fritsche TR, Moet GJ, Biedenbach DJ, Jones RN. Antimicrobial resistance and molecular epidemiology of vancomycin-resistant enterococci from North America and Europe: a report from the SENTRY antimicrobial surveillance program. *Diagnostic Microbiology and Infectious Disease*. 2007;58(2):163-170.

109. Reik R, Tenover FC, Klein E, McDonald LC. The burden of vancomycin-resistant enterococcal infections in US hospitals, 2003 to 2004. *Diagnostic Microbiology and Infectious Disease*. 2008;62(1):81-85.

110. Bonten MJM, Slaughter S, Ambergen AW, et al. The Role of "Colonization Pressure" in the Spread of Vancomycin-Resistant Enterococci: An Important Infection Control Variable. *Arch Intern Med*. 1998;158(10):1127-1132.

111. Lebreton F, van Schaik W, Manson McGuire A, et al. Emergence of Epidemic Multidrug-Resistant Enterococcus faecium from Animal and Commensal Strains. McDaniel LS, ed. *mBio*. 2013;4(4):e00534-13.

112. Lebreton F, Manson AL, Saavedra JT, Straub TJ, Earl AM, Gilmore MS. Tracing the Enterococci from Paleozoic Origins to the Hospital. *Cell*. 2017;169(5):849-861.e13.

113. Huskins WC, Huckabee CM, O'Grady NP, et al. Intervention to Reduce Transmission of Resistant Bacteria in Intensive Care. *The New England Journal of Medicine*. 2011;364(15):1407-1418.

114. Buick S, Joffe AM, Taylor G, Conly J. A Consensus Development Conference Model for Establishing Health Policy for Surveillance and Screening of Antimicrobial-Resistant Organisms. *Clinical Infectious Diseases*. 2015;60(7):1095-1101.

115. Montecalvo MA, Jarvis WR, Uman J, et al. Infection-Control Measures Reduce Transmission of Vancomycin-Resistant Enterococci in an Endemic Setting. *Ann Intern Med*. 1999;131(4):269-272.

116. Siddiqui AH, Harris AD, Hebden J, Wilson PD, Morris JG, Roghmann MC. The effect of active surveillance for vancomycin-resistant enterococci in high-risk units on vancomycin-resistant enterococci incidence hospital-wide. *American Journal of Infection Control*. 2002;30(1):40-43.

117. Rubin LG, Tucci V, Cercenado E, Eliopoulos G, Isenberg HD. Vancomycin-Resistant Enterococcus faecium in Hospitalized Children. *Infection Control & Hospital Epidemiology*. 1992;13(12):700-705.

118. Handwerger S, Raucher B, Altarac D, et al. Nosocomial Outbreak Due to Enterococcus faecium Highly Resistant to Vancomycin, Penicillin, and Gentamicin. *Clinical Infectious Diseases*. 1993;16(6):750-755.

119. Rhinehart E, Smith NE, Wennersten C, et al. Rapid dissemination of beta-lactamase-producing, aminoglycoside-resistant Enterococcus faecalis among patients and staff on an infant-toddler surgical ward. *New England Journal of Medicine*. 1990;323(26):1814-1818.

120. Muto CA, Jernigan JA, Ostrowsky BE, et al. SHEA Guideline for Preventing Nosocomial Transmission of Multidrug-Resistant Strains of *Staphylococcus aureus* and *Enterococcus*. *Infection Control & Hospital Epidemiology*. 2003;24(5):362-386.

121. Slaughter S, Hayden MK, Nathan C, et al. A Comparison of the Effect of Universal Use of Gloves and Gowns with That of Glove Use Alone on Acquisition of Vancomycin-Resistant Enterococci in a Medical Intensive Care Unit. *Ann Intern Med*. 1996;125(6):448-456.

122. Wells CL, Juni BA, Cameron SB, et al. Stool Carriage, Clinical Isolation, and Mortality During an Outbreak of Vancomycin-Resistant Enterococci in Hospitalized Medical and/or Surgical Patients. *Clinical Infectious Diseases*. 1995;21(1):45-50.

123. Zuckerman RA, Steele L, Venezia RA, Tobin EH. Undetected Vancomycin-Resistant *Enterococcus* in Surgical Intensive Care Unit Patients. *Infect Control Hosp Epidemiol*. 1999;20(10):685-686.

124. Ostrowsky BE, Venkataraman L, D'Agata EMC, Gold HS, DeGirolami PC, Samore MH. Vancomycin-Resistant Enterococci in Intensive Care Units: High Frequency of Stool Carriage During a Non–Outbreak Period. *Archives of Internal Medicine*. 1999;159(13):1467-1472.

125. Ostrowsky BE, Trick WE, Sohn AH, et al. Control of Vancomycin-Resistant Enterococcus in Health Care Facilities in a Region. *The New England Journal of Medicine*. 2001;344(19):1427-1433.

126. Calfee DP, Farr BM. Infection Control and Cost Control in the Era of Managed Care. *Infect Control Hosp Epidemiol*. 2002;23(7):407-410.

127. Karanfil LV, Murphy M, Josephson A, et al. A Cluster of Vancomycin-Resistant Enterococcus faecium in an Intensive Care Unit. *Infection Control and Hospital Epidemiology*. 1992;13(4):195-200.

128. Montecalvo MA, Horowitz H, Gedris C, et al. Outbreak of vancomycin-, ampicillin-, and aminoglycoside-resistant Enterococcus faecium bacteremia in an adult oncology unit. *Antimicrob Agents Chemother*. 1994;38(6):1363-1367.

129. Dembry LM, Uzokwe K, Zervos MJ. Control of Endemic Glycopeptide-Resistant Enterococci. *Infection Control and Hospital Epidemiology*. 1996;17(5):286-292.

130. Rupp ME, Marion N, Fey PD, et al. Outbreak of Vancomycin-Resistant *Enterococcus faecium* in a Neonatal Intensive Care Unit. *Infection Control & Hospital Epidemiology*. 2001;22(5):301-303.

131. Malik RK, Montecalvo MA, Reale MR, et al. Epidemiology and control of vancomycin-resistant enterococci in a regional neonatal intensive care unit. *The Pediatric Infectious Disease Journal*. 1999;18(4):352-356.

132. Jochimsen EM, Fish L, Manning K, et al. Control of Vancomycin-Resistant Enterococci at a Community Hospital: Efficacy of Patient and Staff Cohorting. *Infect Control Hosp Epidemiol*. 1999;20(2):106-109.

133. Muto CA, Giannetta ET, Durbin LJ, Simonton BM, Farr BM. Cost-Effectiveness of Perirectal Surveillance Cultures for Controlling Vancomycin-Resistant *Enterococcus*. *Infection Control & Hospital Epidemiology*. 2002;23(8):429-435.

134. Armstrong-Evans M, Litt M, McArthur MA, et al. Control of Transmission of Vancomycin-Resistant Enterococcus faecium in a Long-Term–Care Facility. *Infection Control & Hospital Epidemiology*. 1999;20(5):312-317.

135. Byers KE, Anglim AM, Anneski CJ, et al. A Hospital Epidemic of Vancomycin-Resistant Enterococcus Risk Factors and Control. *Infect Control Hosp Epidemiol*. 2001;22(3):140-147.

136. Martin P, Abou Chakra CN, Williams V, et al. Prevalence of antibiotic-resistant organisms in Canadian Hospitals. Comparison of point-prevalence survey results from 2010, 2012, and 2016. *Infect Control Hosp Epidemiol*. 2019;40(1):53-59.

137. Martischang R, Buetti N, Balmelli C, Saam M, Widmer A, Harbarth S. Nation-wide survey of screening practices to detect carriers of multi-drug resistant organisms upon admission to Swiss healthcare institutions. *Antimicrob Resist Infect Control*. 2019;8(1):37.

138. Furuno JP, McGregor JC, Harris AD, et al. Identifying Groups at High Risk for Carriage of Antibiotic-Resistant Bacteria. *Arch Intern Med*. 2006;166(5):580.

139. Ghosh A, Jiao L, Al-Mutawa F, O'Neill C, Mertz D, Hamilton Health Sciences Infection Prevention and Control Team. Value of an active surveillance policy to document clearance of meticillin-resistant Staphylococcus aureus and vancomycin-resistant enterococci amongst inpatients with prolonged admissions. *Journal of Hospital Infection*. 2014;88(4):230-233.

140. Price CS, Paule S, Noskin GA, Peterson LR. Active Surveillance Reduces the Incidence of Vancomycin-Resistant Enterococcal Bacteremia. *Clinical Infectious Diseases*. 2003;37(7):921-928.

141. Derde LPG, Cooper BS, Goossens H, et al. Interventions to reduce colonisation and transmission of antimicrobial-resistant bacteria in intensive care units: an interrupted time series study and cluster randomised trial. *The Lancet Infectious Diseases*. 2014;14(1):31-39.

142. Shimokura G, Romney M. Assessment of Compliance with Admission Screening for Vancomycin Resistant Enterococci (VRE) and Methicillin Resistant Staphylococcus aureus (MRSA) Colonization. *American Journal of Infection Control*. 2006;34(5):E25.

143. Padiglione AA, Wolfe R, Grabsch EA, et al. Risk Factors for New Detection of Vancomycin-Resistant Enterococci in Acute-Care Hospitals That Employ Strict Infection Control Procedures. *Antimicrobial Agents and Chemotherapy*. 2003;47(8):2492-2498.

144. Linfield RY, Campeau S, Injean P, et al. Practical methods for effective vancomycin-resistant enterococci (VRE) surveillance: experience in a liver transplant surgical intensive care unit. *Infect Control Hosp Epidemiol*. 2018;39(10):1178-1182.

145. Grayson M, Mahony A, Grabsch E, et al. Marked reductions in rates of vancomycin-resistant enterococci (VRE) colonization & disease associated with introduction of a routine hospital-wide bleach cleaning program. *BMC Proc*. 2011;5(S6):P24, 1753-6561-5-S6-P24.

146. Grabsch EA, Mahony AA, Cameron DRM, et al. Significant reduction in vancomycin-resistant enterococcus colonization and bacteraemia after introduction of a bleach-based cleaning–disinfection programme. *Journal of Hospital Infection*. 2012;82(4):234-242.

147. Bearman GML, Marra AR, Sessler CN, et al. A controlled trial of universal gloving versus contact precautions for preventing the transmission of multidrug-resistant organisms. *American Journal of Infection Control*. 2007;35(10):650-655.

148. Perencevich EN, Fisman DN, Lipsitch M, Harris AD, Morris, Jr. JG, Smith DL. Projected Benefits of Active Surveillance for Vancomycin-Resistant Enterococci in Intensive Care Units. *CLIN INFECT DIS*. 2004;38(8):1108-1115.

149. Mac S, Fitzpatrick T, Johnstone J, Sander B. Vancomycin-resistant enterococci (VRE) screening and isolation in the general medicine ward: a cost-effectiveness analysis. *Antimicrob Resist Infect Control*. 2019;8:168.

150. Huang SS, Rifas-Shiman SL, Pottinger JM, et al. Improving the Assessment of Vancomycin-Resistant Enterococci by Routine Screening. *J INFECT DIS*. 2007;195(3):339-346.

151. Smith RA. *Inference of Infectious Disease Dynamics from Genetic Data via Sequential Monte Carlo*. University of Michigan; 2018.

152. Jo I, Song CE, Park KG, Park YJ. Comparison of Three Chromogenic Media for Recovery of Vancomycin-Resistant Enterococci from Rectal Swab Samples. *Ann Clin Microbiol*. 2015;18(3):82-86.

153. Suwantarat N, Roberts A, Prestridge J, et al. Comparison of Five Chromogenic Media for Recovery of Vancomycin-Resistant Enterococci from Fecal Samples. *Journal of Clinical Microbiology*. 2014;52(11):4039-4042.

154. Anderson NW, Buchan BW, Young CL, et al. Multicenter Clinical Evaluation of VRESelect Agar for Identification of Vancomycin-Resistant Enterococcus faecalis and Enterococcus faecium. *Journal of Clinical Microbiology*. 2013;51(8):2758-2760.

155. Fyodorova AV, Klyasova GA. Detection of vancomycin-resistant enterococci using chromogenic selective medium. *Clinical Microbiology and Antimicrobial Chemotherapy*. 2018;20(1):55-61.

156. Cheah ALY, Cheng AC, Spelman D, Nation RL, Kong DCM, McBryde ES. Mathematical modelling of vancomycin-resistant enterococci transmission during passive surveillance and active surveillance with contact isolation highlights the need to identify and address the source of acquisition. *BMC Infect Dis*. 2018;18(1):511.

157. Escaut L, Bouam S, Frank-Soltysiak M, et al. Eradication of an outbreak of vancomycin-resistant Enterococcus (VRE): the cost of a failure in the systematic screening. *Antimicrob Resist Infect Control*. 2013;2(1):18.

158. Ortiz-Prado E, Simbaña-Rivera K, Gómez- Barreno L, et al. Clinical, molecular, and epidemiological characterization of the SARS-CoV-2 virus and the Coronavirus Disease 2019 (COVID-19), a comprehensive literature review. *Diagnostic Microbiology and Infectious Disease*. 2020;98(1):115094.

159. GitHub—nytimes/covid-19-data: An ongoing repository of data on coronavirus cases and deaths in the U.S. GitHub.

160. COVID-19 Map. Johns Hopkins Coronavirus Resource Center.

161. Wang H. Estimation of total and excess mortality due to COVID-19. Institute for Health Metrics and Evaluation.

162. Cao J, Tu WJ, Cheng W, et al. Clinical Features and Short-term Outcomes of 102 Patients with Coronavirus Disease 2019 in Wuhan, China. *Clinical Infectious Diseases*. 2020;71(15):748-755.

163. Tu WJ, Cao J, Yu L, Hu X, Liu Q. Clinicolaboratory study of 25 fatal cases of COVID-19 in Wuhan. *Intensive Care Med*. 2020;46(6):1117-1120.

164. Basu A. Estimating The Infection Fatality Rate Among Symptomatic COVID-19 Cases In The United States. *Health Affairs*. 2020;39(7):1229-1236.

165. Hutchins SS, Truman BI, Merlin TL, Redd SC. Protecting Vulnerable Populations From Pandemic Influenza in the United States: A Strategic Imperative. *Am J Public Health*. 2009;99(S2):S243-S248.

166. Celentano DD, Szklo M. *Gordis Epidemiology*. 6TH ed. Elsevier; 2018.

167. Lyu W, Wehby GL. Shelter-In-Place Orders Reduced COVID-19 Mortality And Reduced The Rate Of Growth In Hospitalizations. *Health Affairs*. 2020;39(9):1615-1623.

168. Yang J, Zheng Y, Gou X, et al. Prevalence of comorbidities and its effects in patients infected with SARS-CoV-2: a systematic review and meta-analysis. *Int J Infect Dis*. 2020;94:91-95.

169. Ji Y, Ma Z, Peppelenbosch MP, Pan Q. Potential association between COVID-19 mortality and health-care resource availability. *The Lancet Global Health*. 2020;8(4):e480.

170. Millar JA. COVID19_CFR: Repository of data and code for calculation COVID-19 county level CFRs. GitHub.

171. Nguyen-Van-Tam JS, Openshaw PJM, Hashim A, et al. Risk factors for hospitalisation and poor outcome with pandemic A/H1N1 influenza: United Kingdom first wave (May–September 2009). *Thorax*. 2010;65(7):645-651.

172. Chan-Yeung M, Xu RH. SARS: epidemiology. *Respirology*. 2003;8(s1):S9-S14.

173. Park JE, Jung S, Kim A, Park JE. MERS transmission and risk factors: a systematic review. *BMC Public Health*. 2018;18(1):574.

174. National Center for Health Statistics. U.S. Census Populations With Bridged Race Categories. Centers for Disease Control and Prevention.

175. US Census Bureau. American Community Survey 5-Year Data (2009-2018). US Census Bureau.

176. Flanagan BE, Hallisey EJ, Adams E, Lavery A. Measuring Community Vulnerability to Natural and Anthropogenic Hazards: The Centers for Disease Control and Prevention's Social Vulnerability Index. *J Environ Health*. 2018;80(10):34-36.

177. Flanagan BE, Gregory EW, Hallisey EJ, Heitgerd JL, Lewis B. A Social Vulnerability Index for Disaster Management. *Journal of Homeland Security and Emergency Management*. 2011;8(1):3.

178. Adams E. Social vulnerability and disaster-related health outcomes. Poster presented at: Esri User Conference; 2016; San Diego, CA.

179. Lavery A. Mapping mortalities following Hurricane Harvey, Harris County, TX, August–September 2017. Poster presented at: Centers for Disease Control and Prevention/Agency for Toxic Substances and Disease Registry GIS Day; 2017; Atlanta, GA.

180. Kolling J, Wilt G, Berens A, Strosnider H, Devine O. Social and environmental risk factors associated with county-level asthma emergency department visits. Poster presented at: American Public Health Association Annual Meeting; 2017; Atlanta, GA.

181. Bakkensen LA, Fox-Lent C, Read LK, Linkov I. Validating Resilience and Vulnerability Indices in the Context of Natural Disasters. *Risk Analysis*. 2017;37(5):982-1004.

182. Schulte F, Lucas E, Rau J, Szabo L, Hancock J. Millions Of Older Americans Live In Counties With No ICU Beds As Pandemic Intensifies. Kaiser Health News.

183. Centers for Disease Control and Prevention. Interactive Atlas of Heart Disease and Stroke. Centers for Disease Control and Prevention.

184. Centers for Disease Control and Prevention. U.S. Diabetes Surveillance System. Centers for Disease Control and Prevention.

185. Centers for Disease Control and Prevention. CDC - BRFSS. https://www.cdc.gov/brfss/index.html.

186. National Cancer Institute. State Cancer Profiles. National Cancer Institute.

187. Keystone Strategy. Covid19-intervention-data. GitHub.

188. Galvin G. CDC Issues New Guidance for Americans As States Reopen From Coronavirus Closures. US News & World Report.

189. NYC Health. COVID-19: Data by Borough - NYC Health. NYC Health.

190. Nishiura H, Klinkenberg D, Roberts M, Heesterbeek JAP. Early Epidemiological Assessment of the Virulence of Emerging Infectious Diseases: A Case Study of an Influenza Pandemic. *PLOS ONE*. 2009;4(8):e6852.

191. Russell TW, Hellewell J, Jarvis CI, et al. Estimating the infection and case fatality ratio for coronavirus disease (COVID-19) using age-adjusted data from the outbreak on the Diamond Princess cruise ship, February 2020. *Eurosurveillance*. 2020;25(12):2000256.

192. Lewnard JA, Liu VX, Jackson ML, et al. Incidence, clinical outcomes, and transmission dynamics of severe coronavirus disease 2019 in California and Washington: prospective cohort study. *BMJ*. 2020;369:m1923.

193. Wulff JN, Ejlskov L. Multiple Imputation by Chained Equations in Praxis: Guidelines and Review. *Electronic Journal of Business Research Methods*. 2017;15(1):41-56.

194. Bursac Z, Gauss CH, Williams DK, Hosmer DW. Purposeful selection of variables in logistic regression. *Source Code for Biology and Medicine*. 2008;3(1):17.

195. Chen Y, Millar JA. Machine Learning Techniques in Cancer Prognostic Modeling and Performance Assessment. In: Matsui S, Crowley J, eds. *Frontiers of Biostatistical Methods and Applications in Clinical Oncology*. Springer; 2017:193-230.

196. Friendly M. *Discrete Data Analysis with R: Visualization and Modeling Techniques for Categorical and Count Data*. Chapman and Hall/CRC; 2015.

197. Moral RA, Hinde J, Demétrio CGB. Half-Normal Plots and Overdispersed Models in R: The hnp Package. *Journal of Statistical Software*. 2017;81(10):1-23.

198. Nisbet R, Miner G, Yale K. Model evaluation and enhancement. In: Miner G, Yale K, Nisbet R, eds. *Handbook of Statistical Analysis and Data Mining Applications*. Elsevier; 2018:215-234.

199. Yezli S, Khan A. COVID-19 pandemic: it is time to temporarily close places of worship and to suspend religious gatherings. *Journal of Travel Medicine*. 2021;28(2):taaa065.

200. Rocklöv J, Sjödin H. High population densities catalyse the spread of COVID-19. *Journal of Travel Medicine*. 2020;27(3):taaa038.

201. Saidan MN, Shbool MA, Arabeyyat OS, et al. Estimation of the probable outbreak size of novel coronavirus (COVID-19) in social gathering events and industrial activities. *Int J Infect Dis*. 2020;98:321-327.

202. Bengtson VL, Silverstein M, Putney NM, Harris SC. Does Religiousness Increase with Age? Age Changes and Generational Differences Over 35 Years. *Journal for the Scientific Study of Religion*. 2015;54(2):363-379.

203. Kang M, Wei J, Yuan J, et al. Probable Evidence of Fecal Aerosol Transmission of SARS-CoV-2 in a High-Rise Building. *Ann Intern Med*. 2020;173(12):974-980.

204. Niu J, Tung TCW. On-site quantification of re-entry ratio of ventilation exhausts in multi-family residential buildings and implications. *Indoor Air*. 2008;18(1):12-26.

205. Flores EO, Padilla A. Hidden threat: California COVID-19 surges and worker distress. Community and labor center at the University of California Merced. *Community and Labor Center at the University of California Merced*. Published online 2020.

206. Karaca-Mandic P, Sen S, Georgiou A, Zhu Y, Basu A. Association of COVID-19-Related Hospital Use and Overall COVID-19 Mortality in the USA. *J GEN INTERN MED*. Published online 2020.

207. Rushing WA, Wade GT. Community-structure constraints on distribution of physicians. *Health Serv Res*. 1973;8(4):283-297.

208. Gowrisankaran G, Town RJ. Competition, Payers, and Hospital Quality1. *Health Services Research*. 2003;38(6p1):1403-1422.

209. Rader B, Astley CM, Sy KTL, et al. Geographic access to United States SARS-CoV-2 testing sites highlights healthcare disparities and may bias transmission estimates. *Journal of Travel Medicine*. 2020;27(7):taaa076.

210. Ioannidis JPA, Axfors C, Contopoulos-Ioannidis DG. Population-level COVID-19 mortality risk for non-elderly individuals overall and for non-elderly individuals without underlying diseases in pandemic epicenters. *Environmental Research*. 2020;188:109890.

211. Centers for Disease Control and Prevention. COVID-19 Hospitalization and Death by Age. Centers for Disease Control and Prevention.

212. Oster AM. Trends in Number and Distribution of COVID-19 Hotspot Counties — United States, March 8–July 15, 2020. *MMWR Morb Mortal Wkly Rep*. 2020;69(33):1127-1132.

213. Centers for Disease Control and Prevention. People with Moderate to Severe Asthma. Centers for Disease Control and Prevention.

214. Wu Z, McGoogan JM. Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72 314 Cases From the Chinese Center for Disease Control and Prevention. *JAMA*. 2020;323(13):1239-1242.

215. Chhiba KD, Patel GB, Vu THT, et al. Prevalence and characterization of asthma in hospitalized and nonhospitalized patients with COVID-19. *Journal of Allergy and Clinical Immunology*. 2020;146(2):307-314.e4.

216. Wu X, Nethery RC, Sabath MB, Braun D, Dominici F. Air pollution and COVID-19 mortality in the United States: Strengths and limitations of an ecological regression analysis. *Science Advances*. 2020;6(45):eabd4049.

217. National Association of Counties. Coronavirus (COVID-19). National Association of Counties.

218. Dalsania AK, Fastiggi MJ, Kahlam A, et al. The Relationship Between Social Determinants of Health and Racial Disparities in COVID-19 Mortality. *J Racial and Ethnic Health Disparities*. 2022;9(1):288-295.

219. Golestaneh L, Neugarten J, Fisher M, et al. The association of race and COVID-19 mortality. *EClinicalMedicine*. 2020;25:100455.

220. Egede LE, Walker RJ. Structural Racism, Social Risk Factors, and Covid-19 — A Dangerous Convergence for Black Americans. *N Engl J Med*. 2020;383(12):e77.

221. Li B, Yang J, Zhao F, et al. Prevalence and impact of cardiovascular metabolic diseases on COVID-19 in China. *Clin Res Cardiol*. 2020;109(5):531-538.

222. Apicella M, Campopiano MC, Mantuano M, Mazoni L, Coppelli A, Del Prato S. COVID-19 in people with diabetes: understanding the reasons for worse outcomes. *Lancet Diabetes Endocrinol*. 2020;8(9):782-792.

223. Battegay M, Kuehl R, Tschudin-Sutter S, Hirsch HH, Widmer AF, Neher RA. 2019-novel Coronavirus (2019-nCoV): estimating the case fatality rate – a word of caution. *Swiss Medical Weekly*. 2020;150:w20203.

224. Kucirka LM, Lauer SA, Laeyendecker O, Boon D, Lessler J. Variation in False-Negative Rate of Reverse Transcriptase Polymerase Chain Reaction–Based SARS-CoV-2 Tests by Time Since Exposure. *Ann Intern Med*. 2020;173(4):262-267.

225. McGrory K, Woolington R. Florida medical examiners were releasing coronavirus death data. The state made them stop. Tampa Bay Times.

226. Dyer O. Covid-19: US states are not reporting vital data, says former CDC chief. *BMJ*. 2020;370:m2993.

227. Vatcheva KP, Lee M, McCormick JB, Rahbar MH. Multicollinearity in Regression Analyses Conducted in Epidemiologic Studies. *Epidemiology (Sunnyvale)*. 2016;6(2):227.

228. Hale T, Angrist N, Hale AJ, et al. Government responses and COVID-19 deaths: Global evidence across multiple pandemic waves. *PLOS ONE*. 2021;16(7):e0253116.

229. Florko N, Boodman E. How HHS's new hospital data reporting system will actually affect the U.S. Covid-19 response. STAT.

230. Co DO, Hogan LH, Karman J, Herbath M, Fabry Z, Sandor M. T Cell Interactions in Mycobacterial Granulomas: Non-Specific T Cells Regulate Mycobacteria-Specific T Cells in Granulomatous Lesions. *Cells*. 2021;10(12):3285.

231. Young K, Doernberg SB, Snedecor RF, Mallin E. Things We Do For No Reason: Contact Precautions for MRSA and VRE. *Journal of Hospital Medicine*. 2019;14(3):178-180.

232. Johnstone J, Shing E, Saedi A, et al. Discontinuing Contact Precautions for Vancomycin-Resistant Enterococcus (VRE) Is Associated With Rising VRE Bloodstream Infection Rates in Ontario Hospitals, 2009–2018: A Quasi-experimental Study. *Clinical Infectious Diseases*. 2020;71(7):1756-1759.

233. Czanner G. What to expect when meeting a statistician. *eye news*. 2014;21(2):2.

234. Chen T, Chen X, Zhang S, et al. The Genome Sequence Archive Family: Toward Explosive Data Growth and Diverse Data Types. *Genomics, Proteomics & Bioinformatics*. 2021;19(4):578-583.

235. Wieczorek J, Guerin C, McMahon T. K-fold cross-validation for complex sample surveys. *Stat*. 2022;11(1):e454.

236. West BT, Sakshaug JW, Kim Y. Analytic Error as an Important Component of Total Survey Error. In: *Total Survey Error in Practice*. John Wiley & Sons, Ltd; 2017:487-510.

237. West BT, Sakshaug JW, Aurelien GAS. How Big of a Problem is Analytic Error in Secondary Analyses of Survey Data? *PLOS ONE*. 2016;11(6):e0158120.

238. Ward BW. Analytic Errors in Analysis of Public Health Survey Data Are Avoidable. *Prev Chronic Dis*. 2018;15:E43.

239. Bishop D. How scientists can stop fooling themselves over statistics. *Nature*. 2020;584(7819):9-9.

240. Serão NVL, Petry AL, Sanglard LP, Rossoni-Serão MC, Bundy JM. Assessing the statistical training in animal science graduate programs in the United States: survey on statistical training. *Journal of Animal Science*. 2021;99(5):skab086.

241. Casadevall A, Ellis LM, Davies EW, McFall-Ngai M, Fang FC. A Framework for Improving the Quality of Research in the Biological Sciences. *mBio*. 2016;7(4):e01256-16.

242. Barraquand F, Ezard THG, Jørgensen PS, et al. Lack of quantitative training among early-career ecologists: a survey of the problem and potential solutions. *PeerJ*. 2014;2:e285.

243. Lang TA, Altman DG. Basic statistical reporting for articles published in Biomedical Journals: The "Statistical Analyses and Methods in the Published Literature" or the SAMPL Guidelines. *International Journal of Nursing Studies*. 2015;52(1):5-9.

244. Yan F, Robert M, Li Y. Statistical methods and common problems in medical or biomedical science research. *Int J Physiol Pathophysiol Pharmacol*. 2017;9(5):157-163.

245. Page MJ, Altman DG, McKenzie JE, et al. Flaws in the application and interpretation of statistical analyses in systematic reviews of therapeutic interventions were common: a cross-sectional analysis. *Journal of Clinical Epidemiology*. 2018;95:7-18.

246. Slutsky DJ. Statistical Errors in Clinical Studies. *J Wrist Surg*. 2013;2(4):285-287.

247. Bakker M, Wicherts JM. The (mis)reporting of statistical results in psychology journals. *Behav Res*. 2011;43(3):666-678.

248. Lang TA, Secic M. *How to Report Statistics in Medicine: Annotated Guidelines for Authors, Editors, and Reviewers*. 2nd ed. ACP Press; 2006.

249. Murray GD. Statistical aspects of research methodology. *British Journal of Surgery*. 1991;78(7):777-781.

250. George SL. Statistics in medical journals: A survey of current policies and proposals for editors. *Medical and Pediatric Oncology*. 1985;13(2):109-112.

251. Bansal M. Cardiovascular disease and COVID-19. *Diabetes Metab Syndr*. 2020;14(3):247-250.

252. Abrams EM, 't Jong GW, Yang CL. Asthma and COVID-19. *CMAJ*. 2020;192(20):E551.

253. Liang W, Guan W, Chen R, et al. Cancer patients in SARS-CoV-2 infection: a nationwide analysis in China. *Lancet Oncol*. 2020;21(3):335-337.

254. Zhao Q, Meng M, Kumar R, et al. The impact of COPD and smoking history on the severity of COVID-19: A systemic review and meta-analysis. *J Med Virol*. 2020;92(10):1915-1921.

255. Hussain A, Bhowmik B, do Vale Moreira NC. COVID-19 and diabetes: Knowledge in progress. *Diabetes Res Clin Pract*. 2020;162:108142.

256. Pranata R, Lim MA, Huang I, Raharjo SB, Lukito AA. Hypertension is associated with increased mortality and severity of disease in COVID-19 pneumonia: A systematic review, meta-analysis and meta-regression. *J Renin Angiotensin Aldosterone Syst*. 2020;21(2):1470320320926899.

257. Cai Q, Chen F, Wang T, et al. Obesity and COVID-19 Severity in a Designated Hospital in Shenzhen, China. *Diabetes Care*. 2020;43(7):1392-1398.

258. Aggarwal G, Lippi G, Michael Henry B. Cerebrovascular disease is associated with an increased disease severity in patients with Coronavirus Disease 2019 (COVID-19): A pooled analysis of published literature. *Int J Stroke*. 2020;15(4):385-389.

259. Dobson AP, Carper ER. Infectious Diseases and Human Population History: Throughout history the establishment of disease has been a side effect of the growth of civilization. *BioScience*. 1996;46(2):115-126.

260. CDC COVID-19 Response Team. Severe Outcomes Among Patients with Coronavirus Disease 2019 (COVID-19) — United States, February 12–March 16, 2020. *MMWR Morb Mortal Wkly Rep*. 2020;69(12):343-346.

261. Rubinson L, O'Toole T. Critical care during epidemics. *Crit Care*. 2005;9(4):311-313.

262. Ayanian JZ, Weissman JS, Schneider EC, Ginsburg JA, Zaslavsky AM. Unmet health needs of uninsured adults in the United States. *JAMA*. 2000;284(16):2061-2069.

263. Wilper AP, Woolhandler S, Lasser KE, McCormick D, Bor DH, Himmelstein DU. Health insurance and mortality in US adults. *Am J Public Health*. 2009;99(12):2289-2295.

264. Braveman PA, Cubbin C, Egerter S, Williams DR, Pamuk E. Socioeconomic disparities in health in the United States: what the patterns tell us. *Am J Public Health*. 2010;100(Suppl 1):S186-196.

265. Hahn RA, Truman BI. Education Improves Public Health and Promotes Health Equity. *Int J Health Serv*. 2015;45(4):657-678.

266. Hamad R, Elser H, Tran DC, Rehkopf DH, Goodman SN. How and why studies disagree about the effects of education on health: A systematic review and meta-analysis of studies of compulsory schooling laws. *Soc Sci Med*. 2018;212:168-178.

267. Athar HM, Chang MH, Hahn RA, Walker E, Yoon P, Centers for Disease Control and Prevention (CDC). Unemployment - United States, 2006 and 2010. *MMWR Suppl*. 2013;62(3):27-32.

268. Institute of Medicine (US) Committee on Health Care for Homeless People. *Homelessness, Health, and Human Needs*. National Academies Press; 1988.

269. Tsai J, Wilson M. COVID-19: a potential public health problem for homeless populations. *Lancet Public Health*. 2020;5(4):e186-e187.

270. Sullivan WF. Briefing Note: Primary Care Guidance Regarding COVID-19 for Those Who Care for Adults with Intellectual and Developmental Disabilities. Surrey Place.

271. Baker MG, McDonald A, Zhang J, Howden-Chapman P. *Infectious Diseases Attributable to Household Crowding in New Zealand: A Systematic Review and Burden of Disease Estimate*. He Kainga Oranga/Housing and Health Research Programme, University of Otago; 2013.

272. Nations U, Bonini A, Mukherjee S, Roig M, Lee YF. Responses to the COVID-19 Catastrophe Could Turn the Tide on Inequality. *UN Department of Economic and Social Affairs (DESA) Policy Briefs*. 2020;21:65.

273. Armitage R, Nellums LB. The COVID-19 response must be disability inclusive. *Lancet Public Health*. 2020;5(5):e257.

274. Shi L, Lebrun LA, Tsai J. The influence of English proficiency on access to care. *Ethn Health*. 2009;14(6):625-642.

275. Wimalawansa SJ. Global epidemic of coronaviruses--COVID-19: What can we do to minimize risks. *European Journal of Biomedical and Pharmaceutical Sciences*. 2020;7(3):432-438.

276. Bayham J, Fenichel EP. Impact of school closures for COVID-19 on the US health-care workforce and net mortality: a modelling study. *The Lancet Public Health*. 2020;5(5):e271-e278.

277. Feagin J, Bennefield Z. Systemic racism and U.S. health care. *Soc Sci Med*. 2014;103:7-14.

278. Katayama S, Tomaru Y, Kasukawa T, et al. Antisense transcription in the mammalian transcriptome. *Science*. 2005;309(5740):1564-1566.

279. Beiter T, Reich E, Williams RW, Simon P. Antisense transcription: a critical look in both directions. *Cell Mol Life Sci*. 2009;66(1):94-112.

280. Georg J, Hess WR. cis-antisense RNA, another level of gene regulation in bacteria. *Microbiol Mol Biol Rev*. 2011;75(2):286-300.

281. Lybecker M, Bilusic I, Raghavan R. Pervasive transcription: detecting functional RNAs in bacteria. *Transcription*. 2014;5(4):e944039.

282. Gripenland J, Netterling S, Loh E, Tiensuu T, Toledo-Arana A, Johansson J. RNAs: regulators of bacterial virulence. *Nat Rev Microbiol*. 2010;8(12):857-866.

283. Gottesman S, Storz G. Bacterial small RNA regulators: versatile roles and rapidly evolving variations. *Cold Spring Harb Perspect Biol*. 2011;3(12):a003798.

284. Kacharia FR, Millar JA, Raghavan R. Emergence of New sRNAs in Enteric Bacteria is Associated with Low Expression and Rapid Evolution. *J Mol Evol*. 2017;84(4):204-213.

285. Waters LS, Storz G. Regulatory RNAs in bacteria. *Cell*. 2009;136(4):615-628.

286. Thomason MK, Storz G. Bacterial antisense RNAs: how many are there, and what are they doing? *Annu Rev Genet*. 2010;44:167-188.

287. Storz G, Vogel J, Wassarman KM. Regulation by small RNAs in bacteria: expanding frontiers. *Mol Cell*. 2011;43(6):880-891.

288. Updegrove TB, Shabalina SA, Storz G. How do base-pairing small RNAs evolve? *FEMS Microbiol Rev*. 2015;39(3):379-391.

289. Saberi F, Kamali M, Najafi A, Yazdanparast A, Moghaddam MM. Natural antisense RNAs as mRNA regulatory elements in bacteria: a review on function and applications. *Cell Mol Biol Lett*. 2016;21:6.

290. Hoynes-O'Connor A, Moon TS. Development of Design Rules for Reliable Antisense RNA Behavior in E. coli. *ACS Synth Biol*. 2016;5(12):1441-1454.

291. Dutcher HA, Raghavan R. Origin, Evolution, and Loss of Bacterial Small RNAs. *Microbiology Spectrum*. 2018;6(2):6.2.12.

292. Johnson JM, Edwards S, Shoemaker D, Schadt EE. Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. *Trends Genet*. 2005;21(2):93-102.

293. Perocchi F, Xu Z, Clauder-Münster S, Steinmetz LM. Antisense artifacts in transcriptome microarray experiments are resolved by actinomycin D. *Nucleic Acids Res*. 2007;35(19):e128.

294. Georg J, Voss B, Scholz I, Mitschke J, Wilde A, Hess WR. Evidence for a major role of antisense RNAs in cyanobacterial gene regulation. *Mol Syst Biol*. 2009;5:305.

295. Yamaguchi Y, Park JH, Inouye M. Toxin-antitoxin systems in bacteria and archaea. *Annu Rev Genet*. 2011;45:61-79.

296. Georg J, Hess WR. Widespread Antisense Transcription in Prokaryotes. *Microbiology Spectrum*. 2018;6(4):6.4.12.

297. Dornenburg JE, Devita AM, Palumbo MJ, Wade JT. Widespread antisense transcription in Escherichia coli. *mBio*. 2010;1(1):e00024-10.

298. Raghavan R, Sloan DB, Ochman H. Antisense transcription is pervasive but rarely conserved in enteric bacteria. *mBio*. 2012;3(4):e00156-12.

299. Thomason MK, Bischler T, Eisenbart SK, et al. Global transcriptional start site mapping using differential RNA sequencing reveals novel antisense RNAs in Escherichia coli. *J Bacteriol*. 2015;197(1):18-28.

300. Giangrossi M, Prosseda G, Tran CN, Brandi A, Colonna B, Falconi M. A novel antisense RNA regulates at transcriptional level the virulence gene icsA of Shigella flexneri. *Nucleic Acids Res*. 2010;38(10):3362-3375.

301. Lejars M, Kobayashi A, Hajnsdorf E. Physiological roles of antisense RNAs in prokaryotes. *Biochimie*. 2019;164:3-16.

302. Jerome LJ, van Biesen T, Frost LS. Degradation of FinP antisense RNA from F-like plasmids: the RNA-binding protein, FinO, protects FinP from ribonuclease E. *J Mol Biol*. 1999;285(4):1457-1473.

303. Koraimann G. Spread and Persistence of Virulence and Antibiotic Resistance Genes: A Ride on the F Plasmid Conjugation Module. *EcoSal Plus*. 2018;8(1).

304. Arthur DC, Ghetu AF, Gubbins MJ, Edwards RA, Frost LS, Glover JNM. FinO is an RNA chaperone that facilitates sense-antisense RNA interactions. *EMBO J*. 2003;22(23):6346-6355.

305. Glover JN, Chaulk SG, Edwards RA, Arthur D, Lu J, Frost LS. The FinO family of bacterial RNA chaperones. *Plasmid*. 2015;78:79-87.

306. Gerdes K, Wagner EGH. RNA antitoxins. *Curr Opin Microbiol*. 2007;10(2):117-124.

307. Ogata RT, Levine RP. Characterization of complement resistance in Escherichia coli conferred by the antibiotic resistance plasmid R100. *J Immunol*. 1980;125(4):1494-1498.

308. Cox KEL, Schildbach JF. Sequence of the R1 plasmid and comparison to F and R100. *Plasmid*. 2017;91:53-60.

309. Pecota DC, Osapay G, Selsted ME, Wood TK. Antimicrobial properties of the Escherichia coli R1 plasmid host killing peptide. *J Biotechnol*. 2003;100(1):1-12.

310. Gerdes K, Thisted T, Martinussen J. Mechanism of post-segregational killing by the hok/sok system of plasmid R1: sok antisense RNA regulates formation of a hok mRNA species correlated with killing of plasmid-free cells. *Mol Microbiol*. 1990;4(11):1807-1818.

311. Chukwudi CU, Good L. The role of the hok/sok locus in bacterial response to stressful growth conditions. *Microb Pathog*. 2015;79:70-79.

312. Baltrus DA. Exploring the costs of horizontal gene transfer. *Trends Ecol Evol*. 2013;28(8):489-495.

313. Vogwill T, MacLean RC. The genetic basis of the fitness costs of antimicrobial resistance: a meta-analysis approach. *Evol Appl*. 2015;8(3):284-295.

314. Millan AS, MacLean RC. Fitness Costs of Plasmids: a Limit to Plasmid Transmission. *Microbiology Spectrum*. 2017;5(5):5.5.02.

315. Pluta R, Espinosa M. Antisense and yet sensitive: Copy number control of rolling circle-replicating plasmids by small RNAs. *Wiley Interdiscip Rev RNA*. 2018;9(6):e1500.

316. Lacatena RM, Cesareni G. Base pairing of RNA I with its complementary sequence in the primer precursor inhibits ColE1 replication. *Nature*. 1981;294(5842):623-626.

317. Spangler R, Zhang SP, Krueger J, Zubay G. Colicin synthesis and cell death. *J Bacteriol*. 1985;163(1):167-173.

318. Riley MA, Gordon DM. The ecological role of bacteriocins in bacterial competition. *Trends Microbiol*. 1999;7(3):129-133.

319. del Solar G, Espinosa M. Plasmid copy number control: an ever-growing story. *Mol Microbiol*. 2000;37(3):492-500.

320. Sengupta M, Austin S. Prevalence and significance of plasmid maintenance functions in the virulence plasmids of pathogenic bacteria. *Infect Immun*. 2011;79(7):2502-2509.

321. Bourque G, Burns KH, Gehring M, et al. Ten things you should know about transposable elements. *Genome Biol*. 2018;19(1):199.

322. Murray GL, Morel V, Cerqueira GM, et al. Genome-wide transposon mutagenesis in pathogenic Leptospira species. *Infect Immun*. 2009;77(2):810-816.

323. Murray JL, Kwon T, Marcotte EM, Whiteley M. Intrinsic Antimicrobial Resistance Determinants in the Superbug Pseudomonas aeruginosa. *mBio*. 2015;6(6):e01603-01615.

324. Kalindamar S, Lu J, Abdelhamed H, Tekedar HC, Lawrence ML, Karsi A. Transposon mutagenesis and identification of mutated genes in growth-delayed Edwardsiella ictaluri. *BMC Microbiol*. 2019;19(1):55.

325. Arciola CR, Campoccia D, Gamberini S, et al. Search for the insertion element IS256 within the ica locus of Staphylococcus epidermidis clinical isolates collected from biomaterial-associated infections. *Biomaterials*. 2004;25(18):4117-4125.

326. Kiem S, Oh WS, Peck KR, et al. Phase variation of biofilm formation in Staphylococcus aureus by IS 256 insertion and its impact on the capacity adhering to polyurethane surface. *J Korean Med Sci*. 2004;19(6):779-782.

327. Perez M, Calles-Enríquez M, del Rio B, et al. IS256 abolishes gelatinase activity and biofilm formation in a mutant of the nosocomial pathogen Enterococcus faecalis V583. *Can J Microbiol*. 2015;61(7):517-519.

328. Lapierre L, Mollet B, Germond JE. Regulation and adaptive evolution of lactose operon expression in Lactobacillus delbrueckii. *J Bacteriol*. 2002;184(4):928-935.

329. Christie-Oleza JA, Nogales B, Martín-Cardona C, et al. ISPst9, an ISL3-like insertion sequence from Pseudomonas stutzeri AN10 involved in catabolic gene inactivation. *Int Microbiol*. 2008;11(2):101-110.

330. Moffatt JH, Harper M, Adler B, Nation RL, Li J, Boyce JD. Insertion Sequence ISAba11 Is Involved in Colistin Resistance and Loss of Lipopolysaccharide in Acinetobacter baumannii. *Antimicrobial Agents and Chemotherapy*. 2011;55(6):3022-3024.

331. Ma C, Simons RW. The IS10 antisense RNA blocks ribosome binding at the transposase translation initiation site. *EMBO J*. 1990;9(4):1267-1274.

332. Ellis MJ, Trussler RS, Haniford DB. Hfq binds directly to the ribosome-binding site of IS10 transposase mRNA to inhibit translation. *Mol Microbiol*. 2015;96(3):633-650.

333. Echols H. Developmental pathways for the temperate phage: lysis vs lysogeny,. *Annu Rev Genet*. 1972;6:157-190.

334. Touchon M, Bernheim A, Rocha EP. Genetic and life-history traits associated with the distribution of prophages in bacteria. *ISME J*. 2016;10(11):2744-2754.

335. Seed KD, Yen M, Shapiro BJ, et al. Evolutionary consequences of intra-patient phage predation on microbial populations. *Elife*. 2014;3:e03497.

336. León M, Bastías R. Virulence reduction in bacteriophage resistant bacteria. *Front Microbiol*. 2015;6:343.

337. Harrison E, Brockhurst MA. Ecological and Evolutionary Benefits of Temperate Phage: What Does or Doesn't Kill You Makes You Stronger. *BioEssays*. 2017;39(12):1700112.

338. Howard-Varona C, Hargreaves KR, Abedon ST, Sullivan MB. Lysogeny in nature: mechanisms, impact and ecology of temperate phages. *ISME J*. 2017;11(7):1511-1520.

339. Liao SM, Wu TH, Chiang CH, Susskind MM, McClure WR. Control of gene expression in bacteriophage P22 by a small antisense RNA. I. Characterization in vitro of the Psar promoter and the sar RNA transcript. *Genes Dev*. 1987;1(2):197-203.

340. Schaefer KL, McClure WR. Antisense RNA control of gene expression in bacteriophage P22. I. Structures of sar RNA and its target, ant mRNA. *RNA*. 1997;3(2):141-156.

341. Toledo-Arana A, Dussurget O, Nikitas G, et al. The Listeria transcriptional landscape from saprophytism to virulence. *Nature*. 2009;459(7249):950-956.

342. Lebreton A, Cossart P. RNA- and protein-mediated control of Listeria monocytogenes virulence gene expression. *RNA Biol*. 2017;14(5):460-470.

343. Hayashi F, Smith KD, Ozinsky A, et al. The innate immune response to bacterial flagellin is mediated by Toll-like receptor 5. *Nature*. 2001;410(6832):1099-1103.

344. Sesto N, Wurtzel O, Archambaud C, Sorek R, Cossart P. The excludon: a new concept in bacterial antisense RNA-mediated gene regulation. *Nat Rev Microbiol*. 2013;11(2):75-82.

345. Parrow NL, Fleming RE, Minnick MF. Sequestration and scavenging of iron in infection. *Infect Immun*. 2013;81(10):3503-3514.

346. Fonseca MV, Swanson MS. Nutrient salvaging and metabolism by the intracellular pathogen Legionella pneumophila. *Front Cell Infect Microbiol*. 2014;4:12.

347. Chen Q, Crosa JH. Antisense RNA, fur, iron, and the regulation of iron transport genes in Vibrio anguillarum. *J Biol Chem*. 1996;271(31):18885-18891.

348. León-Sicairos N, Angulo-Zamudio UA, de la Garza M, Velázquez-Román J, Flores-Villaseñor HM, Canizalez-Román A. Strategies of Vibrio parahaemolyticus to acquire nutritional iron during host colonization. *Front Microbiol*. 2015;6:702.

349. Naka H, López CS, Crosa JH. Role of the pJM1 plasmid-encoded transport proteins FatB, C and D in ferric anguibactin uptake in the fish pathogen Vibrio anguillarum. *Environ Microbiol Rep*. 2010;2(1):104-111.

350. Mellin JR, Tiensuu T, Bécavin C, Gouin E, Johansson J, Cossart P. A riboswitch-regulated antisense RNA in Listeria monocytogenes. *Proceedings of the National Academy of Sciences*. 2013;110(32):13132-13137.

351. Bobik TA, Xu Y, Jeter RM, Otto KE, Roth JR. Propanediol utilization genes (pdu) of Salmonella typhimurium: three genes for the propanediol dehydratase. *J Bacteriol*. 1997;179(21):6633-6639.

352. Degnan PH, Taga ME, Goodman AL. Vitamin B12 as a modulator of gut microbial ecology. *Cell Metab*. 2014;20(5):769-778.

353. Toraya T, Honda S, Fukui S. Fermentation of 1,2-propanediol with 1,2-ethanediol by some genera of Enterobacteriaceae, involving coenzyme B12-dependent diol dehydratase. *J Bacteriol*. 1979;139(1):39-47.

354. Conner CP, Heithoff DM, Julio SM, Sinsheimer RL, Mahan MJ. Differential patterns of acquired virulence genes distinguish Salmonella strains. *Proceedings of the National Academy of Sciences*. 1998;95(8):4641-4645.

355. Ji Y, Yin D, Fox B, Holmes DJ, Payne D, Rosenberg M. Validation of antibacterial mechanism of action using regulated antisense RNA expression in Staphylococcus aureus. *FEMS Microbiol Lett*. 2004;231(2):177-184.

356. Nikravesh A, Dryselius R, Faridani OR, et al. Antisense PNA accumulates in Escherichia coli and mediates a long post-antibiotic effect. *Mol Ther*. 2007;15(8):1537-1542.

357. Sturino JM, Klaenhammer TR. Engineered bacteriophage-defence systems in bioprocessing. *Nat Rev Microbiol*. 2006;4(5):395-404.

358. Alessandra S, Alessandro T, Flavio S, Alejandro H. Artificial antisense RNAs silence lacZ in E. coli by decreasing target mRNA concentration. *BMB Rep*. 2008;41(8):568-574.

359. Suzuki Y, Ishimoto T, Fujita S, et al. Antimicrobial antisense RNA delivery to F-pili producing multidrug-resistant bacteria via a genetically engineered bacteriophage. *Biochem Biophys Res Commun*. 2020;530(3):533-540.

360. Good L, Stach JEM. Synthetic RNA silencing in bacteria - antimicrobial discovery and resistance breaking. *Front Microbiol*. 2011;2:185.

361. Sealy L, Saunders C, Blume J, Chalkley R. The GRE over the entire range of scores lacks predictive ability for PhD outcomes in the biomedical sciences. *PLoS One*. 2019;14(3):e0201634.

362. Galea S. Rethinking our student admission requirements: A case for equity. Boston University School of Public Health.

363. Association of Schools and Programs of Public Health. GRE scores and public health admissions. Association of Schools and Programs of Public Health.

364. Colorado State University. Colorado School of Public Health drops its GRE requirement for admission. Colorado School of Public Health.

365. Wolfson J. In 2 weeks, @PublicHealthUMN will remove its number from ETS, going #GRExit for all programs. Our decision was largely based on the results of a RANDOMIZED assessment of how GRE scores influence admissions decisions. What we did and what we found: a thread. 1/n. @DrJWolfson.

366. Miller C, Stassun K. A test that fails. *Nature*. 2014;510(7504):303-304.

367. Hu JC. Graduate programs drop GRE after online version raises concerns about fairness. Science Careers.

368. Anderson M. Mobile Technology and Home Broadband 2019. Pew Research Center.

369. Gonzales AL, McCrory Calarco J, Lynch T. Technology Problems and Student Achievement Gaps: A Validation and Extension of the Technology Maintenance Construct. *Communication Research*. 2020;47(5):750-770.

370. American Astronomical Society. AAS statement on limiting the use of GRE scores in graduate admissions in the astronomical sciences. American Astronomical Society.

371. Hall J. Here's the current list of bio/biomedical graduate programs that dropped, or plan to drop, the GRE. My plan is to maintain this list as a resource, so please let me know if there are additions/corrections! @jdhallphd.

372. Langin K. Ph.D. programs drop standardized exam. *Science*. 2019;364(6443):816.

373. Millar JA. I created a list of #PublicHealth programs dropping the GRE requirement. Let me know if there are others that can be added. (And thank you @jdhallphd for creating a bio/biomedical #GRExit list!). @JAMicrobe.

374. Rutgers School of Public Health. Admissions. State University of New Jersey.

375. Rollins School of Public Health. Admissions. Emory University.

376. Millar JA. You're welcome! I'm going through right now and adding programs that are temporarily removing the GRE. @JAMicrobe.

377. Millar JA. Public Health Graduate Programs Dropping GRE Requirement. Google Docs.

378. Petersen SL. Developing a More Predictive Model for Identifying STEM Doctoral Students Who Are Likely to Succeed | Beyond the GRE. Beyond the GRE.

379. Millar JA. I'm updating the public health #GRExit list and it's amazing to see the changes in the last month! @UIowaCPH originally just waived for #COVID19 and now it's permanent for MPH. Also permanent for @OSUPublicHealth MPH. Let's keep it going! #EpiTwitter. @JAMicrobe.

380. Millar JA. More #PublicHealth #GRE waiver news. @CornellMPH & @ETSUCPH converted temp waivers to permanent for MPH. No word if @ETSUCPH will extend the temp waiver for DrPH to Fall 2021. @TulaneSPHTM & @VUmedicine extended temp waivers for masters to Fall 2021. @JAMicrobe.

381. University of Washington School of Public Health. UW Department of Epidemiology Eliminates GRE Requirement in Graduate Program Admissions. University of Washington School of Public Health.

382. UNC Gillings School of Global Public Health. Gillings School drops GRE requirement for 8 graduate degrees. UNC Gillings School of Global Public Health.

383. Wolfson RJ. The Role of the GRE in SPH Graduate Admissions. Presented at: University of Minnesota; May 27, 2020; Virtual.

384. University of Michigan School of Public Health. University of Michigan School of Public Health Removes GRE Requirement. University of Michigan School of Public Health.